DEUTSCHE BUNDESBANK
EUROSYSTEM

Addressing Data Gaps in Climate Change and Sustainability Statistics using Generative AI

Technical Report 2026-05

Maurice Fehr[1]
Simon Oehler[1]
Lisa Reichenbach[1]
Marvin Riemer[1]
Anna Steinberg Schulten[2]
Malte Schierholz[2]
Andreas Dimmelmeier[2]
Felicitas Sommer[3]

[1]Deutsche Bundesbank, Research Data and Service Centre
[2]Ludwig-Maximilians-Universität München (LMU Munich)
[3]Technische Universität München

Disclaimer:

The views expressed in this technical report are personal views of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

Research Data and Service Centre

Abstract

Climate change has emerged as a critical macroeconomic and financial risk. Effective policy responses require high-quality, granular sustainability data - in particular issuer-level greenhouse gas (GHG) emissions - which remain scarce, inconsistent, and difficult to extract from corporate disclosures. This paper presents the "Greenhouse Gas Insights and Sustainability Tracking" (GIST) project, a collaboration between Ludwig-Maximilians-University Munich and Deutsche Bundesbank, which leverages state-of-the-art generative artificial intelligence to address these data gaps. We introduce a Retrieval-Augmented Generation (RAG) pipeline for extracting Scope 1–3 GHG emissions from corporate reports, converting unstructured disclosures into standardized, analysis-ready datasets. The pipeline is validated against a rigorously constructed gold standard dataset involving both non-expert and expert review. Results demonstrate that large language models (LLMs) achieve high extraction accuracy (precision: 0.78; recall: 0.90), though challenges persist in handling ambiguous or fragmented disclosures. We further discuss the development of an open-source Python package enabling broad adoption and customization of the extraction workflow. Real-world applications within Deutsche Bundesbank highlight the pipeline's potential to enhance data quality, reduce manual labor, and support evidence-based policy making.

Acknowledgements: We would like to express our sincere gratitude to the BIS Innovation Hub Eurosystem Centre (Project GAIA) and the Deutsche Bundesbank Directorate General Markets (ICAS) for their support and fruitful exchanges throughout the course of this project.

Use of AI Assistant

Generative AI tools (GPT-based) were used during the preparation of this paper to support language editing, improve clarity, and assist with the structuring of the text. All AI-generated or AI-assisted content was critically reviewed and revised by the authors. The authors remain fully responsible for the content, accuracy, interpretation, and conclusions presented in this paper. Generative AI was not used to generate original scientific findings or replace the authors' academic judgment.

Contact

Deutsche Bundesbank, Research Data and Service Centre (RDSC)
Homepage: https://www.bundesbank.de/en/statistics/sustainability/greenhouse-gas-insights-and-sustainability-tracking-gist--992500

Keywords: Data gaps, sustainable finance data, climate data, sustainability reports, natural language processing, large language models

Citation: Fehr, M., Oehler, S., Reichenbach, L., Riemer, M., Steinberg Schulten, A., Schierholz, M., Dimmelmeier, A., and Sommer, F. (2026). Addressing Data Gaps in Climate Change and Sustainability Statistics using Generative AI, Technical Report 2026-05. Deutsche Bundesbank, Research Data and Service Centre

Contents

1 Introduction **5**

2 The GIST project – contributing to closing data gaps **6**
2.1 Problem statement 6
2.2 GIST – project set-up and goals 7
2.3 RAG-based extraction pipeline 7
2.4 Development of a gold-standard dataset 9

3 The GIST-project: results and deliverables **11**
3.1 Gold-standard dataset 11
3.2 Experiments using gold standard validation data 11
3.3 Typology of data quality issues in reports 14
3.4 The climatextract Python package 14
3.5 Real world use case in collaboration with Markets Department at Deutsche Bundesbank 15
3.6 Improving comparability through unit normalization 16

4 Challenges and opportunities ahead **18**

References **19**

A Appendix **22**
A.1 Prompt templates 22

1 Introduction

Climate change has evolved from a predominantly environmental concern into a macroeconomic and financial risk. Recent literature has underscored the potentially severe macroeconomic consequences of climate change through physical risks – stemming from the increasing frequency and severity of extreme weather events – and transition risks, which arise from the structural adjustments required to move toward a low-carbon economy (Bilal & Stock, 2026).

These risks have far-reaching implications for core concerns of central banks including inflation, output, employment, asset prices, financial stability, and long-term economic growth (Boldrini, Ceglar, Lelli, Parisi, & Heemskerk, 2023; Dafermos, 2022; Dikau, Robins, & Volz, 2020; Faccia, Parker, & Stracca, 2021; Mandel, Battiston, & Monasterolo, 2025). To integrate these aspects into both supervisory activities and monetary policy, the Network for Greening the Financial System (NGFS) has thus argued that central banks must develop a comprehensive understanding of the transmission channels through which climate change affects the economy at the micro- and macroeconomic level (Alogoskoufis et al., 2021; Avgousti et al., 2023; Network for Greening the Financial System, 2026).

As climate change is typically not directly referenced in central banks' mandates, with only 12% of 135 mandates compiled by Dikau & Volz (2021) containing "explicit" references to the promotion of sustainability, the creation of the capacities advocated by the NGFS (see also Bolton, Despres, da Silva, Svartzman, & Samama, 2020) has been accompanied by a legal debate about the role of central banks in governing climate-related matters. Apart from highlighting the impacts of climate change on core monetary and supervisory matters, contributors to this debate have highlighted that central banks with a "secondary" objective to support economic policies such as the ECB have an - albeit often vaguely framed - duty to promote sustainable development if this does not conflict with the primary mandate (on the nuances of the legal argument see Elderson, 2021; de Boer & van't Klooster, 2021; Dikau & Volz, 2021). In addition, central banks' own balance sheets are directly exposed to climate risks, reinforcing the need to systematically incorporate such risks into analytical frameworks, risk management practices, and policy operations (Deutsche Bundesbank, 2026).

Moving from the institutional and legal questions to matters of implementation, central banks and supervisory authorities have started to integrate climate-related information such as Greenhouse Gas (GHG) emissions into instruments like stress testing (European Banking Authority, 2024) and collateral frameworks (European Central Bank, 2026). Developing and implementing such instruments is contingent on the availability of high-quality data at the issuer level. In this context, the lack of structured, complete datasets on emissions data at the issuer level can lead to serious gaps and high data collection efforts by analysts (Doll & Alves Werb, 2023; Seltmann, Kormanyos, & Doll, 2025).

It is against this backdrop that this paper investigates how generative artificial intelligence (AI) can help address these data deficiencies, with a particular focus on improving the availability and quality of issuer-level greenhouse gas emissions data. In doing so, we highlight that while modern AI methods enable the extraction and processing of unstructured data at scale, substantial challenges remain in terms of validation, standardization, and post-processing.

2 The GIST project – contributing to closing data gaps

2.1 Problem statement

Even though many companies publish their greenhouse gas emissions in their annual reporting or through dedicated sustainability reports, aggregating this information into a structured database remains challenging. In order to use granular emissions data for analytical purposes, the data has to be searched for in the individual reports, usually available on the companies' websites. Several use cases in central banks' work such as climate stress testing, climate reporting and economic research require quantitative analysis and therefore an amount of company-level climate data for which regular manual extraction from individual reports would not be feasible. Therefore, data from third-party providers, which compile, structure and clean company disclosures, are often used. The ESCB, for example, has licensed two data vendors for emissions data to account for provider-specific strengths, weaknesses and coverages.

Matching emission data from these vendors, we compiled a dataset where each observation represents a unique issuer in a single year that is included in both data sets. Looking at scope 1 emissions (GHG emissions released from operations that are under direct control of the company) for example, Figure 1 (left) below shows that there are noticeable differences between vendor's data. Figure 1 (right) shows that the variation between data providers is strongly reduced if only reported data (instead of data that is estimated by the data providers) is plotted, underlining the importance of increasing the availability of reported emissions values. It also shows, however, that large differences remain in some cases despite the data vendors both stating that the data is based on reports, showing the necessity of a transparent mechanism for translating corporate reports into usable data sets. This aligns with earlier findings from 2021, using different data vendors (Fehr, Gies, Schlitzer, Triebskorn, & Wengerek, 2021).

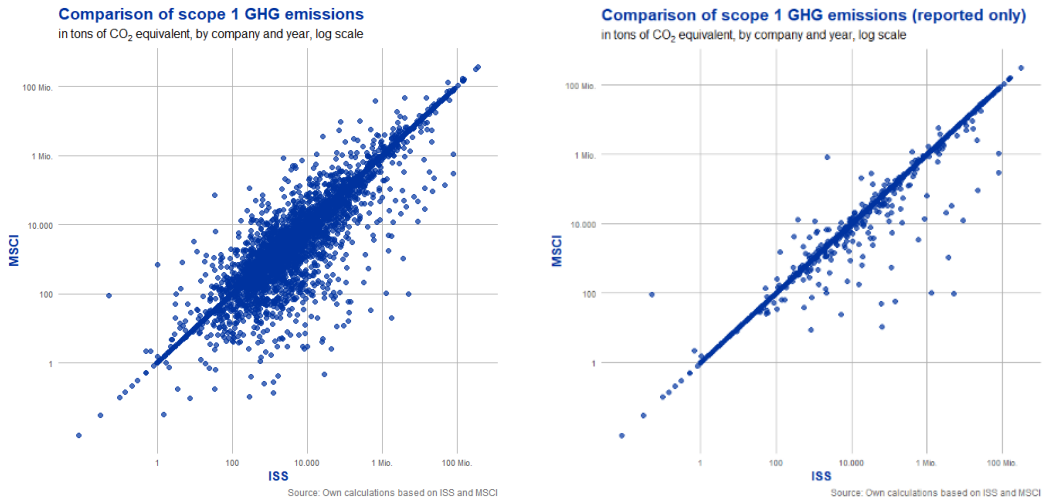


Figure 1: Comparison of scope 1 emissions data provided by ISS ESG and MSCI.

Busch, Johnson, & Pioch (2020) and Berg, Kölbel, & Rigobon (2022) also show that this is not just an issue for the data vendors compared by the Bundesbank but that there are inconsistencies between different data vendors in general and that these depend on scope (information on indirect emissions is less consistent than on direct emissions) and source (data from reports is more consistent than if estimated by the data vendors).

A central limitation from a data user's perspective is the lack of transparency regarding the underlying reasons for discrepancies between numbers, as well as their exact source reports. Furthermore, commercial data vendors do not always provide the necessary data coverage as they usually focus on public companies which suits many of the users who tend to be investors, while smaller companies can also play an important role e.g. in bank's loan books. Finally, there are costs attached to external data licenses and, more importantly, also licensing restrictions when it comes to data sharing in the context of cooperation projects with universities, supervisors, ministries etc.

2.2 GIST – project set-up and goals

The GIST-project is a cooperation between Ludwig-Maximilians University (LMU) Munich and Deutsche Bundesbank initiated in 2023 (Dimmelmeier et al., 2024). One of its main goals is to leverage state-of-the-art AI tools for extracting sustainability-related metrics from company reports and provide such information as structured, analysis-ready tables (Doll, Kormanyos, Walter, & Werb, 2024). Generally, the project's outputs comprise a) software and tools that are used for extraction, preparation and analysis of data, b) high-quality benchmarking data (gold-standard) c) extractions from company reports - currently Scope 1-3 GHG emissions (expressed as CO2 equivalents). This way, the GIST project provides a flexible tool that enables different stakeholders to efficiently extract data from a variety of formats and tailored to their individual requirements. Despite LLMs general ability to automatically extract data from text, review and quality assurance remain important tasks that still require human input and can be labour–intense. In this context, the cooperation between LMU and Bundesbank has proven especially fruitful, e.g. in developing an efficient pipeline set-up and methodologies for post-processing raw data extractions to solve reoccurring data quality problems.

2.3 RAG-based extraction pipeline

We implemented a Retrieval-Augmented Generation (RAG) pipeline to extract firm-level climate-related information from sustainability and annual reports in PDF format. In the current application, the pipeline focuses on greenhouse gas (GHG) emissions that are disclosed in the Scope 1, Scope 2, and Scope 3 accounting categories. The objective is to convert unstructured or semi-structured disclosure text into a standardized table containing, for each report, the reporting year, emission scope, numerical value, unit, and provenance information.

Since corporate reports are often long while relevant emissions disclosures usually appear on only a small number of pages, the pipeline first narrows each document to the most relevant context before applying an LLM. The implementation is asynchronous and supports concurrent processing of multiple reports and LLM calls, subject to API rate limits, making it suitable for scaling to large corpora. The pipeline consists of the following four stages, illustrated in Figure 2.

Document embedding. Each PDF is processed page by page. The textual content of every page is extracted and embedded using a text embedding model. The resulting page embeddings are stored in a persistent embedding repository, which allows the system to reuse embeddings across runs and reduces processing time and costs when evaluating different prompts, models, or retrieval settings.

Semantic retrieval. Relevant pages are identified through semantic search. A fixed query, "What are the total CO2 emissions in different years? Include Scope 1, Scope 2, and Scope 3 emissions if available.", is embedded and compared with the page embeddings using vector similarity. The most relevant pages are selected according to configurable retrieval parameters, such as the maximum number of pages, similarity thresholds, and optional context windows. This retrieval step substantially reduces the amount of text passed to the language model while preserving the pages most likely to contain the desired emissions disclosures.

LLM-based page-level extraction. The selected pages are passed to an LLM for information extraction. The pipeline supports both text-only input and a text-plus-table mode, in which tables on relevant pages are extracted and converted into a structured textual representation. Separate table treatment can be useful because emissions are often disclosed in tables, where preserving layout helps avoid assigning values to the wrong year or scope. Each page is processed independently rather than concatenating all retrieved pages into a single prompt. This enables parallel processing, preserves page-level provenance, and reduces the risk that relevant values are missed in long contexts. We consider two prompt variants (which we present in Appendix A.1): a simple question-style prompt, which is parsed using regular expressions, and a structured prompt, following best practices, which defines the model's role, provides extraction rules, and requests structured output that can be parsed via JSON. In both cases, the model is instructed to extract only relevant absolute emissions values and to return at least the year, scope, value, and unit.

Post-processing. The LLM output is parsed according to the prompt type, validated, and transformed into standardized tables. For all documents analysed in a given pipeline run, the pipeline produces two main consolidated outputs: a detailed technical document overview table and an emissions table containing the structured extracted observations together with document and page identifiers. For the emissions table, post-processing involves deduplicating extracted values, normalizing reported units through a unit dictionary, and converting values, where possible, into a standardized measure expressed in tonnes of CO2 equivalent. Auxiliary outputs, such as invalid or unparsable responses, are stored separately for diagnostic purposes. If available, a benchmark dataset such as the one described in Section 2.4 can be used to evaluate the post-processed emissions table. Since extracted records retain identifiers, predictions can be matched to benchmark entries at the report-year-scope level after unit normalization. This allows standard metrics such as precision, recall, and F1-score to be computed and supports targeted error analysis, for example for missed values, incorrect year or scope assignments, or unit conversion issues. This evaluation step is optional and not required for running the pipeline.

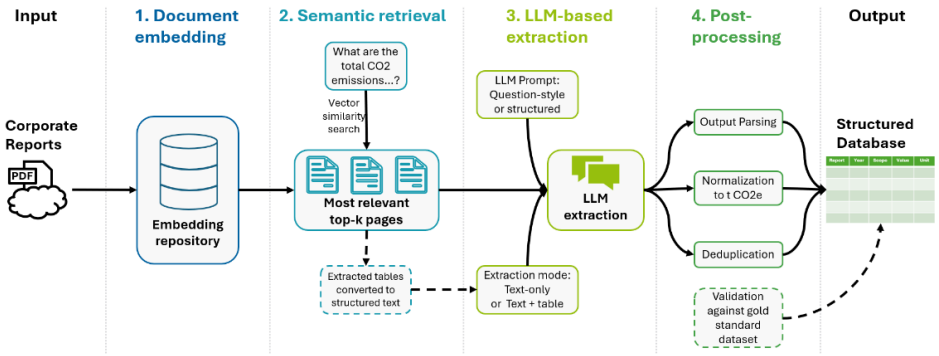


Figure 2: Overview of the RAG-based pipeline for extracting disclosed GHG emissions from corporate reports. The workflow narrows PDF reports to relevant pages using page-level embeddings and semantic retrieval, applies LLM-based extraction, and post-processes the results into a structured database. Dashed elements indicate optional components.

2.4 Development of a gold-standard dataset

To evaluate and compare various extraction approaches, a gold-standard dataset (Beck et al., 2025) was developed. Figure 3 visualizes the data collection process. First, Deutsche Bundesbank data scientists selected around 4000 companies from the international MSCI World Small Cap and German DAX indices and collected annual as well as separate stand-alone sustainability, Environment, Social & Governance (ESG) or Corporate Social Responsibility (CSR) reports, collectively referred to as sustainability reports. A substantial portion of the PDF files was either incomplete or did not conform to the definition of a sustainability report and was therefore discarded.

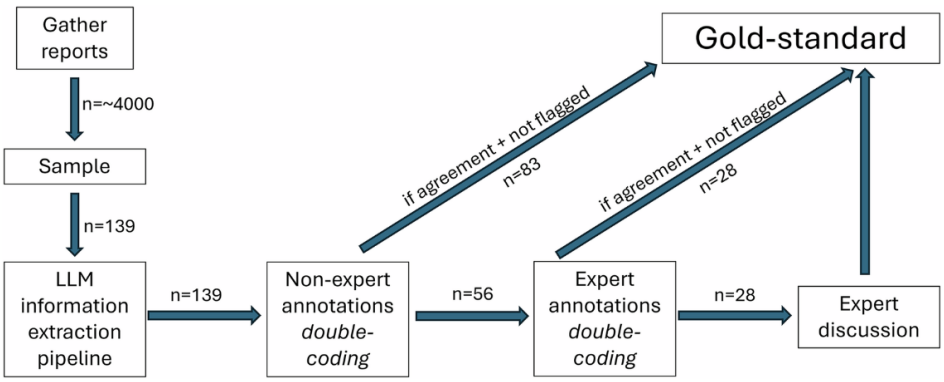


Figure 3: Visualisation of the data collection process from report collection, sampling, LLM extraction to three-stage annotation procedure (Source: Beck et al. 2025).

From these reports, published between 2017 and 2022, a random sample of 139 reports was drawn, stratified by the companies' first letters to match the distribution of the full report set. Seven companies had by chance two of their reports included in the sample. As a result, the gold standard represents a diverse group of 132 companies varying in size, sectors and regions.

An earlier version of the LLM-based information extraction pipeline described in Section 2.3 processed this sample. The output dataframes of emissions data, identified by a year-scope-page combination, form the basis for a three-stage procedure: non-expert annotation, expert annotation, and a final round of expert discussion. Annotation follows a custom convention designed

to keep results comparable across reports, companies, and annotators. Under this convention, an emission value counts as correct if it:

– covers emissions for the entire company,
– is reported according to the operational boundaries of the scope,
– is reported in absolute terms as CO2 or CO2-equivalent emissions, and
– represents a total value rather than a subcategory.

Two non-expert annotators first independently review and correct the LLM-extracted emission values. Reports where they agree on all entries go straight into the gold standard. Reports with disagreements move to two pairs of expert annotators; agreement between both pairs likewise sends those entries through. Annotators at either stage can flag unclear records for further review. Any remaining disagreements get resolved through in-person discussion, with the final entries added to complete the dataset.

3 The GIST-project: results and deliverables

3.1 Gold-standard dataset

The resulting gold standard dataset contains 5646 data records where one record describes a year-scope combination with a value and a unit for a specific report, identified by a company name and a reporting year. If the same scope-year combination appears twice in the report on different pages, the page number identifies the specific data record.

Scope	Average reportings per report
1	3.33
2 lb	2.94
2 mb	1.23
3	1.71

Table 1: Mean number of reported emission values per report by scope type.

In line with Forster et al. (2026), we find that many companies do not comprehensively disclose their carbon emissions: around half of the reports (69 of 139) contain no GHG emission values, the other 70 reports contain 9.2 values on average, and two reports top out at 32 scope-year combinations each. This pattern mirrors the broader disclosure hierarchy documented by Forster et al. (2026) across 600 European firms: Scope 1 and Scope 2 emissions are consistently among the most frequently reported indicators, while coverage drops off sharply for less standardized topics like pollution and biodiversity. Table 1 reflects this same hierarchy, with Scope 1 emerging as the emission type reported most often in our data, meaning that a report on average contains about three (different) scope 1 values.

Generally, tables hold the large majority of emission values, at 80%; graphics account for around 10%, and plain text for close to 4%. For the remaining 6%, non-experts disagreed on the display type, so the gold standard retains both proposed display types for these records.

3.2 Experiments using gold standard validation data

To assess how design choices affect extraction quality, we systematically varied three components of our RAG-based extraction pipeline: the model, the prompt structure, and the context type. We tested four closed-source models (GPT-4o, GPT-4.1, GPT-5, GPT-5.2) and two open-weight models (gpt-oss-120b, Llama-4), each under two prompt designs: a simple question-style prompt and a structured prompt inspired by best-practice recommendations (role, definitions, extraction rules, JSON output schema) and with two context types (text only, or text with tables converted to Markdown format) given the prevalence of emission values in tables. We fixed the number of retrieved pages at k=7 with a context window size of 0, the best-performing configuration identified in earlier experiments with GPT-4o-mini that swept 12+ combinations of k and context window size. Full details from these experiments are expected to be published in the EMNLP Industry Track 2026.

Best configuration. GPT-5.2 with the question-style prompt and additional tabular context achieved the strongest performance, reaching an average F1 score of 0.835 over 5 runs. Earlier

and smaller models consistently traded higher recall for lower precision, for example, GPT-5 achieved the highest overall recall (0.950) but at markedly lower precision (0.644) than GPT-5.2. The two open-weight models we tested (gpt-oss-120b and Llama-4) lagged noticeably behind the closed-source GPT family. Prompt and context design also mattered less than expected: a structured prompt was outperformed by the simpler question-style prompt. An ablation of the structured prompt's components indicates that extraction rules carry almost all of its value, while removing the role description or scope definitions trades a small gain in precision for a slight loss in recall. Adding tabular context to the retrieved text further improved the question prompt, but had the opposite effect on the structured prompt, decreasing both precision and recall. The best configuration of question-style prompt and text and tabular context used 1684756 prompt tokens (includes prompt and context (up to 7 pages)), 839233 output tokens and cost 14.70€ on average for 139 reports.

Prompt type	Input mode	Precision	Recall	F1 score
Question	Text	**0.779 (0.015)**	0.878 (0.084)	0.823 (0.039)
Question	Text + table	0.777 (0.021)	**0.903 (0.014)**	**0.835 (0.017)**
Structured	Text	0.760 (0.008)	0.860 (0.011)	0.807 (0.006)
Structured	Text + table	0.755 (0.020)	0.830 (0.105)	0.787 (0.049)

Table 2: Averages and standard deviations of 5 runs of prompt and context experiments with gpt-5.2.

Error analysis. For a more detailed analysis, we consider the best run with gpt-5.2 with the question-style prompt and the text context. The confusion matrix at the value level shows a marked asymmetry between the two directions of error:

	Gold: value present	Gold: no value / different value
LLM extracts value	483	121
LLM extracts no value	50	4916

Table 3: Confusion matrix with rows indicating the model's extraction (value vs. NA) and columns indicating ground truth (value present vs. no value/different value present) for the best run with gpt-5.2 and question-style prompt and text context.

The pipeline is generally conservative: it correctly abstains from generating values for 4916 of the 5037 cases where no value exists in the gold standard, but it misses 50 disclosed values entirely (false negatives) and extracts 121 unwarranted or false ones (false positives). In relative terms, the model is much more likely to over-report than to fabricate: the false positives are almost always cases where the model extracted a real number from the retrieved page, but not the correct one (or NA) for the requested scope-year combination. This is reflected at the report level in Figure 4: of 70 reports with disclosed emissions, 68 reports were successfully retrieved (for 2 reports at least one page containing values was not identified as relevant); of these, 36 (51%) were completely correctly extracted, 24 partially correctly, and 8 completely incorrectly. In contrast, of the 68 reports without disclosed emissions, 63 (91%) were correctly returned as empty, and only 6 produced incorrect values. The extraction task is thus substantially harder when disclosures exist than when they do not.

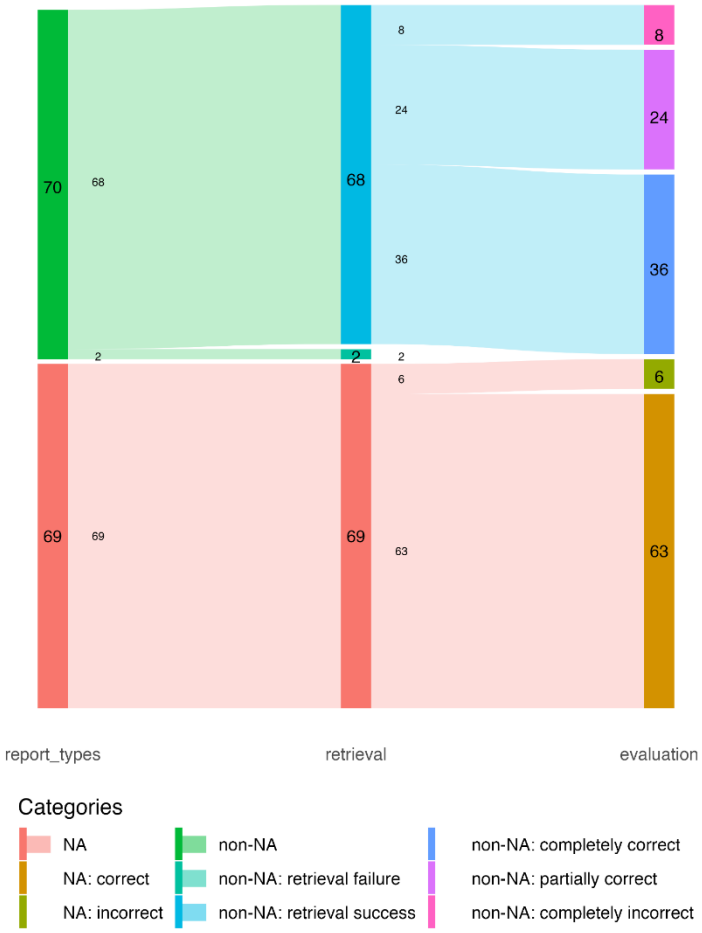


Figure 4: Report-level extraction outcomes across the 139 reports in the gold standard for the best run with question-style prompt and text context.

A manual inspection of the failing extractions reveals three qualitative patterns that explain most of the false positives and partial matches. First, several reports show value–scope or value–year misalignment: the correct emission value is extracted but attributed to the wrong scope or wrong reporting year. Second, extractions which are considered irrelevant according to the annotation criteria: values for subcategories of Scope 3 instead of total, values that cover only a subsidiary or region, values that are sums or intensity ratios. Third, and less frequently, the model interprets footnote numerals as part of a value, e.g., appending a footnote reference "1" to an emissions figure. These patterns are consistent with a model that competently identifies emissions-related passages but struggles to disambiguate among multiple candidate numbers within a passage. Analysis of the additional tabular context, motivated by the misalignment pattern, showed, however, that it resolved misalignment in only 2 reports, at least partially.

Easy vs. difficult reports. Reports with clean, tabular disclosures on a single page are extracted almost perfectly, regardless of prompt and context choice. Reports that are especially difficult tend to exhibit at least one of these three features: (i) multiple candidate values on the same page (subcategories, intensity ratios, etc.) among which the model must select the target; (ii) disclosures fragmented across the report, where different pages report subsidiary-level or partial values rather

than a single consolidated total; and (iii) use of footnotes or ambiguous disclosures. A heuristic post-processing step in the pipeline retains one entry per scope-year combination via three rules (same-page duplicate filtering, target-unit preference, majority-page selection). Addressing this in-prompt, e.g., by instructing the model to filter irrelevant values through explicit extraction rules, may be more effective than heuristic post-processing.

3.3 Typology of data quality issues in reports

Besides report design and format features, data quality issues can influence extraction quality and annotation time. Annotators' comments on incomplete or inconsistent statements in the reports from GIST and from another LLM supported emission extraction project, run by the Swedish Klimatkollen (Climate Checker) platform, were used to create a typology of data errors and issues (Dimmelmeier, Sommer, & Palmquist, 2025). The typology builds on GHG Protocol principles regarding calculation methodologies and the disclosure of results (WBCSD & WRI, 2011) that are integrated in disclosure rules like the CSRD and thus mandatory for reporting companies.

For instance, companies can state the exclusion of certain operations or countries from the overall calculation of emissions in a footnote and hence violate the GHGP principle of "completeness". If the majority or a substantial share of emissions is likely to come from these operations, the statement is not trustworthy. Expert annotators would note this, while LLMs might not be trained to integrate this level of complexity. Other data quality errors and issues pertain to ambiguous boundary specifications (e.g., the inclusion or exclusion of joint ventures); unclear calculation methodologies (e.g. restatements are not disclosed transparently); and calculation errors (e.g. incorrect summations of emission sub-categories). Besides deviations from GHG Protocol principles, different extraction results could be caused by organizational changes like mergers and acquisitions that make "restatements" of prior disclosures necessary or because updates to methodologies require recalculations (Papadopoulos, 2022). Annotation guidelines and extraction rules that dictate how to handle these cases can improve data quality.

3.4 The climatextract Python package

To facilitate reuse beyond the GIST project, the extraction workflow described in Section 2.3 has been made available as climatextract, an open-source Python package published on GitHub and PyPI. The package provides an interface for applying the methodology to corporate reports, for example through a single extraction call on a PDF file, while allowing users to configure key pipeline components such as retrieval settings, model choice, prompts, and post-processing options. Access to an LLM is required.

In addition to the extraction functionality, climatextract includes supporting tools for unit normalization and evaluation. The benchmark dataset described in this paper is compatible with climatextract and available via Zenodo.

Comprehensive documentation covering installation, configuration, the public API, and the underlying research background is publicly available here: climatextract.

3.5 Real world use case in collaboration with Markets Department at Deutsche Bundesbank

The pipeline's ability to correctly extract sustainability metrics from company reports was tested in a collaboration with Bundesbank's in-house credit assessment unit (ICAS). The goal of this exercise was to benchmark LLM-extracted data with an additional source of quality-assured data and evaluate the overall quality of LLM-extracted data. The assessment was based on information provided by the ICAS team: a) a set of approximately 2,000 company reports from German groups covering the period from 2019 to 2024 and b) an accompanying dataset containing Scope 1, Scope 2, and Scope 3 emission values from 2019 to 2024 that were manually extracted from these reports.

More concretely, using specific prompts, the LLM-based pipeline (as described in Section 2.3) was tasked with extracting Scope 1–3 emission data at the company level from these reports for a period from 2013 to 2024. The LLM-outputs were then processed as follows a) cleaning missing values (NAs), b) standardizing units, and c) removing duplicate records, i.e. identical company-year combinations stemming from different reports with the value from the most recent report year being retained. After cleaning, LLM-extracted values were then matched with manually extracted data using a unique company identifier. The final dataset consisted of 2656 standardized unique company-year matches across all scopes and years. For these remaining records, differences between LLM-extracted and manually extracted emission values on the company-level were computed. These differences were subsequently aggregated over all company-year combinations to assess the aggregate quality of the LLM-based extractions relative to the manually curated benchmark data.

Figure 5 shows a scatterplot of manually vs. LLM extracted datapoints for a subsample of German groups (IFRS) from the full dataset. For a large fraction of entities identical values have been extracted by the manual as well the LLM method while values below the 45° line (dashed) the LLM extracted higher emission values than the manual procedure and vice versa.

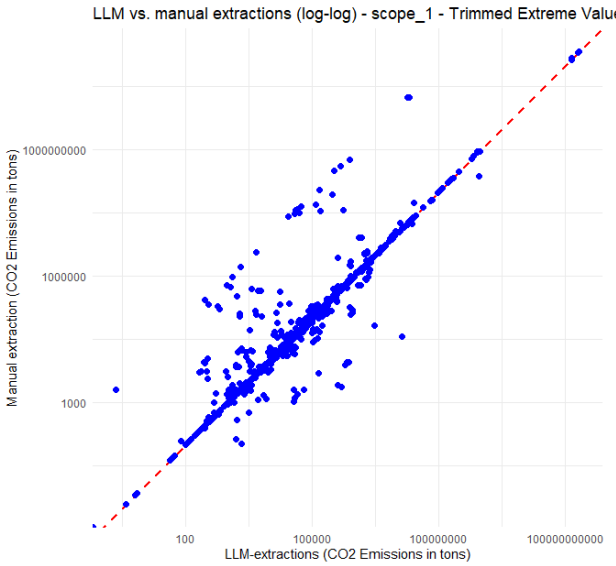


Figure 5: Comparison of manually and LLM-extracted Scope 1 emissions for German company reports (log-log scale). Most data points cluster around the 45° line, indicating high agreement between both extraction methods.

Figure 6 shows results after aggregation of all matching company year combinations (IFRS) by scope. The results indicate that deviations between manual and LLM extractions on aggregate level are relatively small - well below 10% across all scopes. Moreover, for this sample of company reports no bias towards one extraction method can be observed, i.e. none of the approaches seems to consistently over-/underestimate values relative to the other method.

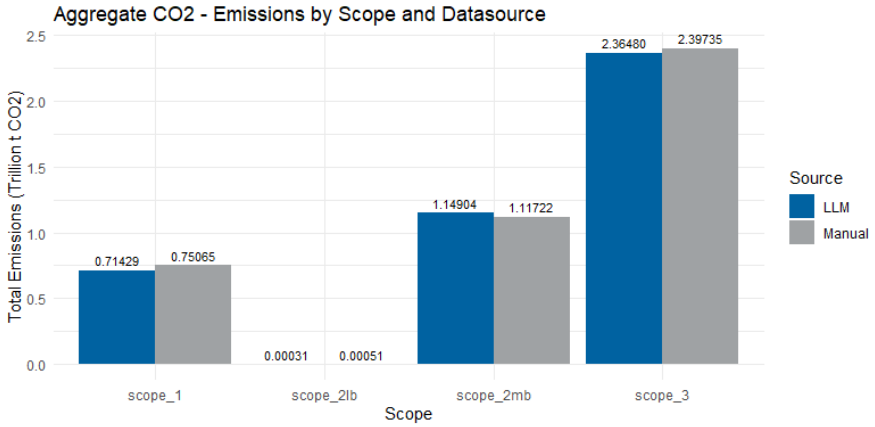


Figure 6: Comparison of aggregated CO_2 emissions by scope and data source. The results are very close for all scopes, showing that the LLM and manual methods produce similar totals.

The results demonstrate that the LLM-based extraction pipeline has significant potential to deliver value for domain experts in real-world applications. The pipeline provided an additional source of information that was instrumental in benchmarking manually extracted Scope 1–3 emission values. The matching of LLM-extracted data with manual extractions allowed for the classification of unambiguous cases, where errors were minimal or exact matches were observed. For the future, the ability to discriminate between certain and rather doubtful cases could potentially free up domain experts to focus on more complex cases e.g. comprising significant deviations in values. For example, LLM extractions could be used to predict which single-annotated reports have a high risk of potential annotation errors and thus may have to be revisited by domain experts (He & Schonlau, 2021). With respect to the number of documents to be analysed, this use-case also highlighted the potential of LLMs to support business units that are confronted with the analysis of large document sets. Given a specific requirement for data quality, such an approach could serve as a fast and efficient way to automate extraction tasks.

3.6 Improving comparability through unit normalization

A practical challenge in transforming LLM-extracted emission values into analysis-ready data is that extracted greenhouse gas emission units appear in highly heterogeneous forms across corporate reports. As the pipeline is applied to larger and more diverse report collections, previously unseen unit strings continue to appear. The dictionary-based normalization currently used in the pipeline is reliable for known variants but leaves many not-yet-observed unit variants unresolved. As illustrated in Table 4, raw unit strings differ not only in spelling and formatting but also in scale. If such units remain unresolved, the corresponding emission values cannot be reliably converted into a common target unit, in our case tonnes of CO2 equivalent, and are therefore difficult to compare or aggregate. At the same time, these scale differences mean that any other normalization must be sufficiently accurate to avoid severe conversion errors.

Raw unit string	Normalized class	Factor
“Tonnes CO2e”	t CO2e	1
“mt CO2e”	t CO2e	1
“kg CO2 equivalent”	kg CO2e	0.001
“kilotonnes CO2e”	kt CO2e	1,000
“1,000 t CO2”	kt CO2e	1,000
“million t CO2e”	Mt CO2e	1,000,000
“US tons CO2e”	USt CO2e	0.907185
“kg CO2/kWh”	Other	–

Table 4: Examples of raw unit strings, normalized classes, and deterministic conversion factors to t CO_2e.

To address this limitation, we have started to investigate a hybrid machine-learning unit-normalization approach as an extension to the current post-processing stage. Known unit strings are first matched against a curated dictionary that maps raw expressions to normalized unit classes and conversion factors. Unit strings that cannot be resolved by dictionary lookup are then passed to a supervised fallback classifier, which predicts categories such as “t CO2e”, “kt CO2e” or “Mt CO2e”. The classifier is trained on the manually curated unit dictionary, complemented by manually validated LLM-generated variants for underrepresented classes, and uses character-level string features.

Preliminary experiments suggest that such a fallback mechanism can substantially improve coverage beyond a static dictionary. The strongest classifier achieves around 95% accuracy and weighted F1-score and remains stable on a separate test set of rare real-world unit strings. Remaining errors mainly involve ambiguous or incomplete unit strings. A detailed evaluation will be provided in a forthcoming companion paper on emission-unit harmonization.

4 Challenges and opportunities ahead

The outcomes of the GIST project presented in this paper enable central banks and other stakeholders to extract GHG emissions data from publicly available corporate sustainability reports. In this regard, currently three levels of uncertainty for data utility of extracted corporate information exist: (1) the varying accuracy of emissions information from corporate sustainability reports, (2) not fully transparent third-party data provider's approaches of extracting and editing GHG emissions information from corporate sustainability reports and (3) new emerging risk topics (e.g. biodiversity) not sufficiently represented in available datasets.

The availability of LLM-extractions, or manually curated gold-standard data allows for the additional benchmarking and evaluation of third-party data. Next important steps are the provision of LLM-extracted values for a larger number of companies or additional metrics that go beyond the current focus of GHG emissions, e.g. water or electricity usage at the company level. Surface and ground water provision are among the most relevant ecosystem services for non-financial corporations (Boldrini et al., 2023). Additionally, more granular data on the asset-level, e.g. factory locations, would be relevant for the measurement of biodiversity risk or transition risk in value chains. Country- or asset-level data from corporate reports can be matched with geospatial data to determine the exposure of companies' assets to transition risks relevant for forward-looking methodologies, such as climate stress tests (Campiglio, Daumas, Monnin, & Jagow, 2022).

In conclusion, the GIST project makes emissions data publicly available based on transparent methods and performance evaluation. It applies state-of-the-art AI tools for information extraction and provides a technical solution as long as no better structured reporting formats or centralized data repositories are available. Beyond closing data gaps, the GIST project outcomes underline the relevance and benefits of interinstitutional cooperations for more accurate corporate GHG emissions and sustainability data.

References

[1] Alogoskoufis, S., Dunz, N., Emambakhsh, T., Hennig, T., Kaijser, M., Kouratzoglou, C., … Salleo, C. (2021). *ECB economy-wide climate stress test: Methodology and results* (ECB Occasional Paper Series No. 281). European Central Bank. Retrieved from European Central Bank website: https://www.ecb.europa.eu/pub/pdf/scpops/ecb.op281~05a7735b1c.en.pdf

[2] Avgousti, A., Caprioli, F., Caracciolo, G., Cochard, M., Dallari, P., Delgado-Téllez, M., … Theofilakou, A. (2023). *The climate change challenge and fiscal instruments and policies in the EU* (ECB Occasional Paper Series No. 315). European Central Bank. Retrieved from European Central Bank website: https://www.ecb.europa.eu/pub/pdf/scpops/ecb.op315~c279c7c290.en.pdf

[3] Beck, J., Steinberg, A., Dimmelmeier, A., Domenech Burin, L., Kormanyos, E., Fehr, M., & Schierholz, M. (2025). Addressing data gaps in sustainability reporting: A benchmark dataset for greenhouse gas emission extraction. *Scientific Data*, *12*(1). https://doi.org/10.1038/s41597-025-05664-8

[4] Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, *26*(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

[5] Bilal, A., & Stock, J. H. (2026). A guide to macroeconomics and climate change. *NBER WORKING PAPER SERIES*. https://doi.org/10.3386/w33567

[6] Boldrini, S., Ceglar, A., Lelli, C., Parisi, L., & Heemskerk, I. (2023). *Living in a world of disappearing nature: Physical risk and the implications for financial stability* (ECB Occasional Paper Series No. 333). European Central Bank. https://doi.org/10.2866/670314

[7] Bolton, P., Despres, M., da Silva, L. A. P., Svartzman, R., & Samama, F. (2020). *The green swan: Central banking and financial stability in the age of climate change*. Bank for International Settlements. Retrieved from https://www.bis.org/publ/othp31.pdf

[8] Busch, T., Johnson, M., & Pioch, T. (2020). Corporate carbon performance data: Quo vadis? *Journal of Industrial Ecology*, *26*(1), 350–363. https://doi.org/10.1111/jiec.13008

[9] Campiglio, E., Daumas, L., Monnin, P., & Jagow, A. von. (2022). Climate□related risks in financial assets. *Journal of Economic Surveys*, *37*(3), 950–992. https://doi.org/10.1111/joes.12525

[10] Dafermos, Y. (2022). Climate change, central banking and financial supervision: Beyond the risk exposure approach. In *The future of central banking* (pp. 175–194). Edward Elgar Publishing. Retrieved from https://www.elgaronline.com/edcollchap/book/9781839100932/book-part-9781839100932-19.xml

[11] de Boer, N., & van't Klooster, J. (2021). *The ECB cannot ignore its secondary mandate*. Verfassungsblog, 25 November 2021. Retrieved from https://verfassungsblog.de/the-ecb-cannot-ignore-its-secondary-mandate/

[12] Deutsche Bundesbank. (2026). *Climate report*. Deutsche Bundesbank. Retrieved from Deutsche Bundesbank website: https://publikationen.bundesbank.de/publikationen-en/reports-studies/climate-reports/climate-related-disclosures-by-the-deutsche-bundesbank-2026-924006

[13] Dikau, S., Robins, N., & Volz, U. (2020). *A toolbox for sustainable crisis response measures for central banks and supervisors* [INSPIRE Briefing Paper]. Grantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science and SOAS Centre for Sustainable Finance. Retrieved from Grantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science and SOAS Centre for Sustainable Finance website: https://soas-repository.worktribe.com/output/381040

[14] Dikau, S., & Volz, U. (2021). Central bank mandates, sustainability objectives and the promotion of green finance. *Ecological Economics*, *184*, 107022. https://doi.org/10.1016/j.ecolecon.2021.107022

[15] Dimmelmeier, A., Doll, H., Schierholz, M., Kormanyos, E., Fehr, M., Ma, B., … Kreuter, F. (2024). Informing climate risk analysis using textual information - A research agenda. In D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, … M. Leippold (Eds.), *Proceedings of the 1st workshop on natural language processing meets climate change (ClimateNLP 2024)* (pp. 12–26). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.climatenlp-1.2

[16] Dimmelmeier, A., Sommer, F., & Palmquist, A. (2025). *Typology of data quality problems in the corporate reporting of GHG emissions*. Zenodo. https://doi.org/10.5281/zenodo.21702666

[17] Doll, H. C., & Alves Werb, G. (2023). Innovation for improving climate-related data—lessons learned from setting up a data hub. *AStA Wirtschafts- Und Sozialstatistisches Archiv*, *17*(3-4), 355–380. https://doi.org/10.1007/s11943-023-00326-w

[18] Doll, H. C., Kormanyos, E., Walter, S., & Werb, G. A. (2024). *The climate data iceberg – a depth of information to integrate* (IFC Bulletin No. 66). Bank for International Settlements. Retrieved from Bank for International Settlements website: https://www.bis.org/ifc/publ/ifcb66_39.pdf

[19] Elderson, F. (2021). *Greening monetary policy*. The ECB Blog, 13 February 2021. Retrieved from https://www.ecb.europa.eu/press/blog/date/2021/html/ecb.blog210213~7e26af8606.en.html

[20] European Banking Authority. (2024). *Annual report*. European Banking Authority. Retrieved from European Banking Authority website: https://www.eba.europa.eu/sites/default/files/2025-06/bee4e97f-91a9-43bd-abdb-bd774e0259bf/2024%20Annual%20Report.pdf

[21] European Central Bank. (2026). *ECB to extend use of climate factors in Eurosystem collateral framework to non-financial corporate credit claims*. ECB Press Release, 24 July 2026. Retrieved from https://www.ecb.europa.eu/press/pr/date/2026/html/ecb.pr260724_4~f082ce289d.en.html

[22] Faccia, D., Parker, M., & Stracca, L. (2021). Feeling the heat: Extreme temperatures and price stability. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3981219

[23] Fehr, M., Gies, K., Schlitzer, C., Triebskorn, E., & Wengerek, S. (2021). *Short and medium-term approaches to closing climate data gaps: German experiences and conclusions on the way forward*. Paper presented at the 9th IMF Statistical Forum: Measuring Climate Change: The Economic and Financial Dimensions. Retrieved from https://www.imf.org/-/media/files/conferences/2021/9th-stats-forum/32finalpaperelenatriebskorn.pdf

[24] Forster, K., Keil, L., Wagner, V., Müller, M. A., Sellhorn, T., & Feuerriegel, S. (2026). Assessing corporate sustainability with large language models: Evidence from europe. *Nature Communications*, *17*(1). https://doi.org/10.1038/s41467-026-75160-z

[25] He, Z., & Schonlau, M. (2021). A model-assisted approach for finding coding errors in manual coding of open-ended questions. *Journal of Survey Statistics and Methodology*, *10*(2), 365–376. https://doi.org/10.1093/jssam/smab022

[26] Mandel, A., Battiston, S., & Monasterolo, I. (2025). Mapping global financial risks under climate change. *Nature Climate Change*, *15*(3), 329–334. https://doi.org/10.1038/s41558-025-02244-x

[27] Network for Greening the Financial System. (2026). *Climate change and monetary policy strategy: A guide for central banks* [NGFS Technical document]. Network for Greening the Financial System. Retrieved from Network for Greening the Financial System website: https://www.ngfs.net/en/publications-and-statistics/publications/climate-change-and-monetary-policy-strategy-guide-central-banks

[28] Papadopoulos, G. (2022). *Discrepancies in corporate GHG emissions data and their impact on firm performance assessment* [JRC Publications Repository]. European Commission, Joint Research Centre. Retrieved from European Commission, Joint Research Centre website: https://publications.jrc.ec.europa.eu/repository/handle/JRC130254

[29] Seltmann, S., Kormanyos, E., & Doll, H. C. (2025). Building a retrieval-augmented generation pipeline to trace administrative data use in academic papers. In *Foundations and advances of machine learning in official statistics* (pp. 347–373). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-10004-7_14

[30] WBCSD, & WRI. (2011). *The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard: Revised Edition*. World Resources Institute; World Business Council for Sustainable Development. Retrieved from World Resources Institute; World Business Council for Sustainable Development website: https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf

A Appendix

A.1 Prompt templates

```
Extract key pieces of information from this sustainability report.
IMPORTANT: Do NOT use any markdown formatting like **bold**, *italic*, or headers.
Write CO2 (not CO₂) when specifying units.
If a particular piece of information is not present, output 'Not specified'.
If the report does not mention if Scope 2 is location-based or market-based,
assume it is Scope 2 location-based. Do not extract it as Scope 2 market-based.

Use the following format:
0. What is the title
1. What are the Scope 1 emissions in 2013
2. What are the Scope 1 emissions in 2014
3. What are the Scope 1 emissions in 2015
4. What are the Scope 1 emissions in 2016
5. What are the Scope 1 emissions in 2017
6. What are the Scope 1 emissions in 2018
7. What are the Scope 1 emissions in 2019
8. What are the Scope 1 emissions in 2020
9. What are the Scope 1 emissions in 2021
10. What are the Scope 1 emissions in 2022
11. What are the Scope 1 emissions in 2023
12. What are the Scope 2 (market-based) emissions in 2013
13. What are the Scope 2 (market-based) emissions in 2014
14. What are the Scope 2 (market-based) emissions in 2015
15. What are the Scope 2 (market-based) emissions in 2016
16. What are the Scope 2 (market-based) emissions in 2017
17. What are the Scope 2 (market-based) emissions in 2018
18. What are the Scope 2 (market-based) emissions in 2019
19. What are the Scope 2 (market-based) emissions in 2020
20. What are the Scope 2 (market-based) emissions in 2021
21. What are the Scope 2 (market-based) emissions in 2022
22. What are the Scope 2 (market-based) emissions in 2023
23. What are the Scope 2 (location-based) emissions in 2013
24. What are the Scope 2 (location-based) emissions in 2014
25. What are the Scope 2 (location-based) emissions in 2015
26. What are the Scope 2 (location-based) emissions in 2016
27. What are the Scope 2 (location-based) emissions in 2017
28. What are the Scope 2 (location-based) emissions in 2018
29. What are the Scope 2 (location-based) emissions in 2019
30. What are the Scope 2 (location-based) emissions in 2020
31. What are the Scope 2 (location-based) emissions in 2021
32. What are the Scope 2 (location-based) emissions in 2022
33. What are the Scope 2 (location-based) emissions in 2023
34. What are the Scope 3 emissions in 2013
35. What are the Scope 3 emissions in 2014
36. What are the Scope 3 emissions in 2015
37. What are the Scope 3 emissions in 2016
38. What are the Scope 3 emissions in 2017
39. What are the Scope 3 emissions in 2018
40. What are the Scope 3 emissions in 2019
41. What are the Scope 3 emissions in 2020
42. What are the Scope 3 emissions in 2021
43. What are the Scope 3 emissions in 2022
44. What are the Scope 3 emissions in 2023

Document:
{context_str}

0. What is the title: Our responsibility. Report {self.min_year}
1. What are the Scope 1 emissions in 2013: <value> <unit>
2. What are the Scope 1 emissions in 2014: <value> <unit>
```

Figure A1: Question-style prompt.

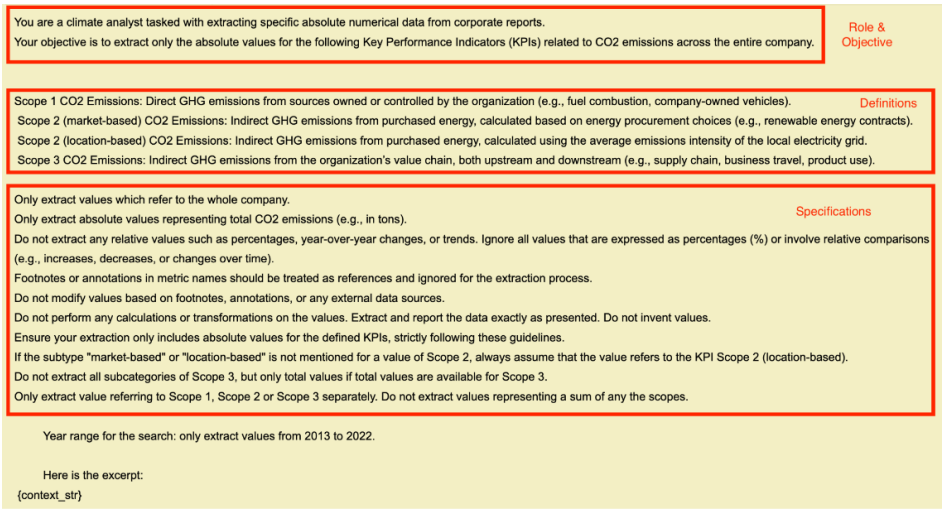


Figure A2: Structured prompt.