DEUTSCHE
BUNDESBANK
EUROSYSTEM

Informing climate risk analysis using textual information – A research agenda

Technical Report 2024-01

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

Andreas Dimmelmeier[1][2][3]
Hendrik Christian Doll[1][4]
Malte Schierholz[1][3]
Emily Kormanyos[4]
Maurice Fehr[4]
Bolei Ma[3][5]
Jacob Beck[3][5]
Alexander Fraser[5][6]
Frauke Kreuter[3][5][7]

Research Data and
Service Centre

Abstract

We present a research agenda focused on efficiently extracting, assuring quality, and consolidating textual company sustainability information to address urgent climate change decision-making needs. Starting from the goal to create integrated FAIR (Findable, Accessible, Inter-operable, Reusable) climate-related data, we identify research needs pertaining to the technical aspects of information extraction as well as to the design of the integrated sustainability datasets that we seek to compile. Regarding extraction, we leverage technological advancements, particularly in large language models (LLMs) and Retrieval-Augmented Generation (RAG) pipelines, to unlock the underutilized potential of unstructured textual information contained in corporate sustainability reports. In applying these techniques, we review key challenges, which include the retrieval and extraction of CO2 emission values from PDF documents, especially from unstructured tables and graphs therein, and the validation of automatically extracted data through comparisons with human-annotated values. We also review how existing use cases and practices in climate risk analytics relate to choices of what textual information should be extracted and how it could be linked to existing structured data.

Keywords: climate risk, transition risk, data gaps, RAG, large language models

DOI: NA

Citation: "Dimmelmeier, A., Doll, H.C., Schierholz, M., Kormanyos, E., Fehr, M., Ma, B., Beck, J,. Fraser, A. & Kreuter, F. (2024). Informing climate risk analysis using textual data – A research agenda. Deutsche Bundesbank Technical Report, 2024-01."

[1] These authors contributed equally.
[2] Corresponding author: A.Dimmelmeier@stat.uni-muenchen.de
[3] Ludwig-Maximilians-University Munich
[4] Deutsche Bundesbank, Data Service Centre
[5] Munich Center for Machine Learning (MCML)
[6] Technical University of Munich
[7] University of Maryland, College Park

1 Introduction

In light of the climate crisis, there is an increasing call to integrate climate risk with the decision-making of companies, banks and regulators. Climate risks for companies and, by extension, financial institutions have been grouped into two types: *transition risks* and *physical risks* (Carney, 2015). Transition risks arise from the transition of the economy towards carbon neutrality and can materialize, e.g., in the form of higher-than-expected carbon prices, stricter regulation, or changes in technology and consumer preferences. These risks affect companies and sectors with high (expected) carbon emissions. Physical risks, on the other hand, denote the direct adverse effects of a changing global climate, such as sea level rise or increases in storms and floods, droughts, and other natural disasters (IPCC, 2022). Unlike transition risks, physical risks do not depend primarily on companies' carbon footprint, but on the vulnerability of their assets and business operations to physical damage based on their geographic location.

Besides the companies themselves, climate risks are relevant to the financial institutions which are exposed to the affected companies through financial instruments such as loans or bonds. A bottleneck in climate risk analysis is the availability of reliable data (NGFS, 2022). Items that can help measure companies' physical or transition risk profiles, such as carbon emissions and transition plans, are scarcely available. As a consequence, institutions like the European System of Central Banks (ESCB) have thus far relied on proprietary datasets from private data providers (Deutsche Bundesbank, 2022). These commercial providers often source their climate risk data from corporate (sustainability) reports through manual annotation. Whenever reported data is not available or deemed insufficiently reliable, these data providers estimate numbers. Often, however, neither the reported nor the estimated data is replicable, since the providers do not disclose their estimation methods, and human annotators can be prone to errors. Despite recent regulatory efforts which have led to an uptick in company sustainability disclosures, the data is most often provided in relatively unstructured sustainability reports. Within these reports, important information is not usually presented in consistent and numeric formats (e.g., in structured tables), but can be presented in any form of text and even graphics.

Beyond corporate sustainability reports, unstructured textual sustainability information on climate risks is also available in the form of newspaper articles, social media comments, and other dispersed sources. The left panel of Figure 1 presents an overview of existing structured and unstructured sources of climate information. In this landscape, recent technological progress in natural language processing (NLP) opens up a range of new opportunities in efficiently extracting relevant data from unstructured textual information, which then can be linked to other data sources. Within the possible sources of textual information, companies' sustainability reports are arguably the most relevant document type for climate risk analysis since some form of sustainability disclosure tends to be mandatory. The information contained in such reports is mostly related to transition risks. This stems from the fact that, while sustainability reports could conceivably also include information on physical risks, the focus (beyond marketing considerations) usually lies on the companies' ecological footprint. Therefore, when referring to climate risks in the context of this paper, we focus on extracting information related to transition risks unless explicitly stated otherwise. For physical risks, unstructured information also exists largely in the form of images, e.g., satellite imagery or street view. In this domain, recent research also aims to convert unstructured information from images into usable data (Alonso-Robisco, Carbó, Kormanyos, and Triebskorn, 2024; Rossi, Byrne, and Christiaen, 2024). Our goal is thus to leverage sustainability reports in order to validate existing data sets, close data gaps by making new variables available, increase the coverage of company-level data, and improve the accessibility of information.

The remainder of this paper develops a research agenda that leverages NLP methods to condense the disparate sources of unstructured information into a structured, comprehensive, accessible, and trustworthy database. We develop this proposal across three sections: The first section discusses the latest research and use cases of NLP in the context of textual sustainability information in general and corporate sustainability reports in particular. The second section further explores the specifications and challenges related to LLM-based extraction pipelines by reporting the results from three initial experiments aimed at extracting emission values from 39 sustainability reports. The third section addresses the questions of (i) *how* data extraction should be organized, (ii) *what* information should be prioritized for extraction, and (iii) *how* data linkage and post-processing should be undertaken in order to create an integrated data infrastructure. The fourth section concludes the paper.

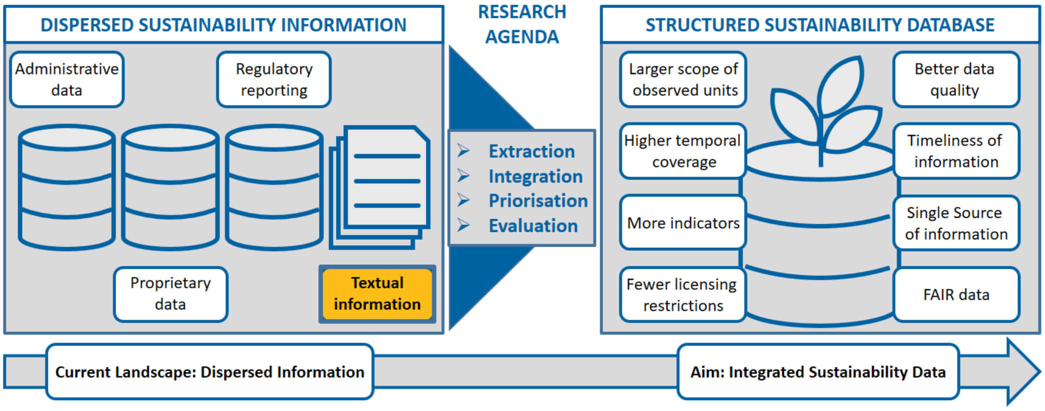
DISPERSED SUSTAINABILITY INFORMATION
Administrative data
Regulatory reporting
Proprietary data
Textual information
RESEARCH AGENDA
Extraction
Integration
Priorisation
Evaluation
STRUCTURED SUSTAINABILITY DATABASE
Larger scope of observed units
Better data quality
Higher temporal coverage
Timeliness of information
More indicators
Single Source of information
Fewer licensing restrictions
FAIR data
Current Landscape: Dispersed Information
Aim: Integrated Sustainability Data

Figure 1: Integration of textual information into existing sustainability data can drive novel use cases and allows enhanced climate risk analysis. Source: Own depiction.

2 Background on NLP for sustainability data

Recent innovations in NLP, especially LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT), have enabled major advances in the availability of research and web-based tools for analyzing documents. Company sustainability and financial reports contain a wealth of data in unstructured, multi-modal (e.g., as tables, graphs, *and* text), and only partially standardized formats. As such, they provide a strong use case for the application of this new generation of NLP approaches. Their potential is illustrated by the fact that freely available online tools for the analysis of texts have mushroomed recently. Next to general-purpose chat bots including OpenAI's ChatGPT and similar (at times derivative) products such as ChatPDF and PDF.ai, there are also products with an exclusive focus on sustainability. Examples of these tools include the Sustainable Development Goals (SDG) Prospector (Jacouton, Marodon, and Laulanié, 2022), which highlights all SDG-related paragraphs in the uploaded documents, or ChatClimate, which targets the analysis corporate sustainability reports.

These solutions, however, generally focus on interactive chat bots with Graphical User Interfaces (GUIs). Similar in design and usability to OpenAI's ChatGPT, they target human, ad-hoc, infrequent users who can profit from a more time-efficient extraction of specific relevant information from sustainability disclosure – essentially, users who do not wish to read complete documents to find specific types or single pieces of information.

Apart from chat bots, academics from a variety of disciplines leverage NLP methods to systematically gather and evaluate sustainability information from large text corpora. In the field of corporate sustainability research, earlier bag-of-words approaches that relied on word-frequency have been increasingly replaced by more sophisticated methods that take the context of textual documents into account and can be leveraged for the extraction and analysis of various types of information. In this context, one strand of research has developed different extensions to BERT models to perform text classification of sustainability-related information, such as FINBERT-ESG (Huang, Wang, and Yang, 2023), ClimateBERT (Leippold, Bingler, Kraus, and Webersinke, 2022), and ClimateQA (Luccioni, Baylor, and Duchene, 2020).

This class of domain-specific language models expands the general BERT model through a pre-training and a fine-tuning stage: During pre-training, the model is augmented with domain-specific texts. In the context of corporate sustainability research, corporate financial and sustainability reports, financial analyst reports, earning call transcripts, (keyword-filtered) news, and scientific abstracts have been used as pre-training data (Huang et al., 2023; Leippold et al., 2022; Luccioni et al., 2020). In the fine-tuning stage, the model is provided with a set of human-annotated texts which have been assigned to a specific outcome category. Such annotation efforts have been undertaken inter alia with regards to the concept of Environmental, Social and Governance (ESG) issues (Huang et al., 2023), each of its subdomains or pillars E, S, and G *separately* (Schimanski et al., 2024), companies' "environmental claims" (Stammbach, Webersinke, Bingler, Kraus, and Leippold, 2022), and particular sustainability disclosure frameworks, i.e., the Taskforce on Climate-Related Financial Disclosures (TCFD) (Bingler, Kraus, Leippold, and Webersinke, 2022; Luccioni et al., 2020).

Domain-specific models have been applied to a variety of tasks including text classification, sentiment analysis, and "fact-checking". These models have also been found to outperform generic language models with regards to the accuracy of text classification (Bingler et al., 2022; Huang et al., 2023; Leippold et al., 2022, 2024; Luccioni et al., 2020). In addition, first proposals suggest that these models could be applied for text classification tasks related to the identification of "greenwashing" (Bingler, Kraus, Leippold, and Webersinke, 2024; Koch, Cooke, Baadj, and Boyne, 2023; Moodaley and Telukdarie, 2023), i.e., the promulgation of unsubstantiated environmental claims (European Commission, 2024).

While domain-specific models have generally focused on the classification of textual data, a second strand of research has applied language models to find and extract numerical as well as textual data. To this end researchers have deployed so called Retrieval-Augmented Generation (RAG) pipelines that add domain-specific context to an LLM prompt. In the field of sustainability research, applications of RAG include the GPT-4 based ChatClimate (Vaghefi et al., 2023) that extracts information from the Intergovernmental Panel on Climate Change (IPCC) AR 6 based on user prompts and ChatReport (Ni et al., 2023), which extracts information from corporate sustainability reports and checks the alignment of the extracted information with TCFD disclosure rules. Another RAG application for extracting sustainability data from companies' sustainability reports is explored by (Bronzini, Nicolini, Lepri, Passerini, and Staiano, 2023), who use a Llama-2 model for a fine-grained assessment of companies' sustainability-linked topics and actions. In addition, (Zou et al., 2023) have tested the performance of different language models in processing sustainability reports, by adopting a RAG pipeline that extracts the numerical and textual indicators that are defined in the Global Reporting Initiative (GRI) and Sustainability Accounting Standards Board (SASB) disclosure standards from pre-processed company reports.

More recently, a similar workflow has been adopted by the Innovation Hub of the Bank for International Settlement's (BISIH) "Project GAIA" (BIS Innovation Hub, 2024), which develops an application that uses GPT-4 in a RAG setting and a module that integrates indicator definitions from legislative texts to extract numerical and categorical Key Performance Indicators (KPIs) from sustainability reports.

These examples from the prior literature underscore the immense potential of novel NLP methods to facilitate the efficient extraction of sustainability-related information from corporate disclosure documents and – once implemented – to do so at a relatively low cost. Their achievements notwithstanding, there are arguably still important challenges that limit the usefulness of such methods for systematic analysis of climate risks and related issues. First, concerning the technical specifications there remain open questions with respect to the validation of the extracted values as well as to cost and time-efficient set-ups of the extraction pipelines. Second, so far there has been comparatively little discussion on how the obtained values can be meaningfully integrated into existing practices of data analysis in the context of climate risk assessments. In light of these challenges, in the following sections we delve further into both technical and user-related issues and propose first steps in a research agenda that tackles these various challenges together.

3 Preliminary results

From a technical point of view, the automatized extraction of information from sustainability reports faces various challenges. With RAG-based pipelines these challenges include the preprocessing of PDFs and the text therein (i.e., the conversion of PDF files into a machine-readable format), cost-efficient procedures for large numbers of PDF documents, and the validation of the extracted values against benchmarks (BIS Innovation Hub, 2024; Bronzini et al., 2023). Especially with regards to validation, the absence of gold-standard benchmarking data has proved to be challenging as existing datasets on corporate sustainability indicators tend to be proprietary, intransparent, and values vary substantially among commercial providers (Berg, Koelbel, and Rigobon, 2022).

To get a clear overview of the challenges and potential trade-offs along the extraction pipeline, we set up a first experiment that enables us to compare different technical specifications of the model but also focuses on the potential pitfalls of human labelled benchmark data. The first step in this experiment was to annotate 39 sustainability reports from large companies from the years 2010 to 2021. These are randomly sampled from the universe of MSCI World firms that published English language reports. The list of selected reports is presented in Table 2 in the Appendix.

We chose to extract the values for Greenhouse Gas (GHG) emissions in our experiment. Compared to other indicators, GHG emissions disclosures are more frequent and less variable as most companies report according to the GHG Protocol (GHGP) standard that was first introduced (WBCSD, 2004). First introduced in 2004 by the World Resource Institute and the World Business Council for Sustainable Development, the GHGP has since been adopted by most large companies and been integrated into regulatory requirements across the world (Jia, Ranger, and Chaudhury, 2022). The GHGP standardizes emission disclosures through three categories of emissions - so called "Scopes" - that reflect the operational control of the company over the released GHG. Accordingly, "Scope 1" emissions denote GHG releases from sources that are directly controlled and operated by the company. "Scope 2" emissions, meanwhile, refer to emissions from that were generated from the generation of electricity that the company purchased. Finally, "Scope 3" emissions refer to other indirect emissions that occur in the company's value chain such as the extraction and production of purchased materials or the use of sold products and services.

Five human annotators extracted Scope 1, 2, and 3 Greenhouse Gas (GHG) emissions. Annotators were asked to open the .pdf file, search for the term "scope 1" (respectively "scope 2" or "scope 3") and a predefined list of synonyms including "direct/ indirect emissions" and extract (if found) the resulting value, unit, variable name, year, page number, and origin (one of "table", "text", or "graphic") into a spreadsheet (see Appendix C.

Among the pitfalls that were encountered by human annotators, missing information is among the most prominent. We found that eleven reports, or 28 percent of the sample, do not report any emission values. The problem of missing information becomes even more accentuated for Scope 2 and 3 emissions, which are often not contained in older reports. A second pitfall concerns unclear and varying concept definitions. For instance, some reports only report employee travel under their Scope 3 GHG emissions, whereas others use this concept to refer to total upstream and downstream emissions. Thirdly, we encountered different ways of disseminating information including text, tables and infographics. A final pitfall is the presence of different measurement units for GHG emissions. While some of these are easy to convert (e.g., tCO_2eq vs $kgCO_2eq$), other units such as emission intensities as opposed to absolute emissions, or CO_2 equivalents as opposed to separate depiction of single greenhouse gases are more problematic in this regard.

The next step was to set up an automatic data extraction pipeline. We use an LLM to convert raw text from PDFs into a structured, tabular format. Since sustainability reports can be rather long, we first need to search for the most relevant content (e.g., pages, tables) before passing it to the LLM. This coupling of search, typically done via embeddings, with LLMs is a common architectural pattern to enhance LLM capabilities, known as naive Retrieval Augmented Generation (Naive RAG) (Gao et al., 2024). Three approaches were tried to extract all scope 1/2/3 GHG emissions for each year from each report:

First, we search for relevant pages and pass the raw text of the so-found pages to an LLM. Specifically, we embed the search query "What are the total CO_2 emissions in different years? Include Scope 1, Scope 2, and Scope 3 emissions if available." using openai's text embedding model *ada-002* and compare it with the embedding of each page from the pdf report. The two most relevant pages from this search are kept, concatenated, and submitted in a single query to openai's flagship LLM, *GPT-4-Turbo*. Based on the raw text from these two pages, the LLM is prompted to answer a list of 48 questions (16 years × 3 scopes): "1. What are the Scope 1 emissions in 2010: 2. What are the Scope 1 emissions in 2011: ", and so on, for all possible combinations of year (2010 - 2025) and scope (1-3). The search query and the complete LLM prompt are provided in appendix

B. The output from the LLM is typically well structured, meaning that it can be parsed using regular expressions to insert the extracted (value, unit)-tuples into a data frame.

The second approach is very similar: The general pipeline, the models, and the queries remain the same. We only change the selection of pages and their handling. We now keep the three most relevant pages from the search, along with each page's preceding and subsequent page. This gives us at most nine pages per report in total. We do not concatenate the pages as in approach 1, but send each page in separate queries to the LLM because we found during preliminary testing that GPT-4-Turbo overlooks relevant values more often if pages were concatenated. The output from each query gets parsed separately, implying that for a single scope-year combination from a single report we may extract more than one value, as the LLM may extract different values from different pages.

Third, again following the same pipeline, we adopt a table-only approach. Since the CO_2 emission scopes are predominantly presented in tables within sustainability reports, we leverage the Python package *pdfplumber*[8], which enables table extraction from PDF files. After extracting the tables, we apply a similar pipeline as in our second experiment, keeping the ten most relevant tables from the search and feeding them into the LLM. We present the results of the three preliminary experiments in Table 1.

Extraction result	**E1**	**E2**	**E3**
Correct result: No CO_2 emissions found	11	11	11
Correct result: All CO_2 emissions extracted	4	1	0
Correct values but wrong units extracted	4	3	0
Retrieval failure: Incomplete text passed to LLM	10	4	NA
LLM extracts information from wrong page	0	5	NA
LLM fails to find ANY correct values	6	3	25
LLM fails to find ALL correct values	4	12	3
Total (N)	39	39	39

Table 1: Short summary of results in preliminary experimentation. E1-3 denotes the experiment 1,2,3. The numbers in the columns are the numbers of reports. NA means that this metric is not straightforward to calculate in experiment E3.

From the results, we notice that, among the 39 reports, all of the applied approaches still struggle to achieve optimal performance on the annotated data. On the positive side, nothing ever gets returned from eleven reports that do not report GHG emission values. The first approach (E1) correctly outputs all the desired values from eight reports. We include four reports in this tally, where the units are not spelled exactly the same way as it was spelled by the human annotator; a harmonization challenge that should be solvable with little effort. The main drawback of E1 is, however, its retrieval strategy: For ten reports we would have liked the algorithm to extract values from specific pages that were not found during our search and were therefore not passed to the LLM. Our second approach (E2) was designed to alleviate this problem: As we widen the search, we reduce the tally of retrieval failures to just four. This success, unfortunately, is not reflected in the number of correctly extracted values (1+3 reports), because the LLM frequently extracts wrong values (five reports) or, in reverse, does not extract values that should have been extracted (3+12 reports). While the performance is not yet satisfactory, these results suggest that future work is needed in three areas: retrieval, usage of LLMs for extraction tasks, and unit harmonization.

[8] https://pypi.org/project/pdfplumber/

The third approach (E3) yields even poorer results for the task. This indicates the inadequacy of only relying on tables for content extraction, even though based on human annotation we would expect that emission values are usually summarized in tables in the reports.

4 Discussion and Research Agenda

As outlined in section 1, the goal of applying NLP techniques to unstructured corporate sustainability information is to extract high-quality data. Notably, this includes a large coverage to enable comparative assessments of transition risks and related use cases by academics, financial supervisors and other public and private institutions. Based on our analysis of the literature and first findings from experiments with a RAG pipeline, we segment the challenges and research gaps for creating a high-quality, accessible database on corporate sustainability into two *how* and one *what* questions.

The first *how* question relates to the design of the RAG pipeline and covers issues like the the set-up of human annotation, prompt engineering, and the extraction of different presentation formats within the sustainability reports (e.g., tables, graphs). The *what* question, in turn, asks which variables should be contained in the structured database. Answering this question, notably, requires domain-specific expertise as it not only relates to the indicators such as GHG emissions that should be extracted, but also to contextual information that could help users to assess the credibility of the reported data. The second *how* question, finally, refers to the post-processing of the extracted values through data science techniques. These operations can include the creation of new indicators pertaining to the reliability of the company disclosed data as well as to the linkage of the extracted values with other datasets.

4.1 How to apply NLP and LLMs to structured data generation?

Annotation. In the absence of transparent and high-quality datasets on companies' sustainability disclosures, the creation of human-annotated validation data becomes a crucial precondition for the evaluation of automatized information extraction pipelines. To serve as a gold-standard for evaluating a model's performance, the quality of human annotation needs to be ensured. Past research making use of human annotations has addressed this aspect by focusing on annotator training and agreement rates (Stammbach et al., 2022).

Apart from its function in validation, systematically comparing between human annotated and automatically extracted information can, however, also deliver insights about the different error types of humans and machines. Regarding the comparison of error types, we note that although annotations generated by LLMs certainly include errors, human annotators are likewise prone to sources of error such as cognitive biases or fatigue. Thus, both types of annotators are imperfect and are likely to reach their maximum potential when complementing each other.

Beyond looking at annotator errors and negligible deviations between automated and human annotations, comparisons can also point to frequent and major errors made by the automatic extraction algorithm, e.g., values that are part of a background image or diagram might not get extracted because the algorithm only uses text. This could be improved with better versions of the algorithm. The most interesting part from a research and policy perspective will, however, be the detection of imprecision and ambiguities in the sustainability reports, like when a report is self-contradictory and mentions different numbers for what should be the same entity, or if a car manufacturer provides the total emissions for its car manufacturing business but does not clarify if this is the same as the company's total emissions. These types of problems let us learn more about the quality of the published sustainability reports and have potential implications for regulatory and standard-setting authorities.

To address both the validation and the research-informing dimensions of annotation, we plan on creating a small-scale gold-standard dataset of emission annotations. We aim to assure a particularly high level of data quality by creating the dataset from LLM annotations that are subsequently evaluated by human annotators and eventually adjudicated by domain experts. In this process we will additionally gain a better understanding of how the complementary annotation process of humans and LLMs can work. Moreover, we aim to document typical sources of error by the LLM and reasons for disagreement between the LLM and the human annotator. In addition, the gold-standard nature of the dataset allows for further evaluations of annotation quality, e.g., through experimental research. The learnings from this small-scale annotation exercise will then also serve as a cornerstone to eventually derive a scalable annotation approach, which will be needed to deploy reliable tools for automated information extraction.

LLM-based Information Extraction. Next to validation and annotation issues, the set-up of an information extraction pipeline also involves a range of technical specifications that need to be systematically addressed. While we have been using GPT-4 within a RAG pipeline, we have found that this process is not straightforward. There are many different choices that can be made and it is often unclear what works best within this setting. When we extracted the raw text from PDF documents (see experiments E1 & E2 in Section 3), any information about the layout of pages and tables and the position of characters within the table got lost. This is clearly not optimal and as a resort we tried table extraction from PDF documents (E3). Yet another possibility to maintain the layout would be to convert PDF files/pages to images for further processing. For retrieval, the challenges include choosing between different embedding models to search for relevant text chunks (e.g., pages), setting appropriate parameters to define the size and overlap between text chunks, and the number of text chunks passed to the LLM. Prompt engineering to make optimal use of LLMs is another big task: the exact wording of prompts matters. One might try prompts that make use of examples (few-shot learning), ask for a single emission value of, e.g., scope 1 in the year 20xx or query the LLM more generally for all available emission values of different scope-year combinations. Getting even more complex, LLM agents as formalized by (Wang et al., 2024) could orchestrate diverse, multi-step workflows where multiple LLMs in various roles and using external tools work together to solve a task.

The LLM output can be structured by requiring JSON output formats or by using function calling if one wants to avoid parsing the textual output from the LLM with regular expressions. Since LLM outputs can differ (depending on another parameter, the temperature), it may be worth querying the LLM repeatedly with identical prompts. Finally, we can ask the LLM for an indicator of certainty, or we can obtain log probabilities for each output token; both of these methods are potentially useful to decide whether we can trust the LLM output or if we should run a different query. Setting up a well-designed study to find out about how to best configure such a data extraction pipeline would be extremely helpful.

In terms of structured content extraction like table extraction from the reports, another difficulty we are always encountering lies in the diverse and non-standardized formats of certain content. For example, a table could have different shapes and styles and some are even incorporated into other content types like graphs. This makes a rigidly structured automatic extraction approach difficult. A possible approach is to train a model on a good number of domain-specific annotated data which could capture the variations of tables and then to deploy this model for the desired use case. However, this approach demands significant annotation efforts and training costs. Alternatively, one could engage a subject matter expert to devise a coding scheme covering all table variations. Subsequently, these variations could be used as prompts for an LLM with contextual learning capabilities to perform few-shot table extraction, as suggested by (Choksi et al., 2024) in content extraction using LLMs with the help of subject matter experts.

As the understanding and interpretation of tables typically depend on other, relevant information from the document – so-called contextualized information –, such table-related content could also be helpful for the extraction task (Gemelli, Vivoli, and Marinai, 2023). In our initial experiments conducted on scope extraction based on table-only content (detailed in Section 3), a notable challenge arises: the potential absence of crucial contextual information during the extraction phase. Therefore, a future research direction could be to conduct the scope extraction based on the tables along with their contextualized information. Leveraging this combined information, the RAG technique of LLMs could be employed to extract the required scopes or other table contents more effectively. Issues that need to be explored include approaches to extract contextual information alongside the tables, integrating this contextual data with the tables, and determining optimal prompts for the extraction processes.

4.2 What information to include in the structured database?

The goal in this comprehensive research agenda is to streamline the automated production of climate-related data from dispersed and unstructured sources into unified, FAIR data (Wilkinson et al., 2016). Findable, because data is in a central repository as opposed to the current situation on dispersed websites. Accessible, because fewer licensing restrictions arise than in the current situation characterized by widely used proprietary data. Interoperable, because information can be compared among reports and linked to other sources. And Reusable, because information from past unstructured reports is preserved.

While existing approaches have focused on extracting indicators prescribed by standard setting bodies (Bronzini et al., 2023) or financial supervisors (BIS Innovation Hub, 2024), the heterogeneity in sustainability reporting practices implies that users would also benefit from additional contextual information that allows them to judge the quality and comparability of extracted indicators. Such additional contextual information could, for instance, include information on calculation methodologies and concept definitions for more ambiguous indicators like Scope 3 emissions. Adding contextual information would enhance the value of a structured database, because despite the existence of standards and protocols to measure and report sustainability performance, a great degree of heterogeneity across currently often unknown dimensions persists in sustainability reporting. Even in the case of emissions data, which is reported by most companies according to the Scopes of the GHG Protocol, great variations across time, methods and observation units (i.e., companies and their boundaries) are possible (see Jia et al. (2022) for a detailed discussion).

A further data need that can be derived from the goal to pursue climate risk analysis consists of the extraction of subsidiary companies and physical assets (e.g., production facilities) from company reports. Obtaining such data could help to fill data gaps for bottom-up and geolocalized assessments on both physical (Rossi et al., 2024) and forward looking transition risks (Bingler, Colesanti Senni, and Monnin, 2021; Kruitwagen, Klaas, Baghaei Lakeh, and Fan, 2021). Their importance notwithstanding, asset-level data are – with few sector-specific exceptions – to date mostly sourced from commercial providers (Kruitwagen et al., 2021).

Another use case for the application of NLP to companies' sustainability reports lies in evaluating the credibility of the disclosed information. In this context, the literature that has proposed to investigate the textual characteristics of sustainability documents to detect instances of greenwashing (Koch et al., 2023; Moodaley and Telukdarie, 2023) could be a starting point. This emerging literature has drawn attention to generic and vague sentences or paragraphs as possible indicators of greenwashing. Further developing the classification of such text snippets could thus contribute to the development of indicators that convey information about the credibility of a sustainability report. In addition, one could think of attributing measures of vagueness and generic nature to specific items and metrics (e.g., CO_2 emissions, decarbonisation targets) to break down credibility assessments to a more granular level.

4.3 How to link the extracted data and assess its quality?

The questions of how to organize the data extraction and what data to extract are also interlinked with considerations about how the data should be treated after extraction. Two key issues in this context are data linkage and post-processing through statistical techniques. Linkage to other structured company information including financial indicators is relatively straightforward, as this concerns mostly large global companies, where company names are relatively standardized and unique identifiers (often ISINs) prevail.

Another possibility of linkage that would be useful for checking the quality of reported information would be to link it to external independent sources such as earth-observation or administrative registers. This could be especially valuable for sectors with high (and sometimes under or misreported, cf. García Vega, Hoepner, Rogelj, and Schiemann (2023)) emissions profiles such as oil and gas extraction, which have already been assessed via remote sensing methods (He et al., 2024). The discrepancies between reported and externally observed values could then feed into the creation of new indicators that alert users about potential reliability issues with the company reported values. Another potential source of such reliability indicators would be to compare the consistency of company reporting over time. By way of example, in the post-processing stage one could compare companies' emission reduction targets over the course of time, i.e., comparing revisions of emission targets for the future as the commitment date nears.

Furthermore, insights regarding data quality and possible inconsistencies can be obtained by linking the extracted information to the offerings of third-party data providers. Ensuring data quality and increasing coverage goes in both directions here: Third-party data providers often draw emissions data from corporate reports too, so the results should, in theory, be unambigious. In reality, however, we have observed that different data providers provide different numbers for the same variable and company even when they all refer to corporate reports. Data drawn from reports via LLMs can be used to verify third-party data and the other way around. Furthermore, third-party providers usually have an estimation method for undisclosed emissions. This can close data gaps that are left open by LLMs, whereas LLMs can close data gaps left by third-party providers due to their lack of interest in smaller companies or specific jurisdictions.

After linkage, it is necessary to provide users with an evaluation of trustability of the source and to resolve conflicts. This post-processing could consist of taking contextual indicators on the data quality of the reports into account. In addition, in line with current market practices, statistics from the obtained structured database itself (e.g., sector averages, deviation from past values) could be used to assess the plausibility of the reported information.

5 Outlook and conclusion

As companies and other stakeholder produce an ever increasing volume of climate and sustainability information, we are confronted with the paradoxical situation, where a wealth of data is freely available, while climate risk analysts simultaneously point to data gaps.

Technological progress in LLMs offers an opportunity to overcome this apparent gulf, by turning dispersed unstructured information into FAIR data. Creating integrated FAIR data, however, comes with technical challenges and domain-specific choices regarding the data infrastructure, both of which should be addressed systematically and transparently as part of an integrated research agenda.

Limitations

Throughout the paper we have highlighted various research gaps, existing shortcomings, and challenges that the research community will need to overcome before high-quality, simple-to-analyze climate-related data extracted from sustainability reports will find more widespread acceptance in fields of research which work more directly %than ourselves on tackling the climate crisis.

Concerning limitations of our extraction pipeline approaches, we note that we have not explicitly addressed questions on the conversion of different units of measurement (e.g., kg vs ktons of GHG). In addition, cost aspects have not been incorporated into our experiments nor in the discussion, although they will be significant to consider when scaling up the proposed extraction pipelines. Since we may need to make over a million LLM requests to extract different indicators and their respective contexts from tens of thousands of reports in order to create an integrated sustainability database, the cost efficiency and – in relation to this – energy efficiency of the computing operations need to be ensured.

Contributions

Andreas Dimmelmeier: Conceptualization, methodology, writing – original draft.

Hendrik Christian Doll: Conceptualization, methodology, visualization, writing – original draft.

Malte Schierholz: Methodology, investigation, software, writing – original draft.

Emily Kormanyos: Conceptualization, methodology, data curation, writing – review & editing.

Maurice Fehr: Resources, writing – review & editing.

Bolei Ma: Software, investigation, data curation.

Jacob Beck: Methodology, data curation.

Alexander Fraser: Conceptualization, supervision, resources, writing – review & editing.

Frauke Kreuter: Conceptualization, supervision, resources, writing – review & editing.

References

Alonso-Robisco, A., Carbó, J. M., Kormanyos, E., and Triebskorn, E. (2024). Houston, we have a problem: Can satellite data bridge the climate-related data gap? *Proceedings of the IFC Workshop on "Addressing Climate Change Data Needs: The Global Debate and Central Banks' Contribution"*.

Berg, F., Koelbel, J. F., and Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, *26*(6), 1315–1344.

Bingler, J. A., Colesanti Senni, C., and Monnin, P. (2021). *Climate transition risk metrics: Understanding convergence and divergence across firms and providers*. https://doi.org/10.2139/ssrn.3923330

Bingler, J. A., Kraus, M., Leippold, M., and Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, *47*, 102776.

Bingler, J. A., Kraus, M., Leippold, M., and Webersinke, N. (2024). How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 107191.

BIS Innovation Hub. (2024). Project Gaia: Enabling climate risk analysis using generative AI. *BIS Technical Report*.

Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., and Staiano, J. (2023). *Glitter or gold? Deriving structured insights from sustainability reports via large language models*.

Carney, M. (2015). Breaking the tragedy of the horizon–climate change and financial stability. *Speech Given at Lloyd's of London*, *29*, 220–230.

Choksi, M., Aubin Le Quéré, M., Lloyd, T., Tao, R., Grimmelman, J., and Naaman, M. (2024). Under the (neighbor)hood: Hyperlocal surveillance on nextdoor. *CHI Conference on Human Factors in Computing*.

Deutsche Bundesbank. (2022). Climate-related data successfully procured. *Press Release*.

European Commission. (2024). *Green claims*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., … Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2312.10997

García Vega, S., Hoepner, A. G. F., Rogelj, J., and Schiemann, F. (2023). *Abominable greenhouse gas bookkeeping casts serious doubts on climate intentions of oil and gas companies* [{SSRN Scholarly Paper}]. Rochester, NY. https://doi.org/10.2139/ssrn.4451926

Gemelli, A., Vivoli, E., and Marinai, S. (2023). CTE: A dataset for contextualized table extraction. *arXiv Preprint arXiv:2302.01451*.

He, M., Ditto, J. C., Gardner, L., Machesky, J., Hass-Mitchell, T. N., Chen, C., … Gentner, D. R. (2024). Total organic carbon measurements reveal major gaps in petrochemical emissions reporting. *Science*, *383*(6681), 426–432. https://doi.org/10.1126/science.adj6233

Huang, A. H., Wang, H., and Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841.

IPCC. (2022). Climate change 2022: Impacts, adaptation and vulnerability. *Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.

Jacouton, J.-B., Marodon, R., and Laulanié, A. (2022). The proof is in the pudding. *AFD Research Papers*, (262), 1–48.

Jia, J., Ranger, N., and Chaudhury, A. (2022). *Designing for comparability: A foundational principle of analysis missing in carbon reporting systems* [SSRN Scholarly Paper]. https://doi.org/10.2139/ssrn.4258460

Koch, N. S., Cooke, D., Baadj, S., and Boyne, M. (2023). Market review of environmental impact claims of retail investment funds in Europe – 2DII. Retrieved from https://2degrees-investing.org/resource/market-review-of-environmental-impact-claims-of-retail-investment-funds-in-europe/

Kruitwagen, L., Klaas, J., Baghaei Lakeh, A., and Fan, J. (2021). *Asset-level transition risk in the global coal, oil, and gas supply chains* [SSRN Scholarly Paper]. Rochester, NY. https://doi.org/10.2139/ssrn.3783412

Leippold, M., Bingler, J. A., Kraus, M., and Webersinke, N. (2022). ClimateBert: A pretrained language model for climate-related text. *University of Zurich Working Paper. Available at Https://Www.zora.uzh.ch/Id/Eprint/235046/.*

Leippold, M., Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., … Huggel, C. (2024). *Automated fact-checking of climate change claims with Large Language Models*. arXiv. Retrieved from http://arxiv.org/abs/2401.12566

Luccioni, A., Baylor, E., and Duchene, N. (2020). Analyzing sustainability reports using Natural Language Processing. *Tackling Climate Change with Machine Learning Workshop at NeurIPS 2020*. https://doi.org/10.48550/arXiv.2011.08073

Moodaley, W., and Telukdarie, A. (2023). Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review. *Sustainability*, *15*(2), 1481.

NGFS. (2022). Final report on bridging data gaps. *Network for Greening the Financial System Technical Document*.

Ni, J., Bingler, J., Colesanti Senni, C., Kraus, M., Gostlow, G., Schimanski, T., et al.others. (2023). CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. *arXiv Preprint arXiv:2307.15770*.

Rossi, C., Byrne, J. GD., and Christiaen, C. (2024). Breaking the ESG rating divergence: An open geospatial framework for environmental scores. *Journal of Environmental Management*, *349*, 119477. https://doi.org/10.1016/j.jenvman.2023.119477

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., and Leippold, M. (2024). Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, *61*, 104979.

Stammbach, D., Webersinke, N., Bingler, J. A., Kraus, M., and Leippold, M. (2022). A dataset for detecting real-world environmental claims. *Center for Law & Economics Working Paper Series*, *2022*(07).

Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., et al.others. (2023). ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, *4*(1), 480.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., … Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

WBCSD, W. (2004). The greenhouse gas protocol.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., … Zhou, W. (2023). *ESGReveal: An LLM-based approach for extracting structured data from ESG reports*. arXiv. Retrieved from http://arxiv.org/abs/2312.17264

A Annotated reports

Company	Year	Language
AbbVie	2019	en
Amazon	2020	en
Apple	2021	en
ASML	2016	en
ASML	2018	en
BASF	2015	en
BASF	2018	en
Chevron	2020	en
Cocacola	2016	en
Continental	2013	en
Continental	2021	de
Deutsche Bank	2015	en
Deutsche Bank	2016	en
Deutsche Bank	2017	en
Deutsche Post	2012	en
Eli Lilly	2010	en
E.ON	2010	en
E.ON	2015	en
Exxon Mobil	2014	en
Fresenius medical care	2021	en
Infineon	2014	en
Infineon	2020	en
JP Morgan Chase	2014	en
Mercedes-Benz group	2014	en
Mercedes-Benz group	2021	en
Microsoft	2010	en
Microsoft	2019	en
Novo Nordisk	2019	en
Novo Nordisk	2020	en
Pepsico	2015	en
Pepsico	2019	en
Pfizer	2019	en
Puma	2013	en
Puma	2014	en
Puma	2018	en
RWE	2014	en
Samsung	2018	en
Volkswagen	2019	en
Walmart	2017	en

Table 2: Overview of the 39 annotated sustainability reports in the preliminary study.

B Prompts used with experiment E1

Search query used with ada-002

```
What are the total CO2 emissions in different years?
Include Scope 1, Scope 2, and Scope 3 emissions if available.
```

LLM prompt used with GPT-4-Turbo

```
Extract key pieces of information from this sustainability report.
If a particular piece of information is not present, output \"Not specified\".
Always include unit of measurement in your answer.

Use the following format:
0. What is the title
1. What are the Scope 1 emissions in 2010
2. What are the Scope 1 emissions in 2011
3. What are the Scope 1 emissions in 2012
4. What are the Scope 1 emissions in 2013
5. What are the Scope 1 emissions in 2014
6. What are the Scope 1 emissions in 2015
7. What are the Scope 1 emissions in 2016
8. What are the Scope 1 emissions in 2017
9. What are the Scope 1 emissions in 2018
10. What are the Scope 1 emissions in 2019
11. What are the Scope 1 emissions in 2020
12. What are the Scope 1 emissions in 2021
13. What are the Scope 1 emissions in 2022
14. What are the Scope 1 emissions in 2023
15. What are the Scope 1 emissions in 2024
16. What are the Scope 1 emissions in 2025
17. What are the Scope 2 emissions in 2010
18. What are the Scope 2 emissions in 2011
19. What are the Scope 2 emissions in 2012
20. What are the Scope 2 emissions in 2013
21. What are the Scope 2 emissions in 2014
22. What are the Scope 2 emissions in 2015
23. What are the Scope 2 emissions in 2016
24. What are the Scope 2 emissions in 2017
25. What are the Scope 2 emissions in 2018
26. What are the Scope 2 emissions in 2019
27. What are the Scope 2 emissions in 2020
28. What are the Scope 2 emissions in 2021
29. What are the Scope 2 emissions in 2022
30. What are the Scope 2 emissions in 2023
31. What are the Scope 2 emissions in 2024
```

```
32. What are the Scope 2 emissions in 2025
33. What are the Scope 3 emissions in 2010
34. What are the Scope 3 emissions in 2011
35. What are the Scope 3 emissions in 2012
36. What are the Scope 3 emissions in 2013
37. What are the Scope 3 emissions in 2014
38. What are the Scope 3 emissions in 2015
39. What are the Scope 3 emissions in 2016
40. What are the Scope 3 emissions in 2017
41. What are the Scope 3 emissions in 2018
42. What are the Scope 3 emissions in 2019
43. What are the Scope 3 emissions in 2020
44. What are the Scope 3 emissions in 2021
45. What are the Scope 3 emissions in 2022
46. What are the Scope 3 emissions in 2023
47. What are the Scope 3 emissions in 2024
48. What are the Scope 3 emissions in 2025

For example, answer as follows:
0. What is the title: Our responsibility. Report 2014
1. What are the Scope 1 emissions in 2010: <value> <unit>
2. What are the Scope 1 emissions in 2011: <value> <unit>
Please continue with your answer:
```

Regular expression used with this LLM prompt

The following regular expression extracts scope, year, value and unit:

```
What are the Scope ([123]{1}) emissions in (20[12]\d): ([0-9\.,]+) (.{0,50})
```

A separate regular expression extracts whether the LLM outputs ”not specified”:

```
What are the Scope ([123]{1}) emissions in (20[12]\d): (Not specified)$
```

C Annotation Guide for sustainability reports

Annotators were provided with the company reports in .pdf format alongside with an E

1. Open the Excel file with the list of sustainability reports
2. For each line with your name, open the relevant pdf of the sustainability report
3. Open the search field in "Adobe Reader" by pressing "ctrl+f"/ "strg+f"
4. Find each term "Scope 1", "Scope 2", "Scope 3" into the search form.
   - Scope 1 can also be called: "direct emissions", "GHG emissions".
   - Scope 2 can also be called: "indirect emissions".
   - Scope 3 can also be called: "carbon footprint".
   - If no results, fill columns D – I in that line with "Na" and go to the next line.
5. If step 4 yields results, go through the search results until a number value with an emission value shows
   - If Scope 1 and Scope 2 are calculated together, use the Scope 2 row in Excel.
6. Extract the information found by Copy/ Pasting the values into columns D – I into the excel file "daten.xlsx"
   - Value (e.g. "260,2")
     - Remove separators for thousands.
     - If there are "larger than" operators ("<" or ">"), include them.
     - Do not include relative values (e.g. "26% lower").
     - If the information is contained in a Graphic, write "Na".
   - Unit (e.g. "tons CO2 eq")
   - Variable Name (e.g. "Scope 1 CO2 equivalents")
   - Year (e.g. "2010" or "1998-2001")
     - Write down all years that are in the Report by adding a newline to the Excel sheet.
   - Page number (Take the page number that is shown in Adobe Reader, where you found the information)
   - Type (one of "Table", "Text", or "Graphic")