# GRI-QA: a Comprehensive Benchmark for Table Question Answering over Environmental Data

**Michele Luca Contalbo**, **Sara Pederzoli**, **Francesco Del Buono**,
**Valeria Venturelli**, **Francesco Guerra**, **Matteo Paganelli**

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

## Abstract

Assessing corporate environmental sustainability with Table Question Answering systems is challenging due to complex tables, specialized terminology, and the variety of questions they must handle. In this paper, we introduce GRI-QA, a test benchmark designed to evaluate Table QA approaches in the environmental domain. Using GRI standards, we extract and annotate tables from non-financial corporate reports, generating question-answer pairs through a hybrid LLM-human approach. The benchmark includes eight datasets, categorized by the types of operations required, including operations on multiple tables from multiple documents. Our evaluation reveals a significant gap between human and model performance, particularly in multi-step reasoning, highlighting the relevance of the benchmark and the need for further research in domain-specific Table QA. Code and benchmark datasets are available at `https://github.com/softlab-unimore/gri_qa`.

## 1 Introduction

Sustainability accounting is crucial to global regulatory efforts for corporate environmental transparency. Initiatives like the European Green Deal (European Commission, 2024) and the United Nations' sustainability agenda (SDGS, 2024) require publicly listed companies to disclose environmental data through non-financial reports adhering to consolidated standards, such as the Global Reporting Initiative (GRI) framework (GRI, 2024). These reports provide crucial information, in particular within many and large tables, to determine whether companies are adopting responsible environmental practices or engaging in greenwashing, i.e., overstating their sustainability achievements to appear more environmentally responsible than they truly are (Nemes et al., 2022; Moodaley and Telukdarie, 2023; de Freitas Netto et al., 2020).

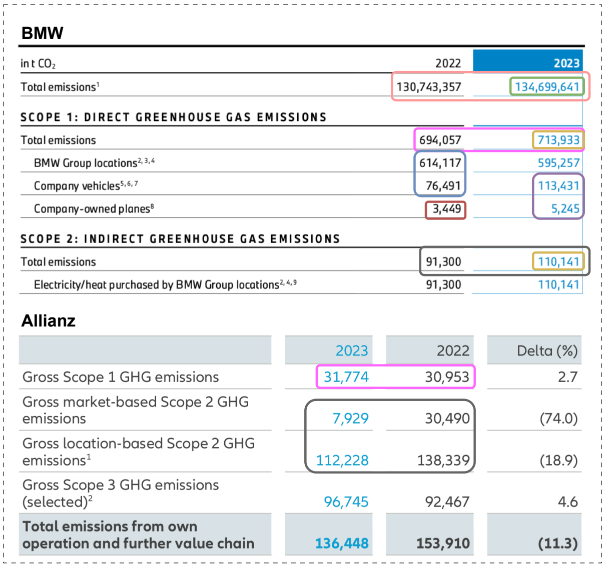
However, automatically analyzing non-financial reports introduces many challenges due to (1) the **format and structure of the tables**, which lack standardization across companies and often feature hierarchical layouts combining top and side headers, (2) **specialized terminologies**, as environmental tables frequently include industry-specific terms which, combined with performance metrics with varying measurement conventions complicates their interpretation, and (3) the **nature and variety of the questions** posed by analysts, which can range from simple value extraction to complex calculations involving multiple elements within the same table, across tables within the same document, or even across tables from different documents.

These challenges highlight the need for models capable of understanding complex table structures, applying context-aware reasoning, and combining general semantic knowledge with domain expertise. While similar datasets exist in other sectors, as highlighted in Section 2, an environmental domain-specific benchmark is still missing and could represent a significant step forward. Such a benchmark would be valuable for both domain specialization in table question answering (QA) models and for its technical features, requiring reasoning across varying numbers of tables, an uncommon aspect in other datasets.

To address this gap, we introduce GRI-QA, a new Table QA test benchmark on environmental tables extracted from corporate reports, with questions categorized according to the GRI standard. GRI-QA was designed through a methodology that requires input from domain experts (Section 3) to reflect the specific information needs in the sector. Figure 1 shows the question types present in GRI-QA and provides an example of question-answer using two tables containing BMW and Allianz $CO_2$ emissions statistics from the 2023 non-financial report. These question types feature *extractive questions* that require straightforward data retrieval, and

*Findings of the Association for Computational Linguistics: ACL 2025*, pages 15764–15779
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

| Question type | Question | Answer |
|---|---|---|
| Extractive | What is the total CO2 emissions in tons generated in 2023? | 134,699,641 |
| Extractive hierarchical | What are the direct emissions from company owned planes in 2022? | 3,449 |
| Comparison | Are the CO2 emissions in 2023 lower than those in 2022? | No |
| Superlative | What is the maximum CO2 value among direct and indirect emissions in 2023? | 713,933 |
| Ranking | What are the 2023 emissions of company vehicles, and planes in ascending order? | 5,245; 113,431 |
| Sum | What is the sum of the CO2 emissions across 2022 and 2023? | 265,442,998 |
| Average | What is the average total CO2 emission across 2022 and 2023? | 132,721,499 |
| Percentage change | What is the percentage increase in CO2 emissions in 2023 compared to 2022? | 3.26% |
| Multi-step Superlative | Which company has the highest amount of Scope 2 emissions over 2023 and 2022? | Allianz |
| Multi-step Ranking | What are the two average values of direct GHG emissions in the years 2022 and 2023 in ascending order among all the companies? | 31,363.5; 703,995 |
| Multi-step Sum | What is the total amount of direct emissions in 2022 and 2023 among all the companies? | 1,470,717 |
| Multi-step Average | What is the average percentage variation of direct emissions from 2023 to 2022 among all the companies? | 2.48% |



BMW

| in t $CO_2$ | 2022 | 2023 |
|---|---|---|
| Total emissions[1] | 130,743,357 | 134,699,641 |
| SCOPE 1: DIRECT GREENHOUSE GAS EMISSIONS | | |
| Total emissions | 694,057 | 713,933 |
| BMW Group locations[2, 3, 4] | 614,117 | 595,257 |
| Company vehicles[5, 6, 7] | 76,491 | 113,431 |
| Company-owned planes[8] | 3,449 | 5,245 |
| SCOPE 2: INDIRECT GREENHOUSE GAS EMISSIONS | | |
| Total emissions | 91,300 | 110,141 |
| Electricity/heat purchased by BMW Group locations[2, 4, 9] | 91,300 | 110,141 |

Allianz

| | 2023 | 2022 | Delta (%) |
|---|---|---|---|
| Gross Scope 1 GHG emissions | 31,774 | 30,953 | 2.7 |
| Gross market-based Scope 2 GHG emissions | 7,929 | 30,490 | (74.0) |
| Gross location-based Scope 2 GHG emissions[1] | 112,228 | 138,339 | (18.9) |
| Gross Scope 3 GHG emissions (selected)[2] | 96,745 | 92,467 | 4.6 |
| **Total emissions from own operation and further value chain** | **136,448** | **153,910** | **(11.3)** |

Figure 1: Examples of tables with related questions in the GRI-QA benchmark. The colored boxes on the tables indicate the input values considered to compute the respective answer. The answers may span multiple tables.

*hierarchical questions* which require disambiguating terms based on the table's hierarchical structure. GRI-QA contains calculated questions that can be *relational*, focusing on understanding numerical relationships between table entries, such as comparisons (blue box in the Figure), superlatives (yellow box), or rankings (purple box). It also includes *quantitative questions*, which demand precise computations (i.e., sums, averages, or percentage variations in pink boxes in the Figure) using numerical data. Moreover, GRI-QA proposes *multi-step* questions which require to process multiple operations on (multiple) tables from (multiple) documents (fuchsia boxes). Some of these questions may require a textual value as an answer, as in the example for the multi-step superlative question.

To assess the benchmark's complexity, we evaluated several state-of-the-art tabular question-answering systems and GPT models. The experimental results in Section 4 highlight a significant performance gap between humans and models, particularly in multi-step and multi-table reasoning. While GPT-based models with CoT prompting achieve strong results on simpler tasks, they still struggle with more complex ones, especially when dealing with multi-table scenarios. Financial models show promise but exhibit greater variability depending on the dataset. These findings underscore the benchmark's relevance and suggest the need for further research, particularly in addressing complex, multi-step and multi-table questions.

In summary, our contributions are threefold: (1) we introduce GRI-QA, a publicly available benchmark designed for Question Answering on environmental tables from corporate reports, including multi-table and multi-document reasoning; (2) we propose a methodology to create the benchmark with the support of domain experts; and (3) we evaluate state-of-the-art Table QA models on GRI-QA, highlighting their limitations and outlining directions for future research.

## 2 Related Work

### 2.1 Table question answering

In recent years, Table QA has become a prominent research area, driving the development of various approaches. Most methods rely on tabular language models (Badaro et al., 2023) and large language models (Sui et al., 2024; Zhang et al., 2024a; Xie et al., 2023; Zhu et al., 2024; Wang et al., 2024b), which enable a deep understanding of queries, tables, and their relations. To improve the performance of these methods across diverse scenarios and domains, several benchmarks have been introduced, with key statistics summarized in Table 1. Many of the proposed datasets are based on tables extracted from Wikipedia and focus on different methods of answer generation, such as direct answer generation (e.g., WTQ Pasupat and Liang, 2015, NQ-Tables Herzig et al., 2021), SQL query generation (e.g., WikiSQL Zhong et al., 2017, SPIDER Yu et al., 2018), free-form text generation (e.g., FeTaQA Nan et al., 2022), and multi-hop question answering using both tabular and textual contexts (e.g., OTT-QA Chen et al., 2021a,

| Dataset | Domain | Data type | Task | Num. | Hier. | Multi-table |
|---|---|---|---|---|---|---|
| WTQ (Pasupat and Liang, 2015) | Wikipedia | Table | Table QA | ✗ | ✗ | ✗ |
| NQ-Tables (Herzig et al., 2021) | Wikipedia | Table | Table QA | ✗ | ✗ | ✗ |
| WikiSQL (Zhong et al., 2017) | Wikipedia | Table | Text-to-SQL | ✓ | ✗ | ✓ |
| Spider (Yu et al., 2018) | Wikipedia | Table | Text-to-SQL | ✓ | ✗ | ✓ |
| BIRD (Li et al., 2023) | Kaggle | Table | Text-to-SQL | ✓ | ✗ | ✓ |
| FeTaQA (Nan et al., 2022) | Wikipedia | Table | Table QA | ✗ | ✗ | ✗ |
| HybridQA (Chen et al., 2020) | Wikipedia | Table & Text | Hybrid Table QA | ✓ | ✗ | ✗ |
| OTT-QA (Chen et al., 2021a) | Wikipedia | Table & Text | Hybrid Table QA | ✓ | ✗ | ✗ |
| TAT-QA (Zhu et al., 2021) | Finance | Table & Text | Hybrid Table QA | ✓ | ✓* | ✗ |
| Fin-QA (Chen et al., 2021b) | Finance | Table or Text | QA | ✓ | ✓* | ✗ |
| PACIFIC (Deng et al., 2022) | Finance | Table & Text | Conversational TQA | ✓ | ✓* | ✗ |
| ConvFinQA (Chen et al., 2022) | Finance | Table or Text | Conversational QA | ✓ | ✓* | ✗ |
| DocFinQA (Reddy et al., 2024) | Finance | Table & Text | Long-document QA | ✓ | ✓* | ✗ |
| AIT-QA (Katsis et al., 2022) | Airlines | Table | Table QA | ✓ | ✓ | ✗ |
| HiTab (Cheng et al., 2022) | Stat. reports, Wiki | Table | Table QA | ✓ | ✓ | ✗ |
| MMQA-QA (Wu et al., 2025) | Wikipedia | Relational Table | Table QA | ✓ | ✗ | ✓ |
| MultiTabQA (Pal et al., 2023) | Wikipedia | Relational Table | Table QA | ✓ | ✗ | ✓ |
| MultiHiertt (Zhao et al., 2022) | Finance | Table | Hybrid Table QA | ✓ | ✓* | ✓ |
| **GRI-QA** (ours) | Environment | Table | Table QA | ✓ | ✓ | ✓ |

Table 1: Comparison of Table QA benchmarks. "Num." refers to questions requiring numerical reasoning and "Hier." to hierarchical questions. The symbol ✓* refers to hierarchical questions that have not been explicitly annotated.

Hybrid-QA Chen et al., 2020). Alongside these general-purpose datasets, several domain-specific datasets have also been introduced, focusing on areas such as finance, airlines (e.g., AIT-QA Katsis et al., 2022), and a combination of several domains (e.g., HiTab Cheng et al., 2022, TableLLM-bench Zhang et al., 2024b, TableInstruct-QA Zhang et al., 2024a, TableBench Wu et al., 2024, FLARE-QA Xie et al., 2023). The datasets in the financial domain are particularly challenging because they require advanced numerical reasoning (e.g., FinQA Chen et al., 2021b), analyzing long corporate documents (e.g., DocFinQA Reddy et al., 2024, ConvFinQA Chen et al., 2022) and solving hybrid QA scenarios where both text and table content need to be aligned (e.g., TAT-QA Zhu et al., 2021, PACIFIC Deng et al., 2022). A common limitation of these benchmarks is their focus on queries involving either single tables (in most cases) or multiple tables with fixed relational schemas, as in MMQA (Wu et al., 2025), MultiTabQA (Pal et al., 2023), and text-to-SQL benchmarks (Zhong et al., 2017; Yu et al., 2018; Li et al., 2023). The only exception is MultiHiertt (Zhao et al., 2022), which includes queries spanning multiple non-relational tables within the financial domain. While similar to GRI-QA in handling multi-table queries, MultiHiertt extracts the tables from a single document, leading to less variability in table structures and vocabulary compared to those considered in GRI-QA.

### 2.2 Environmental data analysis

Environmental data analysis encompasses a broad range of tasks as ESG (**E**nvironmental, **S**ocial and **G**overnance) text classification (Xia et al., 2024; Mehra et al., 2022; Pavlova et al., 2024; Webersinke et al., 2021; Schimanski et al., 2023, 2024), topic detection (Varini et al., 2020; Nugent et al., 2021), claim detection and verification (Stammbach et al., 2022; Diggelmann et al., 2020), question answering (Luccioni et al., 2020), and greenwashing detection (Nemes et al., 2022; Moodaley and Telukdarie, 2023; de Freitas Netto et al., 2020; Mahdavi et al., 2024).

The development of these methods is often supported by specialized datasets. An instruction-tuned ESG news classification dataset was introduced in Xia et al. (2024) to train the ESGLlama model. Similarly, Schimanski et al. (2024) proposed several datasets for pre-training and fine-tuning ESG models[1]. ClimateBERT (Webersinke et al., 2021) utilizes over 2 million climate-related paragraphs for text classification, while ESG-FTSE (Pavlova et al., 2024) and multilingual ESG issues dataset (Chen et al., 2023) focus on ESG topic categorization. For claim verification, the Environmental Claims dataset (Stammbach et al., 2022) and CLIMATE-FEVER (Diggelmann et al., 2020) provide labeled claims with supporting evidence. Finally, ontologies with environmental standards (Zhou and Perzylo, 2023; Usmanova and Usbeck, 2024) have been developed to improve the

[1] https://huggingface.co/ESGBERT

| GRI-QA datasets | Question types | Count (#) | GRI (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 301 | 302 | 303 | 304 | 305 | 306 | 308 |
| `extra` | Extractive (100%) | 1503 | 3.5 | 24.4 | 6.1 | 2.1 | 50.1 | 7.6 | 6.2 |
| `hier` | Extractive hierarchical (100%) | 502 | 3.2 | 17.9 | 11.2 | 0.4 | 55.6 | 9 | 2.8 |
| `rel` | Comparison (40.3%), Superlative (27%), Ranking (32.7%) | 248 | 0.8 | 30.2 | 8.1 | 0 | 44.8 | 16.1 | 0 |
| `quant` | Sum (21.4%), Average (22.6%), Percentage change (56%) | 266 | 1.8 | 27.6 | 7.7 | 0 | 48.6 | 14.3 | 0 |
| `step` | *Multi-Step* Superlative (35.5%), Ranking (32.5%), Sum (10.2%), Average (21.7%) | 166 | 3.6 | 29.5 | 13.9 | 0 | 38.6 | 14.5 | 0 |
| `mrel` | *Multi-Table* Superlative (38.3%), Ranking (61.7%) | 619 | 0 | 20 | 7 | 0 | 66.9 | 6.1 | 0 |
| `mquant` | *Multi-Table* Sum (43.1%), Average (56.9%) | 197 | 0 | 15.2 | 3.1 | 0 | 71.6 | 10.2 | 0 |
| `mstep` | *Multi-Table*, *Multi-Step* Superlative (54.4%), Ranking (29.4%), Sum (2.6%), Average (13.6%) | 588 | 0 | 34.9 | 8.5 | 0 | 38.4 | 18.2 | 0 |

Table 2: Dataset size, question type frequency and distribution in the GRI topics.

organization and accessibility of ESG data.

To the best of our knowledge, GRI-QA is also the first question-answering benchmark for tabular data in the environmental domain and annotated with industry standards such as GRI.

## 3 The GRI-QA Benchmark

The GRI-QA benchmark consists of a total of 4089 questions spanning 204 tables extracted from corporate English reports[2] published in 2023 from companies in Germany (19), France (7), and Italy (4). The questions are logically organized into eight datasets based on their types, as shown in Table 2. In particular, we divide the question types into *extractive* and *calculated*. The *extractive* questions require the identification of relevant span(s) in a table. We distinguish between extractive questions that require to directly retrieve a value (dataset `extra`) and hierarchical questions (dataset `hier`) that involve tables where the row/column headers (e.g., total amounts) are broken down into their components, requiring an understanding of these relationships to answer. The *calculated* types of question require performing a computation over multiple cells. The computation can refer to identifying *relationships* between cells, such as *comparison*, *ranking*, and *superlative operations* or to generating *quantitative* results by applying *sum*, *average*, or *percentage* calculations. In GRI-QA, we also introduce *multi-step* questions whose resolution requires applying a combination of relational and quantitative operations to cells in the same or in different tables. We call `rel`, `quant`, and `step` the datasets with relational, quantitative, and multi-step questions on single tables, and `mrel`, `mquant`, and `mstep` their variants on multiple tables.

Finally, the questions are annotated with their GRI topic. GRI-QA focuses specifically on environmental topics (see Appendix B for a detailed description of the GRI 300 series), with their distribution across the datasets shown in Table 2. The distribution of topics across the datasets is unbalanced, reflecting their prevalence in the analyzed corporate reports. The *Emissions* 305 topic is the most common in the benchmark, while other GRI topics such as 301 *Materials* or 304 *Biodiversity* are less represented.

The construction of GRI-QA consists of two phases: retrieving the tables related to specified GRI topics in corporate reports (Section 3.1) and generating questions based on the extracted tables (Section 3.2).

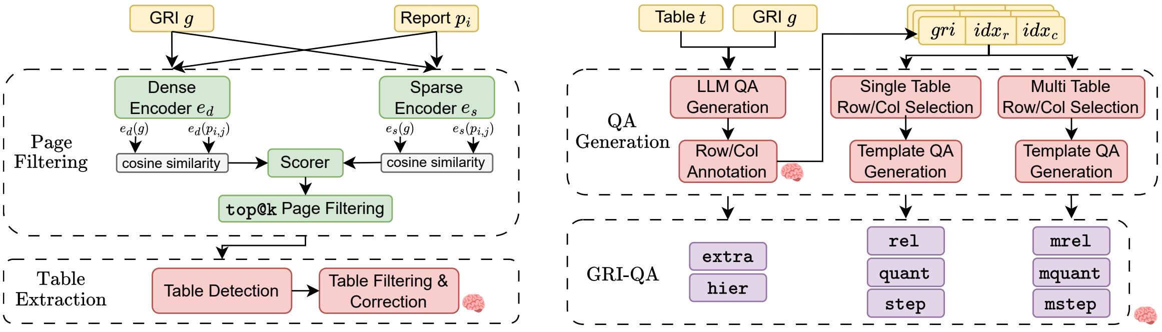
### 3.1 Phase 1: Table extraction

This phase retrieves and extracts relevant tables, i.e., those associated with the GRI topics of interest, from corporate documents. The process begins by selecting pages relevant to the target GRI topic $g$ where $g$ is a textual description taken from Table 8 (Page Filtering in Figure 2a), followed by extracting the tables they contain (Table Extraction).

**Page filtering.** Non-financial corporate reports are typically large documents (500+ pages) covering a wide range of topics. To reduce the search space, this component identifies the top $k$ sections related to a target GRI topic using an information retrieval method that combines sparse (syntactic) and dense (semantic) embeddings, a technique shown to be effective in the BEIR benchmark (Thakur et al., 2021). In particular, given a GRI topic description $g$ and a report page $p_{i,j}$, where $i$ denotes the reports and $j$ the page number, we compute their similarity score $s_t$ with

$$s_d(p_{i,j}, g) = sim(e_d(p_{i,j}), e_d(g))$$
$$s_s(p_{i,j}, g) = sim(e_s(p_{i,j}), e_s(g))$$
$$s_t(p_{i,j}, g) = s_d(p_{i,j}, g) + \lambda \cdot s_s(p_{i,j}, g)$$

[2] www.annualreports.com



(a) Phase 1: Table extraction

(b) Phase 2: Question generation

Figure 2: The two phases of our annotation pipeline. 🧠 boxes contain the operations performed by the annotators.

where $sim$ is the cosine similarity, $e_d$, $e_s$ are respectively a dense retriever[3] (Wang et al., 2024a) and a TF-IDF sparse retriever weighted by $\lambda$, and $s_d$, $s_s$ and $s_t$ are the *dense*, *sparse* and *total* score. If the page content $p_{i,j}$ exceeds the context window of $e_d$, we split $p_{i,j}$ into multiple chunks and repeat the process. Based on empirical evaluation during the annotation phase, we set $\lambda = 0.3$ and $k = 20$, which produced favorable results. This process is repeated for each corporate report and each GRI topic description. The final output identifies the pages containing environmental data along with their associated GRI topics.

**Table extraction & filtering.** We extract tables from the retrieved top $k$ relevant pages using the Unstructured[4] library, combined with Google's Tesseract OCR (Smith, 2007) to accurately recognize the characters in the cells. The tables are then manually corrected for structural and syntactical errors, with annotators verifying the coherence of assigned GRI topics. Among the total 204 clean tables, their dimensions range from 1 to 53 rows and 2 to 15 columns. Furthermore, 32.84% of tables contain at least one hierarchical row index, 17.16% have a hierarchical column index, and 20.59% include data related to multiple GRI topics.

### 3.2 Phase 2: Question generation

In Phase 2, we generate the questions and answers for the GRI-QA datasets from the extracted tables, as shown in Figure 2b. First, we use an LLM to automatically generate extractive questions for the extra and hier datasets. These questions are then reviewed by human annotators, who validate them and add supplementary annotations, such as the corresponding row and column indices. Then, the remaining datasets are created using a template-based approach. This method selects an operation (e.g., maximum, minimum, sum, percentage) and uses the annotated row and column indices to ensure the generated queries compare or combine values in a meaningful way, reflecting real-world scenarios. The specific process for each dataset is detailed in the following.

**Datasets extra, hier.** We use an LLM, specifically gpt-4o-mini with a temperature of 0.2, to generate extractive questions from the tables, ensuring contextual accuracy with minor variations in phrasing. The full prompt used is provided in Appendix A. Human annotators verify the quality of the questions, manually rephrase them for better variability and report the row and column indices $(idx_r, idx_c)$ for the extracted values. Finally, they identify which extra samples require reasoning over hierarchical row structures, contributing to the hier dataset.

**Datasets rel, quant, step.** These datasets contain questions based on operations requiring comparisons, calculations, or a combination of both. In particular, the rel dataset includes comparative, superlative, and ranking operations; the quant dataset includes sum, average, and percentage change; and the step dataset combines these operations. For question generation, similar to Pal et al. (2023), we use expert-defined templates to structure the questions. Each template is designed for a specific operation and requires access to both the input operands and the computed result. To ensure meaningful and reliable datasets, it is crucial that the selected operands are consistent. This consistency is guaranteed by the prior annotation conducted for the extra dataset, which links cells identified by the $(idx_r, idx_c)$ indices in the extracted tables to relevant GRI topics. Leveraging this annotation, we automate the generation of samples for the rel, quant, and step datasets through two al-

[3] intfloat/multilingual-e5-large-instruct

[4] Unstructured Documentation

ternative operand selection strategies. Given an initial operand, **row selection** identifies randomly a second operand from a different column within the same row. Conversely, **column selection** identifies a second operand from the same column but in a different row, ensuring that both values share the same GRI topic. The answer is computed by executing the operation and by also ensuring consistency in the compared GRI topics. To maintain quality, annotators review the generated question-answer pairs, validate their coherence, and manually rephrase questions to enhance linguistic variability and diversity across the dataset.

**Datasets `mrel`, `mquant`, `mstep`.** In the multi-table datasets, tables are extracted from different corporate reports to simulate realistic cross-company comparisons. A selection of 2, 3, or 5 tables is used to generate questions of varying difficulty. Consistency in table selection is ensured using the previously generated GRI associations. As with the `rel`, `quant`, and `step` datasets, a template-based approach is applied to the selected tables to generate relational, quantitative, and multi-step questions for the multi-table `mrel`, `mquant`, and `mstep` datasets. There are, however, two key differences. First, questions may require identifying a specific company name rather than computing a numerical value. Second, due to variations in company size, the same indicator may be presented in different units (e.g., GJ for small firms, GWh for large ones), necessitating implicit unit conversion for meaningful comparisons in some cases.

### 3.3 Quality Control

The annotation of GRI-QA requires a strong understanding of GRI reporting standards and meticulous accuracy in calculating target values. The manual annotation process was conducted by two ICT research fellows, under the guidance of a professor specialized in economic and management science expert in the topic. To ensure that the annotators were well prepared for the task, we provided 1,038 examples that included GRI topics, tables, extractive questions and answers, accompanied by concise documentation of the GRI standards. This preparation allowed the annotators to familiarize themselves with the annotation process and resolve any uncertainties related to domain-specific terminology. The annotation process comprised three distinct phases. Firstly, the annotators refined the structure and content of the automatically extracted tables, ensuring that no table contains information that uniquely identifies people. Secondly, they annotated the row and column indices for the `extra` dataset. Finally, they conducted a thorough review of the generated datasets, ensuring consistency between the questions and the answers while improving the variability of the questions.

## 4 Experimental Evaluation

We assess the quality of GRI-QA using Table QA models trained on both general, and financial-domain datasets. In the follow, we present the baselines (Section 4.1), metrics (Section 4.2), and discuss the results obtained (Section 4.3).

### 4.1 Baselines

As baselines, we select and evaluate state-of-the-art models developed and trained both in general domains and in financial domains close to the ones addressed by GRI-QA. We do not include existing models trained on environmental data (e.g., ESGLlama), as they are not designed for tabular processing, making their assessment unfair.

**General models.** We experiment with four models that have not been specifically developed or trained in environmental and financial domains: (i) *TaPEx* (Liu et al., 2022), a 406M parameters BART-based (Lewis et al., 2020) model trained to predict SQL query results and fine-tuned on textual and tabular inputs; (ii) *OmniTab* (Jiang et al., 2022), a 406M parameters model which incorporates pre-training on synthetic data and demonstrated strong performance on the WTQ dataset (see Table 1); (iii) *TableLLAMA* (Zhang et al., 2024a), a Llama2 (Touvron et al., 2023) 7B model fine-tuned on the Table-Instruct (Zhang et al., 2024a) dataset, achieving state-of-the-art performance on HiTab and FeTaQa (see Table 1); (iv) `gpt-4o-mini` in Zero-Shot (Radford et al., 2019) and Zero-Shot-CoT (Kojima et al., 2024) with temperature of 0.

**Financial-related models.** We consider two models trained on financial data extracted from corporate reports: (i) *FinMA* (Xie et al., 2023), a Llama2 model fine-tuned on several financial tasks, including Table QA; (ii) *TaT-LLM* (Zhu et al., 2024), a Llama2 model fine-tuned on TAT-QA, TAT-DQA and FinQA (Table 1). We evaluate FinMA and TaT-LLM with 7B parameters, using the *step-wise* prompt for TaT-LLM, as indicated by its authors.

**Human-level performance baseline.** We asked three expert users (two professors and one FinTech expert) to answer 400 randomly extracted questions, i.e., 50 from each GRI-QA dataset. The experts had unlimited time to respond to the questions, which were preloaded into an instance of the Label Studio platform (see Appendix C for additional information). In Appendix D, we report the results obtained by the baselines on the human-evaluated samples from the datasets.

### 4.2 Metrics

We use a normalized Exact Match (EM) metric to evaluate the performance of the baselines. We adapted the DROP (Dua et al., 2019) evaluation script to compare the results with the ground truth, considering the specific output formats of each baseline. For instance, TaT-LLM separates list values using a hashtag (e.g., 16.5#16.1), TableLLAMA formats values within angle brackets (e.g., <16.5>,<16.1>). To ensure consistency, we apply custom post-processing operations for each model, normalizing their outputs before comparison. The post-processing operations are limited to handling model-specific output formats and do not modify the predicted values or their units of measurement, which are explicitly required by the question prompts. This normalization allows us to use the models within their intended prompt settings, minimizing the need for significant prompt modifications that could negatively impact performance (Kojima et al., 2024).

### 4.3 Discussion

The analysis of the experimental results in Table 3, broken down into questions requiring a single table (Table 3a) or multiple tables (Table 3b), provides the following insights.

*Humans* make the difference. The accuracy achieved by human annotators surpasses all models by a significant margin, except in the `rel` dataset. This exception is likely due to the fact that answering `rel` questions can be "mechanical" for a human, leading to overconfidence. The performance gap between humans and computational models is relatively small for single-step questions (7.1% on average) but increases substantially for `step` and multi-step datasets (37.2% on average), where multiple calculations across different documents are required to obtain the correct answer. Finally, humans exhibit a lower accuracy variance across datasets, with a standard deviation of 6.0 in single-table questions and 1.7 in multi-table questions.

**Key Takeaway #1.** The significant gap between human and computational performance highlights the need for further research, especially in multi-step and multi-table reasoning tasks.

*GPT-based* models outperform all others. Models based on `gpt-4o-mini`, particularly when using `CoT` prompting, achieve significantly higher accuracy than other models. Focusing exclusively on the results related to single table queries, where the models produce the best performances, there is a difference of about 10% and 30% between the GPT base and CoT models respectively with respect to the best performance for the remaining models: 54% GPT base and 73% GPT CoT versus 41% for TaT-LLM. TaT-LLM emerges as the best non-GPT model, followed by TableLLAMA with an average accuracy of 31.1%. OmniTab and TaPEx achieve slightly lower performances at 29.3% and 26%, respectively, while FinMA records the lowest accuracy at 17.3%. The superiority of GPT models is also confirmed in the multi-table question datasets.

**Key Takeaway #2.** GPT models outperform state-of-the-art tabular QA systems, even specialized in the financial domain. CoT prompting further enhances these performances.

*Financial* training does not always generate better results. Excluding GPT-based models and multi-table datasets, we observe that the average accuracy of general models (28.8%, standard deviation 2.6) is similar to that of financial-related models (29.3%, standard deviation 16.9). However, the financial-based model TaT-LLM consistently outperforms TableLLAMA, its general-domain counterpart based on the same Llama2 LLM. Meanwhile, the second financial-based model in our benchmark, FinMA, achieves the highest non-GPT accuracy on the `rel` dataset but underperforms compared to general models on the other datasets. This variability, also reflected in the higher standard deviation of financial-based models compared to general ones (16.9 vs. 2.6), suggests that while specialized models can excel in certain tasks, general models tend to be more robust across datasets.

**Key Takeaway #3.** Financial-based models can outperform general models but exhibit greater variability depending on the dataset.

|  | extra | hier | rel | quant | step | avg |
|---|---|---|---|---|---|---|
| TaPEx | 55.4 | 46.4 | 27.4 | 1.0 | 0.0 | 26.0 (25.4) |
| OmniTab | 64.7 | 55.4 | 25.4 | 1.1 | 0.0 | 29.3 (30.0) |
| TableLLAMA | 73.1 | 63.3 | 17.7 | 1.5 | 0.0 | 31.1 (34.7) |
| FinMA | 25.7 | 22.9 | 35.1 | 2.6 | 0.0 | 17.3 (15.3) |
| TaT-LLM | 79.7 | 74.3 | 25.4 | 26.7 | 0.0 | 41.2 (34.4) |
| gpt-4o-mini | **86.0** | 78.5 | 61.7 | 43.2 | 0.7 | 54.0 (34.1) |
| gpt-4o-mini CoT | 84.2 | **80.9** | **92.7** | **72.6** | **33.1** | **72.7** (23.3) |
| Dataset avg | 67.0 | 60.2 | 40.8 | 21.2 | 4.8 | 38.8 (26.1) |
| Human† | 89.3 | 92.0 | 87.3 | 90.0 | 76.7 | 87.1 (6.0) |

(a) Single-table questions.

|  | mrel | mquant | mstep | avg |
|---|---|---|---|---|
| TaPEx | 0.3 | 0.0 | 0.2 | 0.2 (0.2) |
| OmniTab | 1.1 | 0.0 | 0.0 | 0.4 (0.6) |
| TableLLAMA | 4.7 | 1.0 | 9.5 | 5.1 (4.3) |
| FinMA | 1.8 | 0.0 | 0.7 | 0.8 (0.9) |
| TaT-LLM | 5.5 | 0.0 | 8.4 | 4.6 (4.3) |
| gpt-4o-mini | 16.1 | 1.5 | 13.6 | 10.4 (7.8) |
| gpt-4o-mini CoT | **37.2** | **29.9** | **35.2** | **34.1** (3.8) |
| Dataset avg | 9.5 | 4.6 | 9.8 | 8.0 (2.9) |
| Human† | 70 | 70.3 | 67.3 | 69.2 (1.7) |

(b) Multi-table questions.

Table 3: Accuracy (EM score) and standard deviation (in brackets) for the GRI-QA benchmark. **Bold** values indicate the best results. † indicates the results obtained from 50 randomly extracted samples.

|  | **(%) Error for each question type** |
|---|---|
| rel | Comparison (4.0%), Superlative (10.5%), Ranking (8.6%) |
| quant | Sum (22.8%), Average (23.3%), Percentage change (34.1%) |
| step | *Multi-step* Superlative (61.0%), Ranking (71.9%), Sum (64.7%), Average (73.1%) |
| mrel | *Multi-Table* Superlative (51.2%), Ranking (70.3%) |
| mquant | *Multi-Table* Sum (73.0%), Average (74.0%) |
| mstep | *Multi-Table*, *Multi-Step* Superlative (55.2%), Ranking (77.2%), Sum (80.0%), Average (80.0%) |

Table 4: Percentage of gpt-4o-mini CoT errors per question type.

*Not* all operations are equally complex. Even within single-table datasets, which result easier for the models in the benchmark, accuracy varies significantly between datasets. The performance on hier is constantly lower than in extra, ranging from 60.2% to 67% on average, indicating that reasoning on hierarchical rows is a difficult task for computational models (Katsis et al., 2022), while humans generalize better. Accuracy on *calculated* questions is even worse, with quantitative questions appearing more complex than relational ones, both in single- and multi-table datasets. Multi-step questions finally show the worst performances, with GPT models reaching around 30% accuracy only when using CoT prompting.

**Key Takeaway #4.** Hierarchical and quantitative questions are moderately challenging, while multi-step questions pose the highest complexity. CoT's "divide and conquer" approach helps improve multi-step reasoning.

*More* documents, more complexity. Accuracy drops significantly when operands come from different documents. This is evident when comparing the results on *calculated* operation datasets rel, quant, and step with their multi-document counterparts, mrel, mquant, and mstep, where operands are drawn from tables in different documents. Table 4 shows the error breakdown for the types of questions in the datasets for the gpt-4o-mini CoT model. *Ranking* questions appear to be the most affected by the multi-table setting, showing a 61.7% increase in error rate between rel and mrel, while the increase for the superlative questions is less marked (40.7%, from 10.5% to 51.2%). *Sum* and *average* questions in the quant dataset behave similarly, with both showing an increase in error rates of about 67%. The mstep dataset may exhibit an anomaly, achieving an overall higher accuracy than step. This can be explained by the presence of questions expecting company names as answers instead of numerical values, which are easier for baseline models to handle. When excluding these questions, the accuracy of gpt-4o-mini CoT on mstep drops to 23.5%.

**Key Takeaway #5.** Accuracy drops when operands span multiple documents.

*Vocabulary* and units of measurement matter. Models struggle when domain-specific terminology is involved, even in tasks that do not require complex reasoning. To isolate the effect of vocabulary and avoid confounds related to structural complexity, we analyzed the errors on the extra dataset designed for direct value retrieval. Moreover, we considered only questions referring to tables where rows represent indicators and columns contain values across years. This ensures the analysis is centered on query intent interpretation and term disambiguation, removing complexities related to table format understanding. The amount of

| Model | Errors | Errors (%) | T1 | T2 | $\frac{(T1+T2)}{Errors}$ |
|---|---|---|---|---|---|
| `TaPEx` | 452 | 67.36% | 66 | 298 | 80.53% |
| `OmniTab` | 334 | 63.02% | 33 | 203 | 70.66% |
| `TableLLAMA` | 254 | 62.87% | 19 | 163 | 71.65% |
| `FinMA` | 800 | 71.62% | 32 | 531 | 70.38% |
| `TaT-LLM` | 198 | 64.92% | 8 | 165 | 87.37% |
| `gpt-4o-mini` | 144 | 68.25% | 5 | 135 | 97.22% |
| `gpt-4o-mini CoT` | 177 | 74.06% | 1 | 173 | 98.31% |

Table 5: Types of error on the `extra` dataset (T1, T2). Term misinterpretations cause the majority of errors.

errors considered for each model is indicated in the "Errors" columns in Table 5. Despite the simplicity of the task, the results reveal that a significant proportion of errors are due to terminology interpretation issues. We categorize errors into three types: (T1) retrieving a value from an unrelated row (indicative of misunderstanding the indicator name), (T2) retrieving a value not present in the table (often due to hallucinated terms or misinterpretation), and (T3) retrieving the correct row but from the wrong column (e.g., incorrect year). Only the first two types reflect vocabulary-related issues. As shown in Table 5, these two categories account for over 70% of all errors across all models. with models like TaT-LLM and gpt-4o-mini reaching 87% and 97%. These values indicate that the models are proficient in identifying simple indicators (e.g. the years in the column schema), but provide hallucinated responses due to term ambiguity and domain-specific vocabulary (e.g. terms related to GRI topics). In addition to vocabulary, models also falter when faced with inconsistent or implicit units of measurement across tables. To assess this, we analyzed the first 50 questions from the multi-table, multi-step setting involving five tables, where unit conversion is often required. For each sample, we manually reviewed the Chain-of-Thought reasonings of `gpt-4o-mini` and annotated the errors related to unit handling. We identified 38 total errors, of which 22 (57.9%) involved incorrect or missing unit conversions.

**Key Takeaway #6.** Domain-specific terminology and unit conversions remain major obstacles, even in seemingly simple retrieval or reasoning tasks.

## 5 Conclusion

We presented GRI-QA, a new single- and multi-table question answering benchmark on environmental data. GRI-QA is composed of eight datasets that focus on different types of questions, providing a new challenging test bed to assess the quality of Table QA models. Furthermore, GRI-QA provides a set of questions on multiple non-relational tables belonging to different corporate reports, a setting only partially explored in previous works. The results show that while current models are proficient in *extractive* questions, they fail in *calculated* questions, which require performing computations over multiple cells. This gap is further increased in multi-step and multi-table questions, where the only model obtaining non-negligible accuracy is `gpt-4o-mini CoT`. We made the datasets and the annotation pipeline publicly available, to promote and support further research in the area.

## 6 Limitations

The paper does not present a new custom baseline model capable of addressing GRI-QA. We motivate this decision by the fact that (i) GRI-QA is supposed to provide a test benchmark rather than training data, and (ii) the number of samples would likely be insufficient to fine-tune existing models and be competitive with larger foundation models such as `gpt-4o-mini`. However, it would still be interesting to leverage GRI-QA, or its data collection pipeline, to improve the performance of small LLMs (e.g. 7B models) or define new prompting techniques to improve foundation LLMs. Moreover, although we tried to make the dataset creation pipeline as automated as possible, a lot of human effort is still needed. As a result, while GRI-QA contains a significant number of questions, it is limited in the number of corporate reports considered. We plan to address these limitations in future work.

OpenAI models accessed via API calls are known to provide non-deterministic outputs even when setting the temperature to 0. This behavior is aggravated in Chain-of-Thought prompting, where the selection of different tokens may lead to different reasoning paths and outputs. As a consequence, the Chain-of-Thought results shown in Table 3 and Figure 7 may slightly differ between different runs.

## 7 Risks

A potential risk with the use of GRI-QA is the growing focus on models that maximize accuracy, while disregarding computational effort and energy consumption (see Appendix D). Although these models can help environmental practitioners extract relevant information from corporate reports, they can also contribute to environmental impact.

Another issue concerns the sources of corporate reports considered in this study. Our paper considers only French, German and Italian companies with reports written in English (see Section 3). As a result, GRI-QA can disadvantage practitioners analyzing companies located in other geographical regions, as well as stakeholders relying on different languages.

## 8 Use of AI assistants

When writing this paper, we used AI assistants, such as ChatGPT and Writefull, 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. Furthermore, we used `gpt-4o-mini` to help us debug our code.

## 9 Acknowledgments

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

Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Transformers for tabular data representation: A survey of models and applications. *Trans. Assoc. Comput. Linguistics*, 11:227–249.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115, Macao. -.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021a. Open question answering over tables and text. In *International Conference on Learning Representations*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

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. 2021b. 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.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *EMNLP*, pages 6279–6292. Association for Computational Linguistics.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. Hitab: A hierarchical table dataset for question answering and natural language generation. In *ACL (1)*, pages 1094–1110. 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.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

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.

European Commission. 2024. A european green deal - european commission.

GRI. 2024. Global reporting initiative website.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.

Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. OmniTab: Pretraining with natural and synthetic data for few-shot table-based question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 932–942, Seattle, United States. Association for Computational Linguistics.

Yannis Katsis, Saneem A. Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael R. Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2022. AIT-QA: question answering dataset over complex tables in the airline industry. In *NAACL-HLT (Industry Papers)*, pages 305–314. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin Chen-Chuan Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can LLM already serve as A database interface? A big bench for large-scale database grounded text-to-sqls. In *NeurIPS*.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. TAPEX: Table pre-training via learning a neural SQL executor. In *International Conference on Learning Representations*.

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

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).

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

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).

Timothy Nugent, Nicole Stelea, and Jochen L. Leidner. 2021. Detecting environmental, social and governance (ESG) topics using domain-specific language models and data augmentation. In *FQAS*, volume 12871 of *Lecture Notes in Computer Science*, pages 157–169. Springer.

Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. MultiTabQA: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6322–6334, Toronto, Canada. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Mariya Pavlova, Bernard Casey, and Miaosen Wang. 2024. ESG-FTSE: A corpus of news articles with ESG relevance labels and use cases. 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 137–149, Torino, Italia. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. In *ACL (Short Papers)*, pages 445–458. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.

SDGS. 2024. Sustainable development goals.

Ray 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. IEEE.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims. *CoRR*, abs/2209.00507.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA. Association for Computing Machinery.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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.

Francesco S. Varini, Jordan L. Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. *CoRR*, abs/2012.00483.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024b. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations*.

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

Jian Wu, Linyi Yang, Dongyuan Li, Yuliang Ji, Manabu Okumura, and Yue Zhang. 2025. MMQA: Evaluating LLMs with multi-table multi-hop complex questions. In *The Thirteenth International Conference on Learning Representations*.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. 2024. Tablebench: A comprehensive and complex benchmark for table question answering. *CoRR*, abs/2408.09174.

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.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018*

*Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024a. TableLlama: Towards open large generalist models for tables. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6024–6044, Mexico City, Mexico. Association for Computational Linguistics.

Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, Jifan Yu, Shu Zhao, Juanzi Li, and Jie Tang. 2024b. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *CoRR*, abs/2403.19318.

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.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

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.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 310–318, New York, NY, USA. Association for Computing Machinery.

## A Examples of prompts

Examples of prompts are shown for: (1) the creation of the extra dataset (Figure 3); (2) generating the tests shown in Table 3a and Table 3b (Figure 4); (3) generating the tests shown in Table 3a and Table 3b using the chain-of-thought technique (Figure 5).

## B GRI 300 topics and disclosures

Table 8 shows the descriptions of the GRI categories, topics and disclosures. GRI-QA focuses on tables related to topics from category 300.

## C Labeling Interface

Figure 6 shows the Label Studio interface used to obtain the human results shown in Table 3 and Table 6. The text inside Figure 6 are the guidelines we provided to the annotators.

## D Supplementary results

Table 6a and Table 6b directly compare the results obtained by three human annotators on 50 samples extracted from each dataset of GRI-QA, with the results obtained using the baseline models on the complete datasets. To ensure a fair comparison, in Table 6 we re-evaluate the baseline models on the dataset samples, showing similar results to the ones obtained in Table 3a and Table 3b.

Figure 7 shows the accuracy breakdown in multi-table datasets as the number of tables increases.

Figure 7 shows the energy consumption of each model tested. Even though TaT-LLM, FinMA and Table-LLAMA share the same Llama2 backbone, TaT-LLM leads to higher energy consumption due to its *step-wise* prompting. The results for `gpt-4o-mini`, in both Zero-Shot and Zero-Shot CoT prompting, are omitted as the energy consumption of API calls cannot be measured.

**extra generation prompt**

You will be given a table (in HTML) and a GRI description.

Extract the values that can reply to the given description. Then, select one of the values and generate an unambiguous question that can be replied by the value. By unambiguous question we mean that, based on the table, the question can have only one answer. Hence, make sure to include the correct column header in the question.

As output, provide a Python dictionary that has the questions as keys and the extracted values are the values of the dictionary. Do not write anything else. Do not provide any Markdown formatting.

Table: {table}
GRI description: {description}
Question:

Figure 3: Prompt to generate the samples of `extra`.

**GPT prompt template**

You must answer the following question given the provided table.

If the question is boolean, write exclusively a ’yes’ or ’no’ answer. If the question asks for a list of values, separate them with a comma. Write the numerical values with exactly 2 decimal values. Do not write anything else. Do not write any Markdown formatting.

Question: {question}
Table: {table}
Answer:

Figure 4: GPT prompt template used for the tests in Table 3a and Table 3b. Tables are provided in their HTML representation.

**GPT CoT prompt template**

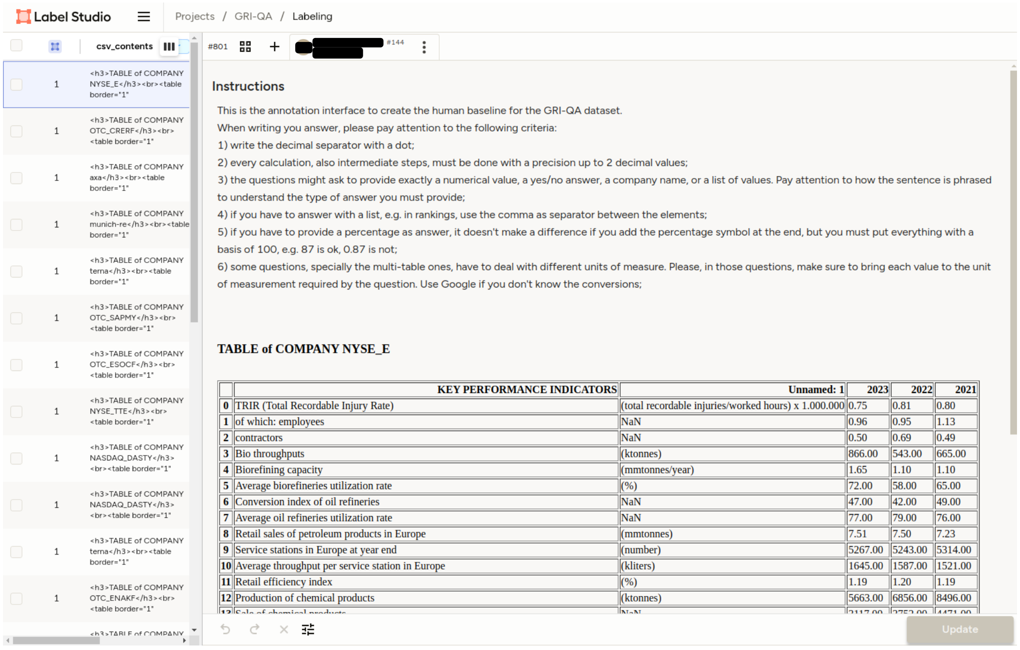
You must answer the following question given the provided table.

First write your reasoning. Then, in the end, write "The final answer is:" followed by the answer. If the question is boolean, write exclusively a ’yes’ or ’no’ answer. If the question asks for a list of values, you must answer with a list of values separated with a comma. Write the numerical values with exactly 2 decimal values. Do not write any Markdown formatting.

Question: {question}
Table: {table}

Let’s think step-by-step.

Figure 5: GPT CoT prompt template used for the tests in Table 3a and Table 3b. Tables are provided in their HTML representation.



| | | csv_contents |
|---|---|---|
| □ | 1 | <h3>TABLE of COMPANY NYSE_E</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY OTC_CRERF</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY axa</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY munich-re</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY terna</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY OTC_SAPMY</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY OTC_ESOCF</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY NYSE_TTE</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY NASDAQ_DASTY</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY NASDAQ_DASTY</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY terna</h3><br><table border="1" |
| □ | 1 | <h3>TABLE of COMPANY OTC_ENAKF</h3><br><table border="1" |

Instructions

This is the annotation interface to create the human baseline for the GRI-QA dataset.
When writing you answer, please pay attention to the following criteria:
1) write the decimal separator with a dot;
2) every calculation, also intermediate steps, must be done with a precision up to 2 decimal values;
3) the questions might ask to provide exactly a numerical value, a yes/no answer, a company name, or a list of values. Pay attention to how the sentence is phrased to understand the type of answer you must provide;
4) if you have to answer with a list, e.g. in rankings, use the comma as separator between the elements;
5) if you have to provide a percentage as answer, it doesn't make a difference if you add the percentage symbol at the end, but you must put everything with a basis of 100, e.g. 87 is ok, 0.87 is not;
6) some questions, specially the multi-table ones, have to deal with different units of measure. Please, in those questions, make sure to bring each value to the unit of measurement required by the question. Use Google if you don't know the conversions;

**TABLE of COMPANY NYSE_E**

| | KEY PERFORMANCE INDICATORS | Unnamed: 1 | 2023 | 2022 | 2021 |
|---|---|---|---|---|---|
| 0 | TRIR (Total Recordable Injury Rate) | (total recordable injuries/worked hours) x 1.000.000 | 0.75 | 0.81 | 0.80 |
| 1 | of which: employees | NaN | 0.96 | 0.95 | 1.13 |
| 2 | contractors | NaN | 0.50 | 0.69 | 0.49 |
| 3 | Bio throughputs | (ktonnes) | 866.00 | 543.00 | 665.00 |
| 4 | Biorefining capacity | (mmtonnes/year) | 1.65 | 1.10 | 1.10 |
| 5 | Average biorefineries utilization rate | (%) | 72.00 | 58.00 | 65.00 |
| 6 | Conversion index of oil refineries | NaN | 47.00 | 42.00 | 49.00 |
| 7 | Average oil refineries utilization rate | NaN | 77.00 | 79.00 | 76.00 |
| 8 | Retail sales of petroleum products in Europe | (mmtonnes) | 7.51 | 7.50 | 7.23 |
| 9 | Service stations in Europe at year end | (number) | 5267.00 | 5243.00 | 5314.00 |
| 10 | Average throughput per service station in Europe | (kliters) | 1645.00 | 1587.00 | 1521.00 |
| 11 | Retail efficiency index | (%) | 1.19 | 1.20 | 1.19 |
| 12 | Production of chemical products | (ktonnes) | 5663.00 | 6856.00 | 8496.00 |

Figure 6: Label Studio annotation interface (www.labelstud.io), along with the textual guidelines provided to the annotators to answer the questions.

| | GRI-QA | | | | | avg |
|---|---|---|---|---|---|---|
| | `extra` | `hier` | `rel` | `quant` | `step` | |
| TaPEx | 54 | 36 | 32 | 2 | 0 | 24.8 (23.3) |
| OmniTab | 52 | 48 | 28 | 0 | 0 | 25.6 (25.1) |
| TableLLAMA | 74 | 54 | 22 | 2 | 0 | 30.4 (32.6) |
| FinMA | 22 | 26 | 22 | 4 | 0 | 14.8 (11.9) |
| TaT-LLM | 86 | 70 | 24 | 25 | 0 | 41.0 (35.7) |
| gpt-4o-mini | **90** | **80** | 56 | 44 | 2 | 54.4 (34.6) |
| gpt-4o-mini CoT | 84 | **80** | **98** | **70** | **31** | **72.6** (25.3) |
| Human | 89.3 | 92.0 | 87.3 | 90.0 | 76.7 | 87.1 (6.0) |

(a) Single-table questions.

| | GRI-QA | | | avg |
|---|---|---|---|---|
| | `mrel` | `mquant` | `mstep` | |
| TaPEx | 0 | 0 | 0 | 0.0 (0.0) |
| OmniTab | 0 | 0 | 2 | 0.7 (1.2) |
| TableLLAMA | 6 | 0 | 0 | 2.0 (3.5) |
| FinMA | 2 | 0 | 0 | 0.7 (1.2) |
| TaT-LLM | 6 | 0 | 20 | 8.7 (10.3) |
| gpt-4o-mini | 14 | 2 | 20 | 12.0 (9.2) |
| gpt-4o-mini CoT | **40** | **30** | **38** | **36.0** (5.3) |
| Human | 70 | 70.3 | 67.3 | 69.2 (1.7) |

(b) Multi-table questions.

Table 6: Accuracy (EM score) on 50 samples extracted from each GRI-QA dataset.

| | GRI-QA | | | avg |
|---|---|---|---|---|
| | `mrel` | `mquant` | `mstep` | |
| *2 tables* | 56.6 | 58.7 | 43.7 | 53.0 |
| *3 tables* | 34.3 | 20.8 | 32.7 | 29.3 |
| *5 tables* | 19.5 | 0.0 | 25.5 | 15.0 |

Table 7: EM scores of `gpt-4o-mini` CoT for multi-table questions in the `mrel`, `mquant` and `mstep` datasets.

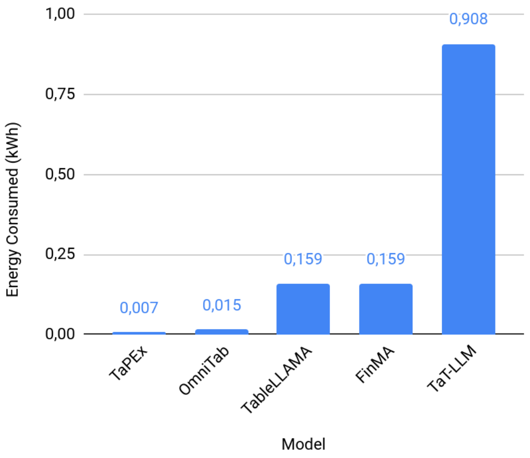


Figure 7: Total energy consumption (kWh) of each model on the `extra` dataset, measured using an NVIDIA L40S.

| **Topics** | **Disclosures** | **Descriptions** |
|---|---|---|
| 301<br>(*Materials*) | 301-1 | Materials used by weight or volume |
| | 301-2 | Recycled input materials used |
| | 301-3 | Reclaimed products and their packaging materials |
| 302<br>(*Energy*) | 302-1 | Energy consumption within the organization |
| | 302-2 | Energy consumption outside the organization |
| | 302-3 | Energy intensity |
| | 302-4 | Reduction of energy consumption |
| | 302-5 | Reductions in energy requirements of products and services |
| 303<br>(*Water and Effluents*) | 303-1 | Interactions with water as a shared resource |
| | 303-2 | Management of water discharge-related impacts |
| | 303-3 | Water withdrawal |
| | 303-4 | Water discharge |
| | 303-5 | Water consumption |
| 304<br>(*Biodiversity*) | 304-1 | Operational sites owned, leased, managed in, or adjacent to, protected areas and areas of high biodiversity value outside protected areas |
| | 304-2 | Significant impacts of activities, products and services on biodiversity |
| | 304-3 | Habitats protected or restored |
| | 304-4 | IUCN Red List species and national conservation list species with habitats in areas affected by operations |
| 305<br>(*Emissions*) | 305-1 | Direct (Scope 1) GHG emissions |
| | 305-2 | Energy indirect (Scope 2) GHG emissions |
| | 305-3 | Other indirect (Scope 3) GHG emissions |
| | 305-4 | GHG emissions intensity |
| | 305-5 | Reduction of GHG emissions |
| | 305-6 | Emissions of ozone-depleting substances (ODS) |
| | 305-7 | Nitrogen oxides (NOx), sulfur oxides (SOx), and other significant air emissions |
| 306<br>(*Waste*) | 306-1 | Waste generation and significant waste-related impacts |
| | 306-2 | Management of significant waste-related impacts |
| | 306-3 | Waste generated |
| | 306-4 | Waste diverted from disposal |
| | 306-5 | Waste directed to disposal |
| 308<br>(*Supplier Environmental Assessment*) | 308-1 | New suppliers that were screened using environmental criteria |
| | 308-2 | Negative environmental impacts in the supply chain and actions taken |

Table 8: Summary of GRI 300 topics and disclosures