O EPJ Data Science

a SpringerOpen Journal

RESEARCH   **Open Access**

Check for updates

Glitter or gold? Deriving structured insights from sustainability reports via large language models

Marco Bronzini[1,2]* (iD), Carlo Nicolini[2] (iD), Bruno Lepri[2,3] (iD), Andrea Passerini[1] (iD) and Jacopo Staiano[1] (iD)

*Correspondence:
marco.bronzini-1@unitn.it
[1] University of Trento, Trento, Italy
[2] Ipazia S.p.A., Milan, Italy
Full list of author information is
available at the end of the article

Abstract

Over the last decade, several regulatory bodies have started requiring the disclosure of non-financial information from publicly listed companies, in light of the investors' increasing attention to Environmental, Social, and Governance (ESG) issues. Publicly released information on sustainability practices is often disclosed in diverse, unstructured, and multi-modal documentation. This poses a challenge in efficiently gathering and aligning the data into a unified framework to derive insights related to Corporate Social Responsibility (CSR). Thus, using Information Extraction (IE) methods becomes an intuitive choice for delivering insightful and actionable data to stakeholders. In this study, we employ Large Language Models (LLMs), In-Context Learning, and the Retrieval-Augmented Generation (RAG) paradigm to extract structured insights related to ESG aspects from companies' sustainability reports. We then leverage graph-based representations to conduct statistical analyses concerning the extracted insights. These analyses revealed that ESG criteria cover a wide range of topics, exceeding 500, often beyond those considered in existing categorizations, and are addressed by companies through a variety of initiatives. Moreover, disclosure similarities emerged among companies from the same region or sector, validating ongoing hypotheses in the ESG literature. Lastly, by incorporating additional company attributes into our analyses, we investigated which factors impact the most on companies' ESG ratings, showing that ESG disclosure affects the obtained ratings more than other financial or company data.

Keywords: ESG dimensions; Non-financial disclosures; Large language models; In-context learning; Knowledge graphs; Bipartite graph analyses; Interpretability

1 Introduction

Public health, climate change, social inequalities, diversity, and inclusiveness are challenges that need global attention as well as innovative and collaborative solutions. However, building a sustainable society requires defining a common set of sustainable-related issues to disclose, measure and comply with. ESG, which stands for Environmental, Social, and Governance, is an established set of principles used to monitor the sustainability and ethical practices of businesses within society. These three E/S/G aspects are further described via more granular indicators (both qualitative and quantitative) concerning, for

Springer

example, waste management, emissions, labour rights, and diversity. These indicators aid in evaluating the degree to which a corporation contributes to achieving societal goals. Assessing these ESG aspects can also help monitor the progress of the seventeen Sustainable Development Goals (SDGs) included in the United Nations' 2030 Agenda for Sustainable Development [1] which sets ambitious goals for building a sustainable society such as gender equality, responsible consumption and production, and climate action.

Over the last decade, there has been a growing demand for disclosing companies' non-financial information. This demand comes from legislation such as the European Union's Non-Financial Reporting Directive (NFRD, [2]) which requires all public-interest companies with more than 500 employees to do so. The more recent European Union's Corporate Sustainability Reporting Directive (CSRD, [3]) further increases this demand by enlarging the pool of companies concerned by a factor of 4: from roughly 12 thousand to 50 thousand companies [3].

Non-financial disclosures are typically reported in sustainability reports, web pages, social media posts, news, press releases or earning calls. To overcome this variety of sources, stakeholders generally rely on a third-party assessment of corporations' ESG performance to inform their decisions: ESG ratings provided by agencies such as *Sustainalytics*, *MSCI*, *S&P Global*, *Moody's* and *Refinitiv* [4]. These rating agencies rely on proprietary assessment methodologies with different perspectives on the measurement, scope and weight of different ESG aspects. This creates divergences in companies' evaluations across agencies and thus unsatisfactory degrees of explainability, transparency or fairness [5–8]. Stakeholders might overcome this issue by directly accessing non-financial information and imposing their scope and weight to assess corporates' ESG performance [4, 9]. However, extracting meaningful insights from ESG-related data sources can be challenging and laborious, often including lengthy documents. Consequently, the stakeholders can face significant obstacles in evaluating companies' ESG performance from opaque and divergent assessments to verbose textual documents. We posit that a data-driven approach, coupled with state-of-the-art Natural Language Processing (NLP) techniques, can provide automatic tools to extract insights from companies' disclosures such as sustainability reports. Further, the proposed methodology allows us to better understand the companies' ESG assessments and to unveil relationships between what companies disclose and their ratings.

The purpose of this work is to automatically extract and analyse the ESG initiatives disclosed by companies in their sustainability reports, and to investigate how these impact their ESG performance assessment. Table 1 presents the definitions of the ESG-related terms adopted henceforth. Our proposed methodology relies on Large Language Models (LLMs, [10–12]) for Information Extraction purposes, on graph-based representations for data analysis, and on the SHapley Additive exPlanations (SHAP, [13]) framework for the interpretability of the ESG scores. Specifically, we employed a generative LLM to extract structured insights from companies' sustainability reports, bipartite graphs[1] to conduct analyses on them, and the SHAP framework on linear regression to investigate the impact of companies' disclosures on ESG scores. Turning the unstructured information from sustainability reports into a structured and unified format allows us to build graph-based representations; these, in turn, are directly usable and exploitable for a diverse set of tasks,

---

[1]A bipartite graph is a graph whose nodes can be divided into two distinct and independent partitions [14, 15].

Table 1 Definitions of the terms used in this study

Term	Definition	Examples
Initiatives	Textual disclosures related to companies' ESG-related efforts	*Investment in recycling technologies, partnership with non-governmental organizations*
Topic/Category	An aspect related to Corporate social responsibility (CSR)	*Supply Chain, Waste, Packaging, Philanthropy*
Predicate	A nominalized verb that affects the ESG-related aspect	*Investment in, Advisory support for*
Action	Semantic concatenation between an ESG topic and its associated predicate from the generated triple	*WASTE:Investment in, PHILANTHROPY: Advisory support for*

from exploring and navigating the data to discovering emerging patterns via statistical and interpretability analyses.

As mentioned before, for this first Information Extraction step, we employ an instruction-finetuned generative LLM (WizardLM, [16]), leveraging both In-Context Learning (ICL, [17]) and Retrieval Augmented Generation (RAG, [18]), to extract ESG-related actions as triples from companies' sustainability reports. LLMs have consistently been shown to hold semantic understanding and store factual knowledge [18–21]; Instruction-tuning (that is, providing task descriptions via natural language instructions) further enhances their capabilities to address downstream NLP tasks, such as Information Extraction [10, 16, 22–24]; LLMs have also demonstrated a remarkable ability as few-shot learners using In-Context Learning [17], a technique that relies on providing a few input-output samples within the model instruction.

The triple format allows us to represent the document sentences in a standardised format following a pre-defined semantic template tailored to ESG aspects. These extracted triples are then used to build a Knowledge Graph (KG), following the same research direction of recent studies based on OpenAI's commercial LLMs [25–28]. This allows us to condense all the companies' disclosures through a graph representation which offers a versatile approach to illustrate structured information [29] as concepts (nodes) and their relationships (edges) [30, 31]. The generated KG is then decomposed into several bipartite graphs, thus two-fold graph representations, to analyse the companies' disclosures by inspecting the extracted information from various angles, including the company and topic perspective.

Our findings include descriptive, similarity and correlation analyses concerning ESG-related actions disclosed by companies in their sustainability reports. These analyses unveiled that companies addressed ESG topics from several perspectives, spotlighting the complexity of ESG-related matters. In addition, similarities in companies' disclosures emerged from companies from the same geographical region and the same sector, remarking the possible influence of exogenous factors [32, 33] and the presence of sector-focused topics [34–36]. Further, our interpretability analysis of ESG scores highlights how a company's disclosures impact its ESG rating more than other company-specific or financial aspects. Finally, our analyses show the rewards of transparency: comprehensive disclosure of non-financial information appears to affect ESG scores positively, whereas, conversely, reporting on a limited set of ESG categories seems detrimental; interestingly, we also observe a significant incidence of other factors not directly related to ESG, such as a company's incorporation year and continent (Europe).

Overall, our work contributes to the literature on sustainability and Corporate Social Responsibility (CSR) by proposing an advanced NLP pipeline for automatically extracting information from companies' sustainability reports and validating some ongoing hypotheses using a data-driven approach and the company perspective.

2 Related work

In this section, we first discuss the state-of-the-art approaches for creating Knowledge Graphs (KGs, Sect. 2.1), encompassing a spectrum ranging from conventional NLP pipelines to the exploitation of Large Language Models (LLMs). In Sect. 2.2, we then summarise the main studies focused on ESG-related textual information. Lastly, we present ongoing discussions within the sustainable finance literature in Sect. 2.3.

2.1 Knowledge graphs generation

Knowledge Graphs (KGs, [30, 31]) offer a versatile method of representing knowledge that can be leveraged in various use cases and domains [37, 38]. They can be applied to question-answering [39], recommendation systems [40], and information retrieval [29]. The task of KG generation, also known as knowledge acquisition, aims to create KGs by extracting information from unstructured, semi-structured or structured sources as well as augmenting existing graphs [38, 41, 42]. Traditional approaches for knowledge acquisition involve several NLP tasks which are generally disjointly learned, a process that is prone to error accumulation. To overcome this problem, new one-stage NLP pipelines have been proposed to jointly extract both entities and relations [43, 44].

In this context, Open Information Extraction (OIE, [45]) emerged as the task of extracting structured information, typically in the form of subject-predicate-object (SPO) triples, without relying on a predefined template or a specific domain. These SPO triples can then be leveraged to generate knowledge graphs based on the subjects, predicates and objects extracted from textual documents. This approach can mitigate the impact of depending on external knowledge, such as patterns and domain-specific heuristic rules present in the training data. Recently, OIE models (e.g., Multi^{2}OIE [46] and DeepEx [47]) have employed transformer-based LLMs (e.g., BERT [48]) to extract both syntactic and semantic information [49].

LLMs trained on large-scale corpora have demonstrated significant potential across diverse NLP tasks thanks to the technique of prompt engineering [50]. Many researchers [25–28] have demonstrated the ability of LLMs (i.e., OpenAI's GPT models) to extract structured data from texts, accomplishing the KG generation task. Furthermore, LLMs can enhance KGs through several other perspectives: (1) enrich entity and relation representations using embeddings, (2) generate new facts (i.e., KG completion [51]), (3) produce natural-language descriptions of KG facts (i.e., KG-to-text [52]) and (4) answer natural-language questions (i.e., question-answering [50]).

We follow this promising research direction by exploiting the semantic understanding, generative abilities and flexibility of LLMs to extract ESG-related structured information. Our methodology adopts a Large Language Model, alongside the Retrieval Augmented Generation (RAG) paradigm [18] and the In-Context Learning (ICL) technique, to overcome the main limitation of conventional OIE methods in achieving our goal of extracting ESG-related structured insights: OIE methods traditionally extract structured information by relying only on the syntactical structure of a sentence, without any predefined template, which poses important limitations in the type of information extracted.

They may fail to extract, for example, information related to specific domains or activities that are not the direct subject of the sentence. An information extraction approach based on generative LLMs allows us, instead, to generate semantically-aware and ESG-oriented triples instead of traditional SPO triples, enabling the creation of a full-fledged ESG-oriented KG.

2.2 Text analytics on ESG-related information

Several works have explored the use of NLP technology to process companies' non-financial textual information and extract meaningful insights concerning statements, facts and actions disclosed by companies. For example, Chou *et al.* [53] applied distributional semantics on 10-K filings (i.e., annual financial reports required by the U.S. Securities and Exchange Commission) for extracting topics related to climate change that are disclosed by companies, aiming to monitor environmental policy compliance. Raghupathi *et al.* [54] applied traditional approaches ranging from trigram co-occurrences to clustering and classification to gain insights into the shareholders' perspectives and objectives regarding sustainability and climate change. LLMs have also been leveraged for Text Analytics in the ESG domain. Jacouton *et al.* [55, 56] introduced *SDG Prospector*, a tool exploiting LLMs to identify paragraphs in Public Development Banks' sustainability reports that address SDGs. Similarly, Webersinke *et al.* [57] introduced *ClimateBert*, a transformer-based language model fine-tuned for climate-related classification tasks. Vaghefi *et al.* [58] adopted the paradigm of RAG [18] to enrich GPT-4 with the ability to reliably answer climate-related questions, by augmenting the posed questions with contextual information retrieved from the Sixth Assessment Report, released by the United Nations Intergovernmental Panel on Climate Change (IPCC). The authors also released a conversational agent [59] based on their proposed approach. Ni *et al.* [60], developed *ChatReport*, an LLM-based tool that evaluates companies' sustainability reports according to the eleven recommendations [61] provided by the Task Force for Climate-Related Financial Disclosures (TCFD). The tool combines semantic search to identify text chunks that are pertinent to each recommendation and LLM prompting to summarize them.

Our work differs from existing approaches by jointly (i) leveraging generative LLMs for Knowledge Graph (KG) generation, (ii) employing an open-sourced generative LLM and (iii) relying directly on companies' sustainability reports. This methodology, alongside the exploitation of bipartite graph representation, allows us to conduct non-trivial analyses concerning ESG categories and actions disclosed by companies.

2.3 Sustainable finance

Companies' non-financial disclosures might be influenced by the diverse regulatory requirements specific to different regions [62], as well as by factors such as the political, labour, and cultural landscape of countries they operate in [32, 33, 63].

European, American and Asian companies could address their Corporate Social Responsibility (CSR) by prioritising different socially responsible efforts, investments, and disclosures based on the demands coming from their region. For instance, Baldini *et al.* [33] found that a country's labour union density positively impacts social and governance disclosure, whereas Yu *et al.* [32] unveiled that the lack of political rights has a negative influence on ESG disclosure. On the other hand, CSR might also be influenced by regulatory agencies [62]. For example, European companies might prioritise environmental-related issues due to more stringent climate regulations in Europe, such as the European

Union's Emissions Trading System [64] or the ambitious "Fit for 55" plan [65] recently proposed by the European Union to achieve climate-neutrality by 2050.

This greater European commitment towards environmental aspects might also cause biases in benchmarking companies' environmental performance. For example, the study of LaBella *et al.* [66] unveiled the important role of ESG rating agencies and their biases towards European companies. The authors [66] discovered a bias among rating agencies when evaluating ESG performance, showing a preference for European companies over their North American, Emerging Markets, and Developed Asian counterparts. They proposed that this bias could stem from variations in formal reporting requirements across different jurisdictions, contributing to differences in the quality of companies' non-financial disclosures [66]. This indirectly spotlights the pioneering and benchmark role of European companies concerning non-financial disclosures and CSR. Additionally, there are other studies [63, 66, 67] that suggest non-financial disclosures, and indirectly, companies' ESG assessments, might also suffer from biases based on company size. This is because generating non-financial disclosures can be both financially and labour-intensive [66]. As a result, larger companies could invest more economic and human resources in improving their non-financial transparency, positively influencing their ESG evaluation.

On the other hand, the company's ESG performance is traditionally assessed relative to industry peers due to industry-specific ESG concerns. This allows rating agencies (e.g., Refinitiv) to assess a company's ESG performance by outweighing the ESG topics relevant to the company's industry. For instance, packaging could be more relevant for companies producing consumer staples (e.g., the beverage industry), while materials companies (e.g., the chemicals industry) might be affected by physical-related topics such as Employee Safety. This factor has already emerged in the ESG literature [34–36] and is named ESG Industry Materiality. In the ESG-related debate [68, 69], materiality generally identifies two types of sustainability issues in an industry: *financial materiality* refers to issues that must be disclosed due to their potential significant effects on financial performance, and *impact materiality* pertains to information on the impact of a company's activities on the surrounding ecosystems and social systems. Whereas, reporting standard organizations, such as the Sustainability Accounting Standards Board (SASB),[2] identify environmentally, socially or financially relevant issues for a specific industry (known as materiality matrix) that necessitate disclosure, such as Water Management within the Non-Alcoholic Beverages industry. This aids companies in enhancing their disclosure of pertinent subjects within their industry. Our data-driven methodology enriches these ongoing discussions in the literature by providing structured insights on ESG topics directly extracted from companies' non-financial textual disclosures (Sect. 4). This facilitates a deeper examination of companies' ESG initiatives compared to the prevalent dependence on proprietary assessments and tools in current literature. In the discussion section (Sect. 5), we leverage these data-driven insights to extend upon existing debates in the literature, concerning, for example, the diversity of actions taken by companies, as well as correlations across sectors and geographic regions.

---

[2]https://sasb.ifrs.org.

3 Materials and methods

This section discusses the data, the approaches, and the methods used in this work. First, we describe the data sources (Sect. 3.1); second, we provide a detailed overview of our approach, from data preparation (Sect. 3.2) to triple generation (Sect. 3.3) and KG generation (Sect. 3.4). Finally, we discuss the methods and approaches used to analyse, compare and evaluate the findings concerning the generated triples (Sect. 3.5).

3.1 Data sources

Here, we describe the three main data sources used in our work, which include (i) sustainability reports (Sect. 3.1.1), (ii) an ESG categorization (Sect. 3.1.2), and (iii) ESG rating scores and other companies' information (Sect. 3.1.3).

3.1.1 Sustainability reports

Sustainability reports are non-financial documents published by companies to disclose information concerning the impact of their activities on the environment and people. Therein are described the actions the company took or expects to take regarding ESG matters – such as respect for human rights, fair treatment of employees, anti-corruption and bribery as well as board diversity [3]. However, ESG reporting can be subjective and opaque due to the complexity of reporting qualitative aspects, particularly those related to social and governance issues [70]. Furthermore, the lack of a standardised framework for ESG reporting makes quantitative/comparative analyses difficult [71, 72].

We initially collected 6456 sustainability reports from 4222 different companies using the report URLs available on two public websites [73, 74]. Although these sustainability reports are mainly written in English (94% of all the available documents), the nationality of the companies is fairly diversified, covering 74 different countries. However, the majority of the available reports (56%) come from North American companies. For our study, we consider only the reports written in English because of its broad coverage and the wide range of pre-trained language models available for this language (many thousands of language models on the Hugging Face platform [75]).

Concerning the period covered, we gather reports up to fiscal year 2022 (9.6% of the available documents), even though the majority of the documents (54%) refer to the fiscal year 2021 (i.e. a fiscal year is a twelve-month period that is generally equal to a calendar year). This temporal distribution displays an expected coverage since we gathered these sustainability reports in February 2023 with the majority of the reports published throughout 2022 disclosing information concerning the previous fiscal year (2021).

Nevertheless, processing more than six thousand sustainability reports from four thousand diverse companies poses a significant computational challenge, especially when using resource-intensive generative LLMs that require a significant increase in computational budget. To showcase our pipeline and present meaningful insights, we consequently select a representative subset of companies by balancing sector, region, company notoriety, size and age representation. For the notoriety criteria, we subjectively consider well-known companies with high capitalization (e.g., NVIDIA), low ESG-related performance (e.g., Saudi Aramco) and recent controversies (e.g., First Republic Bank). This subset compounds to 124 companies spanning 3 continents, 15 countries and 11 GICS (Global Industry Classification Standard)[3] sectors (Table 2). The sample covers companies established

---

[3]https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard.

Table 2 Selected companies by sector. Each sector showcases the number of companies represented. The "Companies" column shows a glimpse of the representative companies within each sector, offering a snapshot of the prominent companies chosen

Sector	Number	Companies
Industrial	20	Airbus, 3M Corporation, Adecco, Daikin Industries, …
Information Technology	16	Apple, STMicroelectronics, Intel, NVIDIA, LG Display, …
Financials	13	Bank of America, Deutsche Bank, Visa, Mastercard, …
Consumer Staples	11	Coca-Cola, British American Tobacco, Walmart, …
Consumer Discretionary	11	Amazon, Adidas, Toyota Motor, Alibaba, Tesla, …
Communication Services	10	Fox, Netflix, Walt Disney, Activision Blizzard, Meta, …
Health Care	10	Amplifon, AstraZeneca, Bayer, Johnson&Johnson, …
Materials	9	United States Steel, ArcelorMittal, DuPont, Croda, …
Energy	9	Saudi Aramco, TotalEnergies, Paramount Resources, …
Real Estate	8	China Evergrande, Park Hotels Resorts, British Land, …
Utilities	7	American Electric Power, Edison, Enel, Uniper, …
	124	

from the 19th to the 21st century exhibiting a wide range of market capitalization from 0.4 USD billion to 2901 USD billion.

For this study we only considered the most recent sustainability report for each company in the selected subset, so to avoid any skewness effect. Note that the proposed methodology can be used as-is for longitudinal studies by simply using sustainability reports referred to several years.

Further details concerning the distribution of the fiscal years considered for this work, the complete list of the considered companies and the original dataset can be found in the Supplementary Material (SM) document (see Additional file 1, Sects. SM1.2, SM1.3 and SM1.4).

3.1.2 ESG categorization

We adopt the ESG categories proposed by Berg *et al.* in [8]. The authors grouped, using a bottom-up methodology, nearly seven hundred ESG-related indicators from six distinct ESG data providers (Sustainalytics, S&P Global, Refinitiv, Moody's ESG, Morgan Stanley Capital International-MSCI, and MSCI-KLD) into a unified categorization comprising sixty-four ESG categories. These categories encompass environmental, social and corporate governance issues including Employee Development, Supply Chain, Climate Risk Management, Energy, Financial Inclusion, Biodiversity, Customer Relationship, Access to Basic Services, and Board Diversity. A complete list of all the ESG categories can be found in the Supplementary Material file (see Sect. SM1.1).

3.1.3 ESG ratings and other company information

Rating agencies such as Refinitiv, MSCI, and Sustainalytics utilise non-financial reports and ESG-related information to systematically assess the impact of companies' activities on the environment and society. This assessment, typically done through numerical scores, offers stakeholders valuable measures for evaluating and comparing companies in terms of their performance related to ESG topics.

Among those, we used the Refinitiv platform,[4] which provides high-quality financial and ESG data. The Refinitiv ESG ratings are given as percentage scores [76], wherein lower values (0-25) indicate poor ESG performance and a lack of transparency in publicly reporting

---

[4]https://eikon.refinitiv.com – recently rebranded as LSEG Data&Analytics.

material on ESG data (i.e., laggard companies); conversely, higher values (75-100) indicate excellent ESG performance and a high level of transparency in publicly reporting material on ESG data (i.e., leader companies). In addition, it is worth mentioning that a zero score is assigned in the rare case that a company does not disclose any metrics or information relevant to its industry [9, 77]. The combined ESG score is a cumulative measure of E/S/G pillars' weights, which differ by industry for the first two pillars (E and S), while the weight of the third pillar (G) remains consistent across all industries [76]. A company's ESG performance is indeed assessed relative to industry peers due to industry-specific ESG concerns (Sect. 2.3).

For each company, we collected twelve company features encompassing ESG scores, unchanging company details and annual financial data (regarding the same fiscal year of the considered sustainability reports). Specifically, we gathered: the combined ESG scores, individual scores for the E/S/G pillars, company sector, industry, country, region and continent, number of employees, market capitalization, EBITDA (Earnings Before Interest, Taxes, Depreciation, and Amortization), and total liabilities.

3.2 Data preparation

Our NLP pipeline consists of several components (Fig. 1), including some pre-processing methods, to extract structured insights from sustainability reports.

In this section, we describe the NLP methods adopted to prepare the data for our subsequent analyses. These pre-processing methods include extracting text from PDF files and segmenting sentences (Sect. 3.2.1) as well as using semantic search to select only sentences related to ESG topics (Sect. 3.2.2). The latter relies on the ESG categorization introduced in Sect. 3.1.2.

3.2.1 NLP pipeline

Sustainability reports are generally visually rich and lengthy PDF documents, for example, our sample has a median page value equal to 61 pages. Unfortunately, the incentive of companies to present visually appealing infographics and tables, results in degraded quality of standard text extraction tools.

Hence, for the textual part, we rely on a PDF parser (PyMuPDF, [78]) and apply standard preprocessing steps to improve the quality of the extracted text and reduce the artefacts generated by the parser. Specifically, regular expressions are used to add a full stop between
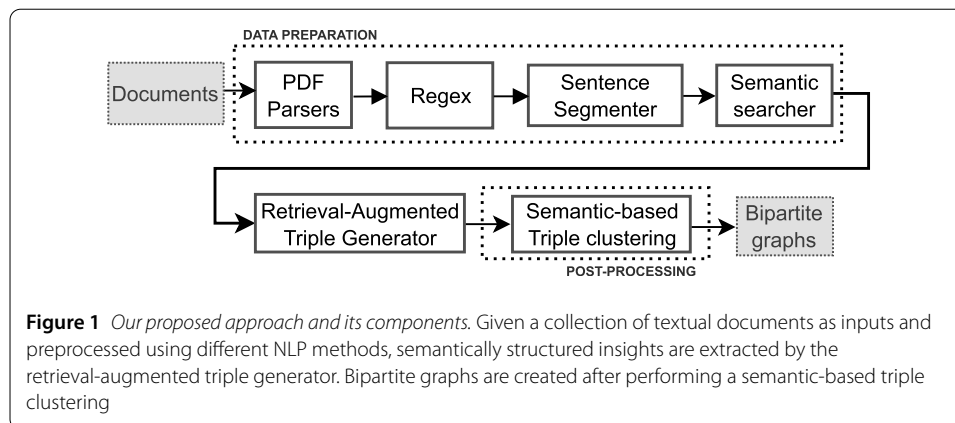

Figure 1 *Our proposed approach and its components.* Given a collection of textual documents as inputs and preprocessed using different NLP methods, semantically structured insights are extracted by the retrieval-augmented triple generator. Bipartite graphs are created after performing a semantic-based triple clustering

two sentences when missing, remove new lines in the middle of sentences and remove duplicate white spaces and lines.

Processing textual data also requires defining the granularity of the input data according to purpose, needs and limitations. A text corpus contains textual items representing singular tokens, words, sentences, paragraphs, or entire documents. We adopt sentence-level textual inputs as a good trade-off between the semantic meaning conveyed by a sentence and technical limitations (e.g., the maximum prompt length of the language model). Consequently, after extracting textual data from the sustainability reports, we decompose each report into sentences with PySBD [79, 80], a tool widely considered the state of the art in Sentence Boundary Disambiguation (SBD).

3.2.2 Asymmetric semantic search

Sustainability reports include generic and vague statements, for example, phrases such as "Air is something that surrounds us 24 hours a day".[5] Accordingly, a filtering process is required to consider only ESG-related sentences for downstream tasks. The two most well-known filtering methods for information retrieval are keyword-based and vector-based search [81]. We adopt the approach of neural semantic search [82–85], a vector-based search method, that exploits text embeddings to represent both documents and queries in the same vector space. Relying on text embeddings allows us to move towards a semantic-oriented filtering approach, reducing the dependency on single keywords, and thus on the ESG categorization adopted. This representation allows one to measure the semantic similarity between a document and a query by simply computing the distance, and contrarily the similarity, between their corresponding embedded vectors [82].

Our retrieval task involves discovering sentences related to each of the ESG categories (Sect. 3.1.2) within each sentence-segmented sustainability report. This implies working in a setting of asymmetric semantic search in which queries (ESG categories) and corpus documents (company sentences) are not interchangeable as they represent different semantic object types and have different lengths; similarly to the question answering framework [83, 84]. In contrast, symmetric semantic search is adopted when queries and corpus documents are interchangeable such as in similar document retrieval systems [86–88]. After extensive testing, we determine that "INSTRUCTOR-xl" [89], an instruction-tuned embedding model [90], is the most suitable choice for our specific tasks. The authors of the model [90] offer a universal instruction template ("Represent the [domain] [text type] for [task objective]:") along with an example list. Since we deal with asymmetric semantic search, the instructions provided to the model vary depending on the type of input. For embedding the ESG categories (queries), we use the instruction "*Represent the title for retrieving relevant statements*". To embed the sentences (corpus documents), the instruction employed is "*Represent the statement for retrieval*".

After generating the text embeddings for a sentence-segmented sustainability report, we retrieve the most semantically relevant sentences for each ESG topic (semantic search, Table 3). We set the similarity cut-off threshold t_{sim} equal to 0.6 to retrieve relevant sentences to a given ESG topic. Empirical experiments show an acceptable sentence relevance with a similarity threshold equal to or above 0.4. Thus, we adopt a cut-off point of 0.6 as a good trade-off between sentence coverage and computational workload. In addition, we

---

[5]Extrapolated from page 2 of Daikin's 2022 sustainability report.

Table 3 Example of the top three sentences selected for two ESG categories. The approach is capable of retrieving sentences that pertain to the two topics. However, it may also pick some meaningless sentences, such as the first one for the `Waste` topic, which comes from an infographic or a tabular layout

ESG topic	Sim score	Company sentence
Labor Practices	0.73	It also guarantees *the implementation of legislation on workers' rights* defining appropriate application standards …
	0.73	All Group workers, both in Italy and in Brazil, are covered by *Collective Labour Agreements* reached with trade union organizations and …
	0.72	In addition to the protections and rights provided by law and *the national collective labour agreement* for the sector …
Waste	0.68	*Total amount of waste by type* and disposal method 306-2 KPI A1.3 & KPI A1.4 Environmental Management
	0.64	At the same time, to *avoid unnecessary material loss and waste*, the Group has formulated loss rate standards for materials …
	0.68	Through continuous publicity and education on *garbage classification*, the project wastes were gradually reduced, recycled and harmless …

consider the top k sentences (with $k = 30$) to limit the number of retrieved sentences following the aforementioned trade-off.

This filtering layer also helps us reduce the computational time of the follow-up steps (e.g., the inference of the generative language model) and prune the resulting KG.
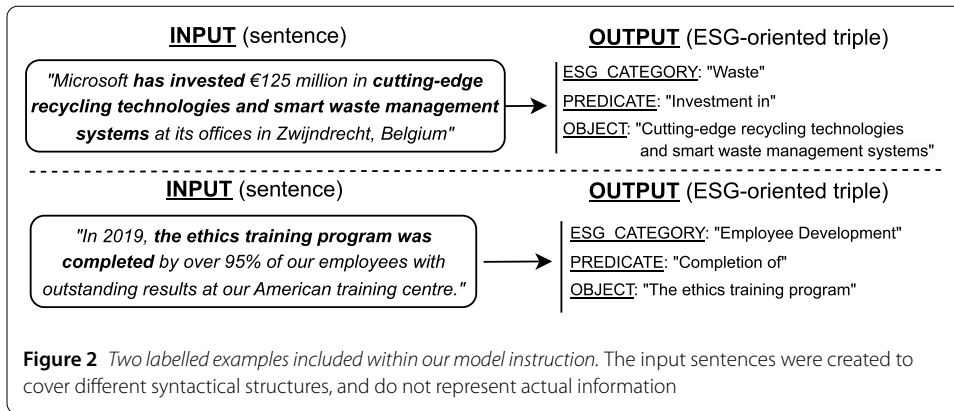
3.3 Retrieval-augmented triple generation

Our work aims to create a KG connecting companies, ESG topics, and actions disclosed by companies related to those topics. To achieve this goal, we need to represent ESG-related sentences in a unified and standardised format: triples with a predefined semantic template. Precisely, each ESG-oriented triple (`cat-pred-obj`) should consist of an ESG category[6] (`cat`) representing an ESG topic mentioned directly or indirectly in the sentence, a predicate affecting that category (`pred`), and an entity (`obj`) related to the ESG category undergoing the predicate. We define an action (`act`) as the concatenation between the ESG category (`cat`) and the predicate (`pred`) of a `cat-pred-obj` triple.

Consequently, our goal requires knowing the semantic meaning of words as well as defining a semantic template to generate ESG-oriented triples. The latest OIE techniques (Sect. 2.1) incorporate semantic information for extracting structured information, yet they rely only on the syntactical structure of the sentence. For example, given the sentence: "*Microsoft has invested 125 million in cutting-edge recycling technologies*",[7] conventional OIE techniques [91] would identify and generate a traditional SPO triple as the following: (`Microsoft, Invested, 125 million`). Although the above SPO triple can well represent the semantic meaning conveyed, it would not be suitable for our goal. Indeed, the ideal triple would have been: (`Waste, Investment in, Cutting-edge recycling technologies`).

Generating the latter requires defining a semantically-aware triple template. Firstly, the entity `Waste`, representing an ESG category, is not explicitly mentioned, although it could be inferred from the term `recycling technologies`. This type of inference jointly involves information extraction and semantic classification tasks. Secondly, our

---

[6]We use the terms "category" and "topic" interchangeably.

[7]This sentence is created for explanatory purposes, not reflecting real information.

Figure 2 *Two labelled examples included within our model instruction.* The input sentences were created to cover different syntactical structures, and do not represent actual information

goal requires extracting ESG-related actions rather than generic statements. Hence, triples should envelop predicates and objects related to an ESG category. For instance, given that ideal triple, the action (`act`) is defined as the ESG category `Waste` concatenated with the predicate `Investment in`, resulting in the action "`Waste:Investment in`".

LLMs have already demonstrated abilities in semantic understanding and handling a broad range of NLP-related tasks [10, 22]. Accordingly, in this work, we employ instruction-tuned LLMs, the In-Context Learning technique and the prominent RAG paradigm [18] to address this challenge. Our work exploits these techniques to provide an LLM with an input (ESG-related sentence) and an external context (input-output examples and a semantic schema) to extract structured information from the sentence.

We choose the `Kor` library [92] to create in-context instructions for LLMs. `Kor` allows to programmatically construct prompts by specifying the semantic data schema for the ideal triples (`cat-pred-obj`) as well as including labelled examples. A labelled example connects an input sentence with the desired output, an ESG-oriented triple. Figure 2 exhibits two labelled examples included in the model instruction to leverage the In-Context Learning abilities of the LLM.

In the model instruction, each element of the triple (`cat, pred` and `obj`) is declared with a unique name and a natural language description conveying its semantic meaning. The `Kor` library then uses this information to generate, by means of a predefined template (Fig. 3), a textual instruction to prompt an instruction-tuned LLM. We integrated the sixty-four ESG categories of the ESG categorization (Sect. 3.1.2) in the description of the attribute `cat`. The LLM leverages this list of ESG topics as semantic guidance, aiding itself in the generation of the ESG category for a given input sentence. It achieves this by mimicking the results of a supervised text classifier, whose labels are those of the ESG categorization, and extrapolating those labels with semantic generalization. Sect. SM2.1 of the Supplementary Material (SM) document exhibits the full model instruction.

We tested different instruction-tuned LLMs such as Google's Flan-T5 [93] and LLaMA-based models [94]. We empirically found that LLaMA-based models (e.g., Alpaca [95]) generate better results when prompted to extract structured information, with `WizardLM-7B` [16, 96] producing the highest-quality results. Appendix C and Appendix D showcase additional information regarding empirical experiments conducted on different LLMs, as well as the selection of specific hyperparameters for the LLM generation process.

```
Your goal is to extract structured information from the user's input that matches

the form described below. When extracting information please make sure it matches

the type of information exactly. Do not add any attributes that do not appear in the

schema shown below.

$DATASCHEMA

Please output the extracted information in JSON format. Do not output anything except

for the extracted information. Do not add any clarifying information. Do not add any

fields that are not in the schema. If the text contains attributes that do not appear

in the schema, please ignore them. All output must be in JSON format and follow the

schema specified above. Wrap the JSON in <json> tags.

$EXAMPLES

input: $INPUT

output:
```

Figure 3 *The instruction template used to prompt the generative LLM.* The instruction template is created and compiled by the `Kor` library to prompt an LLM to extract structured data using In-Context Learning. `DATASCHEMA` is a placeholder for the output data schema. `EXAMPLES` is a placeholder for the labelled examples in the format of input-output pairs. While `INPUT` represents an ESG-related sentence from which structured data needs to be retrieved
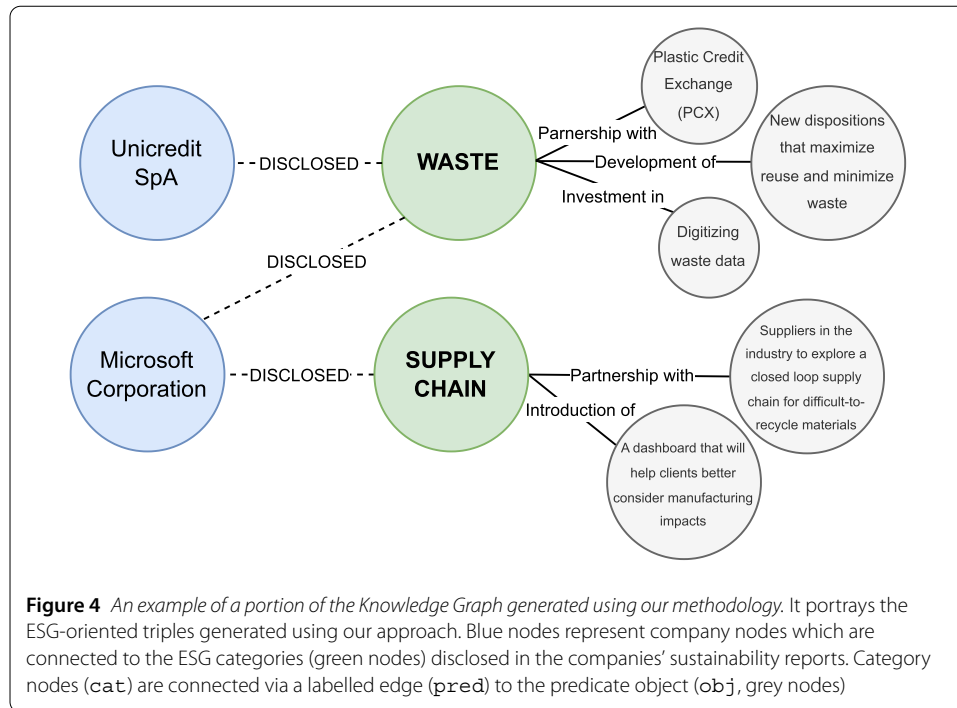
3.4 Knowledge graph generation

Before constructing a Knowledge Graph (KG) using the generated triples `cat-pred-obj`, we apply a data-cleaning process to reduce data redundancy. The redundancy in the KG comes from nodes and edges representing similar concepts and their relationships multiple times.

To achieve this goal, we perform semantic clustering on all the ESG categories (`cat`) and predicates (`pred`) included in the generated triples. Firstly, we generate text embeddings using the "INSTRUCTOR-xl" embedding model [90] with the model instruction "`Represent the title`". Secondly, semantic clusters are discovered as high-density regions in the embedded vector space using cosine similarity as a metric. We conducted several empirical experiments to evaluate the cluster goodness using different similarity cut-off thresholds, ranging from 0.5 to 0.9. Eventually, we adopt a similarity cut-off point of 0.8, as it strikes a good balance between the semantic coherence of the cluster elements (cluster quality) and the cluster sizes.

Finally, we label each cluster with its centroid and use cluster labels to replace the original ESG categories and predicates of the original triples. For instance, the predicate cluster labelled `Partnership with` groups 103 different predicates encompassing `Working together with`, `Partnering with others to` and `Collaborating of`. The Supplementary Material file reports further examples of these clustering operations (Sect. SM2.2). Figure 4 exhibits, for explanatory purposes, a portion of the KG generated using our methodology.

3.5 Approaches for statistical analyses

In the Results Sect. (Sect. 4), we mostly deal with undirected bipartite graphs obtained from the original KG. A bipartite graph is a graph whose vertices can be divided into two distinct and independent sets or partitions [14, 97, 98]. It can be described through its binary *bi-adjacency* matrix **B**, a $\{0, 1\}^{n \times m}$ matrix where n and m are the numbers of nodes in

Figure 4 *An example of a portion of the Knowledge Graph generated using our methodology.* It portrays the ESG-oriented triples generated using our approach. Blue nodes represent company nodes which are connected to the ESG categories (green nodes) disclosed in the companies' sustainability reports. Category nodes (`cat`) are connected via a labelled edge (`pred`) to the predicate object (`obj`, grey nodes)

the two partitions. A bipartite graph can consequently be seen as a special type of knowledge graph whose nodes can be divided into two distinct and independent partitions. Its graph edges are accordingly context-dependent and change based on the perspective used to generate the bipartite graph.

Specifically, we create three distinct bipartite graphs for the analyses of our findings through node and edge filtering of the comprehensive KG; though isolating distinct types of nodes, and their relative connections, from the original graph. The creation of different two-fold representations (bipartite graphs) can help conduct downstream analyses on specific relationships among different types of nodes included in the comprehensive knowledge graph. This allows us to analyse the extracted insights using three distinct perspectives:

1. the predicates (pred) disclosed with each ESG category (cat): analysed using the category-predicate bipartite graph $\mathbb{B}_{\text{catpred}}$;
2. the ESG categories (cat) disclosed by each company: analysed using the company-category bipartite graph $\mathbb{B}_{\text{cocat}}$;
3. the actions (act) disclosed by each company: analysed using the company-action bipartite graph $\mathbb{B}_{\text{coact}}$.

A table encompassing the number of partition nodes, the number of edges and the density for each bipartite graph is provided in the Supplementary Material (see Sect. SM3.2.1).

3.5.1 Bipartite graph statistics
Most of the unipartite graph metrics can be extended to the bipartite case [14, 98]. Specifically, here we compute the bipartite variants of network statistics such as degree centrality, closeness centrality and betweenness centrality [98, 99].

The degree centrality of a partition node is the fraction of the nodes of the other partition connected to it [100]. The closeness centrality of a node [97, 98] is determined by calcu-

lating its average shortest path distance to all other nodes. It represents the efficiency of a node to be connected directly to nodes from the other partition and indirectly to nodes from the same partition [101]. For instance given $\mathbb{B}_{cocat}$, a company node with a high closeness score indicates the company is connected, and thus it is close, to many category nodes which in turn are connected to several other company nodes. Lastly, betweenness centrality [97, 102] assesses the level of influence a node holds over information flow within a graph. In the context of bipartite graphs, it identifies nodes serving as critical mediators in enabling interactions between the two separate node partitions [14].

3.5.2 ESG-related actions' variability

We leverage information theory to assess the entropy of ESG topics based on their associated predicates yielded from companies' sustainability reports. Specifically, we adopt Shannon's entropy (Equation (1), [103]) to measure the information, and thus the variability, present in a set of events $\mathcal{X}$ through their respective occurrence probabilities $p(x)$:

$$H(\mathcal{X}) := -\sum_{x \in \mathcal{X}} p(x) \ln p(x). \tag{1}$$

In our context, the events $\mathcal{X}$ are the predicates disclosed by all companies for a given ESG topic, while $p(x)$ represents their relative occurrences. High entropy denotes high variability in the predicate occurrences, indicating an ESG topic addressed through many actions (predicates) with an almost uniform probability of predicate occurrence. On the other hand, low entropy indicates the predominance of a limited set of predicates for an ESG topic.

3.5.3 Similarity analysis

We estimate company similarities based on jointly disclosed ESG-related actions through the Jaccard similarity coefficient [104], which measures the similarity between two sets as the cardinality of their intersection over the cardinality of their union:

$$J(\mathcal{A}_{c_1}, \mathcal{A}_{c_2}) = \frac{|\mathcal{A}_{c_1} \cap \mathcal{A}_{c_2}|}{|\mathcal{A}_{c_1} \cup \mathcal{A}_{c_2}|}, \tag{2}$$

where $\mathcal{A}_{c_i}$ is the set of ESG-related actions disclosed by company c_i. To mitigate the influence of stochastic fluctuations on the similarity score, we generated a null model [105] with a bootstrapping technique by computing company similarities on the randomised action sets through 1000 simulations, and substracted this null model from the observed company similarities.

3.5.4 Correlation analysis

We evaluate whether company similarities in terms of jointly disclosed ESG-related actions (Sect. 3.5.3) are correlated to similarities in ESG scores or other company characteristics such as market capitalization or geographical location. We first measure feature similarities through different strategies, ensuring the same numerical range and monotonicity. The similarities in numerical features, such as ESG scores, are measured by computing the absolute numerical difference normalised using max-min scaling [106]. While, similarities in textual features, such as company sectors, are first embedded using the

"INSTRUCTOR-xl" embedding model [90], and then their semantic similarities are assessed through the cosine similarity normalised in $[0, 1]$ using min-max scaling [106]. A complete list of all the features and measures used can be found in the Supplementary Material (see Sect. SM2.4).

Afterwards, we perform a bivariate analysis through a correlation analysis to measure the monotonic association between action similarities and similarities of other company features. We rely on Kendall's τ correlation coefficient (Equation (3), [107]), a nonparametric and rank-based statistic computed as:

$$\tau = \frac{n_c - n_d}{n_c + n_d},$$
(3)

where n_c and n_d are the numbers of concordant and discordant pairs respectively. Rank-based correlation methods overcome some limitations of traditional correlation methods such as the well-known Pearson correlation coefficient [108]: they can measure nonlinear monotonic relationships, are more robust against outliers and normality assumption is not required [109]. High positive coefficients express a high level of order consistency in the company similarities sorted according to actions' and other similarities, while high negative coefficients occur when these two similarities are sorted reversely [107].

3.5.5 Interpretability of ESG scores

Lastly, we investigate the interpretability of ESG scores through linear regression and the SHAP (SHapley Additive exPlanations) framework [13]. Here, we investigate the most impacting factors on the ESG scores of companies by exploiting the interpretability of a first-order linear regression model.

The model predictors are based on our findings and other company information (Sect. 3.1.3). We first use as predictors the percentage of the top ten most disclosed ESG categories for each company. For example, if a company has ten percent of its generated triples concerning the ESG category `Waste`, and that is within its top ten most disclosed topics, the feature `Category:Waste` for this observation has a value of 0.1. We also consider the proportion of the E/S/G pillars based on all the disclosed categories for each company. In addition, we compute the category and action entropy for each company, indirectly representing the cardinality of the ESG categories and actions disclosed in its sustainability report (Sect. 3.5.2). Lastly, we consider nine company-related features as further predictors encompassing five company characteristics and four annual financial attributes. Specifically, we include in our predictors the company sector, the country, region and continent of its headquarters, and its incorporation year. On the other hand, we also include financial features concerning the fiscal year of the analysed sustainability report: EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortisation), liabilities, market capitalization and the number of employees.

The descriptive statistics of these report-based and company-based predictors are exhibited in Fig. 5 and Table 4. Categorical variables, such as the company sector, are turned into binary indicator variables (i.e., dummy variables [110]), generating a total of 97 features whereas standardisation is applied to those numerical such as market capitalization. An example of an observation, encompassing a complete list of features, is exhibited in Sect. SM2.5.1 of the Supplementary Material (SM) document.
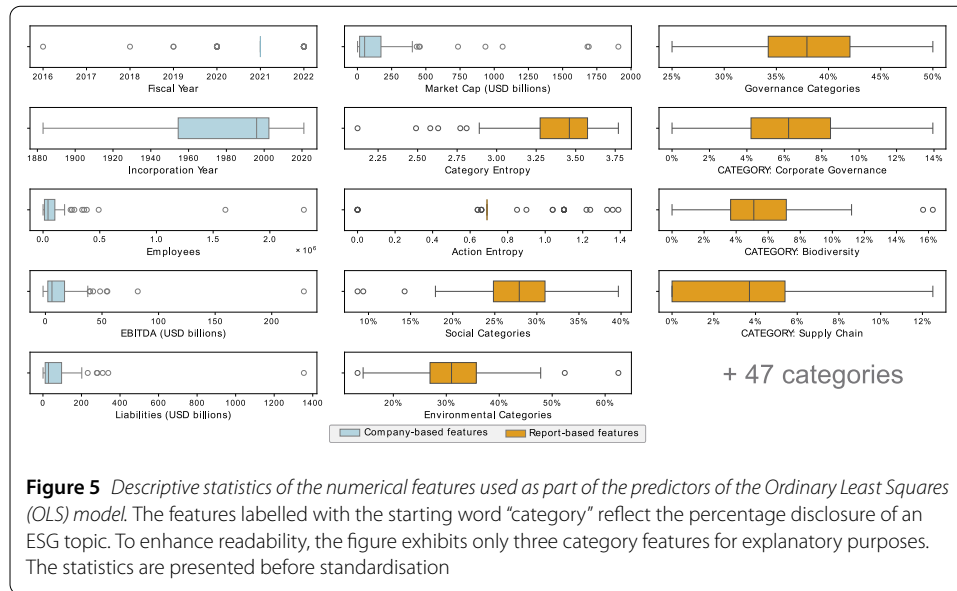
Figure 5 *Descriptive statistics of the numerical features used as part of the predictors of the Ordinary Least Squares (OLS) model.* The features labelled with the starting word "category" reflect the percentage disclosure of an ESG topic. To enhance readability, the figure exhibits only three category features for explanatory purposes. The statistics are presented before standardisation

Table 4 Descriptive statistics of the categorical variables used as part of the predictors of the Ordinary Least Squares (OLS) model. All these variables are transformed into binary indicator variables for each observation

Feature name	Unique	Top frequent categories (%)
Region	3	Americas (56.3%), Europe (29.9 %), Asia (13.8%)
Sub-region	7	Norther America (55.2%), Western Europe (13.8%), Eastern Asia (12.6%), …
Country	15	United States of America (51.7%), Italy (8%), United Kingdom (6.9%), …
Sector	11	Industrials (14.9%), Information Technology (11.5%), …

Then, we adopt an Ordinary Least Squares (OLS, [111]) regression with Elastic Net Regularization [112] to perform inference on companies' ESG scores available. We adopt Elastic Net Regularization, a generalisation of the LASSO method [113], to perform both feature selection and training regularisation. We choose this regularisation method as it improves performance when the number of predictors (|*features*| = 97) is higher than the observations (|*companies*| = 89) as well as in the presence of strong pairwise correlations [112]. We train the OLS model using the Elastic Net cost function and an eight-fold cross-validation approach [114] (see Sect. SM2.5.3). The performance of the OLS model is discussed in Appendix E.

Subsequently, we employ the SHAP framework [13] to investigate which predictors impact the inference of ESG scores most. SHAP is a model-agnostic and additive feature importance measure which is based on cooperative game theory. It provides local interpretations of model predictions as additive sums of the directed effects of each predictor using a conditional expectation function. SHAP starts from the prior knowledge of the expected model output $E[f(X)]$, and then evaluates, for each model prediction, the magnitude and direction changes in this expected value (SHAP values) when conditioned on each predictor. It thus quantifies the magnitude and direction of the observed effects of each predictor.

Predictors with positive SHAP values affect the expected model output with additive increments, conversely, those with negative values have additive decrement impacts.

4 Results

In this section, we first report the network statistics computed at the node level from the three bipartite graphs in Sect. 4.1. Next, a diversity analysis examines how ESG topics are disclosed across companies and different sectors (Sect. 4.2). Section 4.3 reports company similarities according to jointly disclosed ESG actions. The follow-up section (4.4) addresses whether these company similarities are associated with similarities in other company information. Finally, we evaluate the interpretability of ESG scores by investigating the most impacting factual aspects (Sect. 4.5). A qualitative analysis of the generated triples and an ablation study on the model instruction are exhibited in Appendix A and Appendix B.

4.1 Bipartite graphs' analysis

We here present some network statistics concerning the three bipartite graphs $\mathbb{B}_{cocat}$, $\mathbb{B}_{catpred}$, and $\mathbb{B}_{coact}$. Further statistics and extensive tables for all three bipartite graphs are shown in the Supplementary Material document in Sect. SM3.2.

The ESG movement encompasses many socially responsible issues Our data-driven methodology unveils that the one hundred and twenty-four companies disclosed 542 distinct ESG topics/categories in their sustainability reports. The company-category bipartite graph $\mathbb{B}_{cocat}$ has an average degree distribution equal to almost 11%, making this graph relatively connected. There are however some mainstream ESG categories: `Climate Risk Management`, `Supply Chain`, `Energy` and `Corporate Governance` are connected to, and thus disclosed by, almost all the considered companies (degree > 92%, Table 5). Conversely, the list of the least disclosed categories is rather surprising: `Market Responsibility`, `Anti-Discrimination` and `LGBTQ+ Inclusion` are connected to, and thus disclosed by, less than 5% of all the considered companies (Table 5).

ESG topics are addressed from several perspectives, with some frequent ones   The average degree centrality of the category-predicate bipartite graph $\mathbb{B}_{catpred}$ is less than 1%. There are however some predominant predicate nodes (see Sect. SM3.2.3) that are associated with more than ninety ESG categories (degree ≥ 16.6%) such as `Commitment and involvement with` (113 categories), `Advisor support for` (102), `Partnership with` (97), and `Establishment of` (94). These prominent nodes interestingly exhibit high closeness centrality (> 88%) in contrast to their relatively low degree centrality values. This indicates that common category nodes indirectly connect them.

Table 5 Graph metrics of a sample of all the 542 category nodes of the bipartite graph $\mathbb{B}_{cocat}$

Category node (cat)	Degree (%)	Closeness (%)	Betweenness (%)
Supply Chain	95.3	98.9	2.3
Climate Risk Management	94.5	98.5	2.3
Corporate Governance	93.0	98.1	2.1
Community and Society	91.4	98.1	2.1
Philanthropy	88.3	96.8	2.0
Packaging	50.0	77.7	0.6
Human Resources	39.8	74.2	0.4
LGBTQ+ Inclusion	4.7	55.4	0.0
Anti-Discrimination	2.3	52.1	0.0
Marketing Responsibly	0.8	47.8	0.0

Common actions are the exception   The company-action bipartite graph $\mathbb{B}_{coact}$ connects the company nodes to almost twenty thousand different ESG-related actions (19,574) disclosed in companies' sustainability reports. However, there are a few prominent actions disclosed by the majority of the considered companies: `AIR EMISSION: Reduction of` (degree of, and thus connected to, the 70% of the companies), `ENERGY:Reduction of` (61%), `PHILANTHROPY:Donation by` (60%) and `CLIMATE RISK MANAGEMENT:Assessment of` (56%).

4.2 Diversity analysis on disclosing ESG categories

In this section, we analyse the differences in disclosing and addressing ESG topics in companies' sustainability reports.

Companies approach ESG topics from plenty of perspectives, especially those generic and vague   The predicates associated with each ESG category vary significantly with an average Shannon's entropy of 1.5 nats. This broad action variety is predominant in generic and umbrella ESG topics such as `Corporate Governance`, `Human Rights`, and `Supply Chain` (Table 6). For example, when addressing `Product Safety`, companies approach it from different perspectives, ranging from developments (2.3%) to regulatory compliance (2.3%) and assessments (4.4%, Table 6). However, a notable correlation (*corr* = 0.84) exists between the entropy of categories and the number of companies disclosing information about a particular ESG category. This suggests differences in how each company approaches these ESG topics.

Cross-sector vs sector-focused topics   A proportion of almost 12% among all the 542 ESG categories is disclosed across all company sectors, encompassing various umbrella aspects such as `Climate Risk Management`, `Supply Chain` and `Business Ethics`. However, it is noteworthy that certain sectors emphasise specific topics more than others (Table 7). For example, `Packaging` is more emphasised in Consumer Staples companies, such as PepsiCo (18% of all the company triples), Coca-Cola (9%), Monster Beverage (6%), and Tesco (5%). This category accounts for 4.9% of all generated company triples within this sector as exhibited in Table 7. Another example is `Water` which is more stressed

Table 6 A sample of the ESG categories with their entropy values computed. The three most frequent category predicates are reported alongside the percentage of companies disclosing that category

ESG category (`cat`)	Entropy (nats)	Companies (%)	Category predicates (`pred`)
Supply Chain	5.72	95	Partnership with (3.7%), Assessment of (2.4%), Engagement in (2.2%)
Corporate Governance	5.62	94	Establishment of (4.3%), Commitment and involvement with (3.7%), Overseeing (2.9%)
Human Rights	5.62	90	Commitment and involvement with (5.1%), Respect for (2.6%), Assessment of (2.4%)
Product Safety	5.27	69	Assessment of (4.4%), Compliance with (2.3%), Development and implementation of (2.3%)
Food Waste	2.44	6	Commitment and involvement with (14.3%), Reduction of (14.3%), Development of (7.1%)
Anti-Slavery Practices	1.61	4	Undertaking of (20%), Integration of (20%), Required training (20%)

Table 7 Sample of all the ESG categories disclosed by companies in their sustainability reports. This table exhibits the category coverage through different percentages concerning: (i) the company triples including a category, (ii) the companies reporting a category (iii) also aggregated by sector, and (iv) the company triples aggregated by sector

ESG category (cat)	Triples (%)	Companies (%)	Sectors (%)	Triples per company sector (%)
Corporate Governance	6.7	94	100	Industrials (7.8%), Financials (7.5%), Health Care (7.4%)
Air Emissions	3.7	92	100	Energy (6.9%), Materials (4%), Industrials (3.8%)
Water	3.1	85	100	Consumer Staples (6.6%), Materials (4.9%), Energy (4.4%)
Green Buildings	1.5	88	100	Real Estate (3.5%), Consumer Discretionary (1.8%), Information Technology (1.6%)
Packaging	0.9	49	100	Consumer Staples (4.9%), Consumer Discretionary (1.5%), Information Technology (0.9%)
Business Ethics	0.3	61	100	Health Care (0.7%), Utilities (0.5%), Industrials (0.4%)

by companies operating in water-intensive sectors such as Consumer Staples (6.6%, Table 7, e.g., Coca-Cola) and Materials (4.9%, e.g., DuPont) rather than Financials (0.5%, e.g., Goldman Sachs).

Comparison with ESG materiality at the sector and industry level We compare the data-driven evidence of sector-focused ESG topics with the sector-level ESG materiality matrix (Sect. 2.3) identified by Khan *et al.* [35]. We explore the relevant issues identified for the Financials sector for explanatory purposes as is one of the common sectors between the authors' work and ours. There are thirteen financial companies in our sample: nine commercial banks (e.g., Deutsche Bank), two credit service companies (e.g., Mastercard), one insurance company (Assicurazioni Generali) and one asset management company (3i Group). Our data-driven findings unveil that the companies' sustainability reports address the ten issues identified as relevant for this sector with different importance. Financial companies, for example, extensively disclosed actions concerning: *environmental, and social impacts on core assets and operations* (Climate Risk Management: 7% of all the sector triples), *Diversity and Inclusion* (Financial Inclusion: 6.1% and Board Diversity: 5.2%) and *Lifecyle impacts of products and services* (Product Sustainability: 4.1%). On the other hand, companies' non-financial disclosures neglect to address some relevant issues encompassing: *Access and affordability* (Access to Basic Services: 0.8%, Accessibility: 0.3%, and Access to Information: 0.1%), *Fair marketing and advertising* (Marketing and Advertising: 0.1%), and *Business ethics and transparency of payments* (Business Ethics: 0.2% and Anti-Money Laundering: 0.1%). A table exhibiting the comparison with all ten issues can be found in Sect. SM3.6 in the Supplementary Material (SM) document.

Furthermore, we select UniCredit, a financial company operating in the industry of Commercial Banks, as an explanatory example to compare our findings with the relevant disclosure topics outlined in SASB standards for its industry. This reporting standard organization identifies six important disclosure topics for commercial banks [115] which were addressed differently in the company's sustainability report. The company focused on industry-relevant issues encompassing: *Financial Inclusion & Capacity Building*

(Financial Inclusion: 4.6%) and *Data Security* (Data Privacy: 2.4%). It however neglected to disclosure much information concerning issues such as *Financed Emissions* (Product Design: 0.4%) and *Incorporation of Environmental, Social, and Governance Factors in Credit Analysis* (Environmental Risk Assessment: 0.4%).

Different sectors employ tailored actions to address ESG topics   Although there are a few widely disclosed actions (Sect. 4.1), the same action is disclosed, on average, by less than 2% of the considered companies. Whereas, only 15% of the company sectors are, on average, engaged in the same action. This unveils different priorities and a variety of approaches among companies and sectors. For example, the Assessment of aspects concerning Climate Risk Management are more emphasised by Real Estate companies which manage assets vulnerable to climate risks, such as Park Hotels Resorts (2% of all the company triples) and Sun Communities (1%). Conversely, Materials companies such as United States Steel (1%), Yamana Gold (1%) and Aluminum Corporation of China (1%), emphasise instead the Commitment and involvement concerning Employee Safety, a worker-related topic. Further extensive tables are in the Supplementary Material document (see Sect. SM3.1).

4.3 Company similarities based on disclosed ESG-related actions

Here, we discuss company similarities according to jointly disclosed actions using the Jaccard similarity coefficient (see Sect. 3.5).

Companies from the same sectors tend to perform similar actions   For example, as depicted in Fig. 6 and outlined in Table 8, five companies among the top ten similar companies of Deutsche Bank are banks too: Royal Bank of Canada (action similarity equal to 7%), Banco Santander (6%) and UniCredit (6%). Notably, Visa and Mastercard, both operating in the Credit Services industry, form a distinct group (Fig. 6).
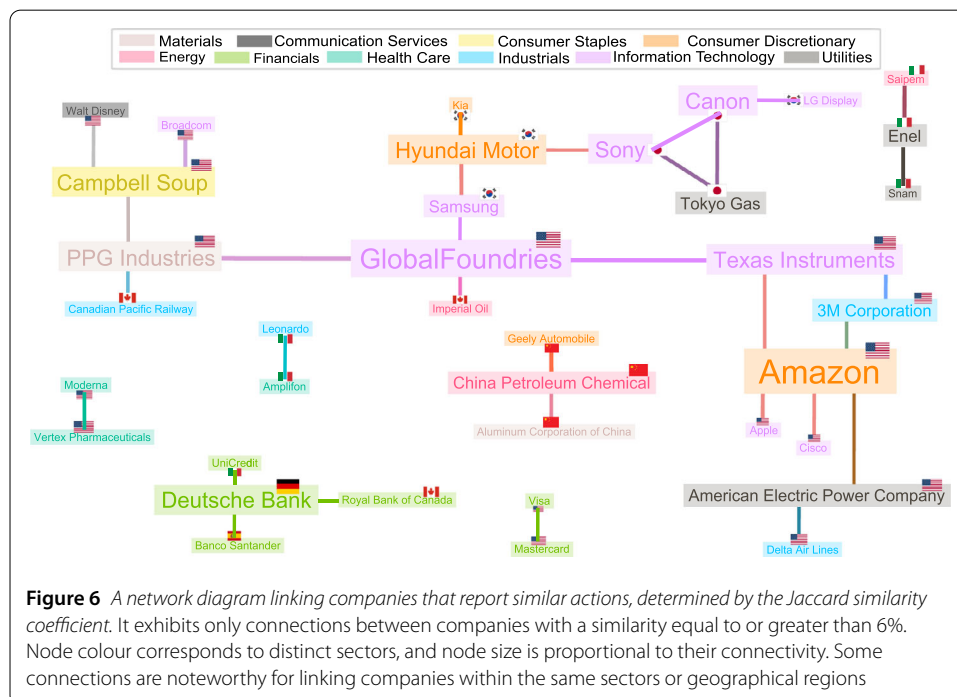

Figure 6 *A network diagram linking companies that report similar actions, determined by the Jaccard similarity coefficient.* It exhibits only connections between companies with a similarity equal to or greater than 6%. Node colour corresponds to distinct sectors, and node size is proportional to their connectivity. Some connections are noteworthy for linking companies within the same sectors or geographical regions

Table 8 A company sample with the top three most reported actions and the most similar companies for each. Company similarities are assessed by computing the Jaccard similarity on the companies' disclosed action set

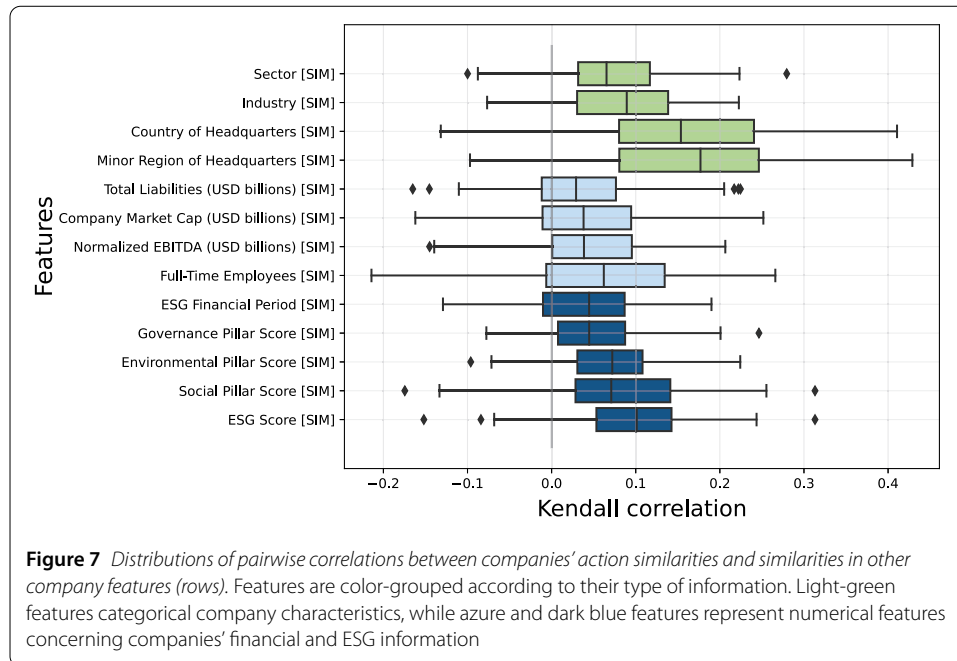| Company	Most reported actions	Most similar companies
Sony	PACKAGING: Reduction of (x7) PHILANTHROPY: Advisory support for (x6) CORPORATE GOVERNANCE: Establishment of (x6)	Canon (7%), Tokyo Gas (6%), Hyundai Motor (6%), Toshiba (6%), Kia (6%)
Deutsche Bank	ENERGY: Reduction of (x4) BIODIVERSITY: Promotion of (x4) CORPORATE GOVERNANCE: Establishment of (x4)	Royal Bank of Canada (7%), Banco Santander (6%), UniCredit (6%)
Global-Foundries	WATER: Use of (x6) AIR EMISSIONS: Reduction of (x7) PHILANTHROPY: Donation by (x7)	Texas Instruments (7%), PPG Industries (7%), Samsung (6%), Visa (6%)
Geely Automobile	PHILANTHROPY: Participation in (x5) SUPPLY CHAIN: Establishment of (x5) CORPORATE GOVERNANCE: Development and implementation of (x5)	China Petroleum Chemical (7%), Baidu (6%), LG Display (6%), Alibaba (5%), Korean Air Lines (6%)
Saudi Aramco	AIR EMISSIONS: Reduction of (x4) BIODIVERSITY: Protection of (x4) ENERGY: Investment in (x4)	Tokyo Gas (5%), Royal Dutch Shell (5%), Yamana Gold (5%), Visa (5%)
Philip Morris	WASTE: Continuous efforts to reduce, reuse, or recycle (x5) BIODIVERSITY: Continuing to set goals and work towards (x4) PHILANTHROPY: Investment in (x3)	Croda (4%), 3M (4%), Coca-Cola (4%), GlobalFoundries (4%) Samsung (4%), United States Steel (4%)

Comparably, action similarities emerge in companies operating in the Health Care sector: Moderna, Vertex Pharmaceuticals and AstraZeneca (Table 8). Further details are visible in Sect. SM3.4 of the Supplementary Material document.

Companies from the same geographical region tend to perform similar actions   For example, as can be visually noted in Fig. 6, 80% of the ten most similar companies of Sony (Japan, Eastern Asia) are companies from the same geographical region: 40% from Japan and 40% from South Korea. Similarly, Geely Automobile (China, Eastern Asia) has 70% of its ten most similar companies from the same region too: 40% from China and 30% from South Korea. On the west side, there are six European companies in the ten most similar companies of Enel (Italy, Southern Europe) with Italian companies representing 40% of the total.

4.4 Correlation analysis among company similarities

This section answers the research question concerning whether company similarities in terms of jointly disclosed actions (Sect. 4.3) are associated with similarities in other company information (Sect. 3.1.3). We perform a bivariate correlation analysis using Kendall's correlation coefficient (Sect. 3.5) for each company with all information available: 81% of the considered companies. Aggregated results are exhibited in Fig. 7 through box plots.

Similarities in companies' disclosed actions are correlated with companies' geographical regions   Action similarities have the highest, yet weak, correlation with the Region and Country of the company headquarters, with a median correlation coefficient of 0.18 and 0.15. This confirms the empirical findings discussed in the previous section (4.3). Moreover, only these two features demonstrate median p-values, resulting from the null hypothesis test of zero monotonic correlation, below the established accepting threshold of

Figure 7 *Distributions of pairwise correlations between companies' action similarities and similarities in other company features (rows).* Features are color-grouped according to their type of information. Light-green features categorical company characteristics, while azure and dark blue features represent numerical features concerning companies' financial and ESG information

5%, respectively 1% and 2%. The p-value distributions for all the features are shown in Sect. SM3.7.1 of the Supplementary Material (SM) document. Taking the previous example companies, `Sony` and `Enel` exhibit a relatively high monotonic correlation between their company similarities in terms of disclosed actions and geographical locations. `Sony` has an action-country similarity correlation equal to 0.22 and an action-region correlation equal to 0.20, while `Enel` exhibits a lower action-country correlation (0.14) and a higher action-region correlation (0.25).

No other statistically significant similarity correlations emerge   Company similarities in `ESG scores` and `Industries` have a median pairwise correlation with companies' disclosed actions equal to 0.1 and 0.09. Their statistical significance appears however relatively weak due to high median p-values (13% and 15%), which suggest accepting the null hypothesis of zero monotonic correlation.

ESG scores are only correlated with their underlying components   After analysing company similarities from the disclosed action perspective, we perform a pairwise monotonic correlation analysis to unveil possible confounding factors for company similarities. Strong monotonic correlations appear between similarities in companies' `Region` and `Country` (median correlation equal to 0.7) as well as between companies' `ESG score` and their `Social` (0.5) and `Environmental Pillar scores` (0.4). No other relevant correlations emerge for ESG scores or other company information. A graphical representation of all the pairwise correlations is exhibited in Sect. SM3.7.2 of the Supplementary Material (SM) document.

4.5 Interpretability of ESG scores

Lastly, we investigate the interpretability of companies' ESG scores by employing a first-order linear regression and the SHAP framework (Sect. 3.5.5). We specifically evaluate

how various factual and corporate aspects impact these scores using features based on the companies' extracted actions (such as the most disclosed ESG topics), and additional financial and company-specific information.

Social-related actions and company transparency have a significant impact on ESG scores On average, the most impacting aspects affecting ESG scores are the percentages of disclosed actions related to `Human Rights` and `Employee Development`, with a mean SHAP value of 2.6 and 2.7. High percentages of the former (colour scale in Fig. 8) positively impact ESG scores, while the latter has the effect of hurting them. High disclosing percentages in actions related to `Philanthropy` (mean SHAP value of 1.1) and `Energy` (0.7) also hurt companies' scores. Similar average magnitude, but an opposite effect, disclosing several actions related to `Waste` (0.8) or `Supply Chain` (0.7) has a positive impact. Furthermore, high variety in the disclosed ESG topics (`Category Entropy`)
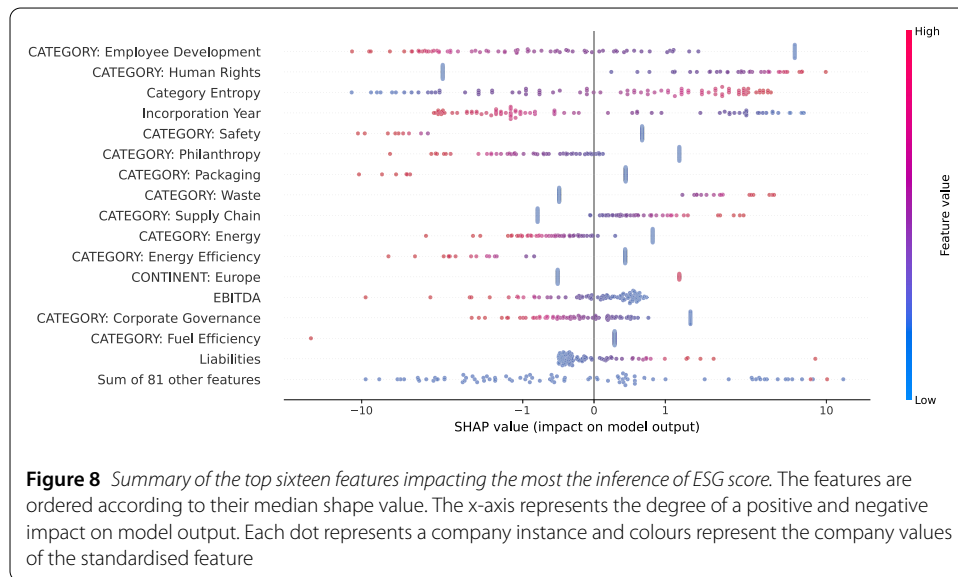
Figure 8 *Summary of the top sixteen features impacting the most the inference of ESG score.* The features are ordered according to their median shape value. The x-axis represents the degree of a positive and negative impact on model output. Each dot represents a company instance and colours represent the company values of the standardised feature
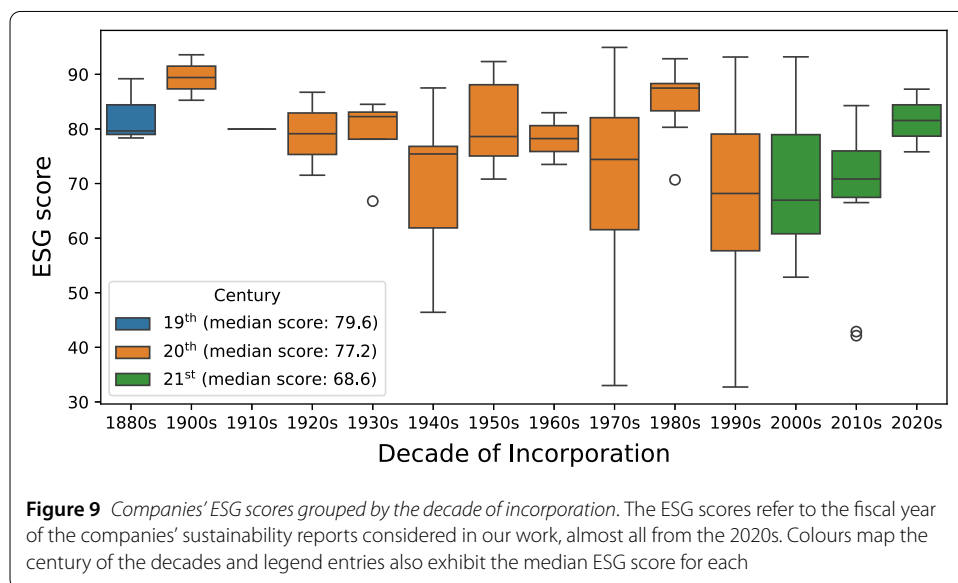

Figure 9 *Companies' ESG scores grouped by the decade of incorporation.* The ESG scores refer to the fiscal year of the companies' sustainability reports considered in our work, almost all from the 2020s. Colours map the century of the decades and legend entries also exhibit the median ESG score for each

positively affects ESG scores with a mean SHAP value of 2.1 (Fig. 8). Sharing a similar average magnitude (1.9), being founded earlier, represented by an older `Incorporation Year`, positively impacts a company's score. This is also validated numerically by a negative Kendall correlation equal to −0.22 (p-value of 0.5%) and visually by Fig. 9 which groups the companies' ESG scores by their decade of incorporation. The median ESG score of the fifty-three companies funded in the 20th century is 77.2, whereas the thirty-one companies established in the current century exhibit a lower median score of 68.6. Notably, the three companies founded in the 19th century exhibit the highest median ESG score equal to 79.6.

Further noteworthy factors positively impacting ESG scores are being a European company (`CONTINENT:Europe`, mean SHAP value of 0.7) and exhibiting a high level of `Liabilities` (0.5). In contrast, high annual earnings (`EBITDA`, 0.6) have a slight negative impact on ESG scores. The positive influence of being European can be also validated by grouping ESG scores by company region: the twenty-six companies from Europe exhibit the highest average ESG score equal to 82, the forty-nine American companies have an average ESG score of 69.7, whereas the average ESG score of the twelve Asian companies is equal to 67 (see Sect. SM1.5 in the Supplementary Material document).

Local interpretability analysis reveals company-specific impacting factors   Moving from global to local interpretability, we choose `Sony` as an example company and investigate the most impacting factors for its ESG score (Fig. 10). The `Incorporation Year` and `Category Entropy` features, respectively far below (standardised value of −1.03, representing the year 1946) and above (0.87, 3.6 nats) the average of company values, positively affect its score. In addition, disclosing several actions related to `Human Rights` (0.57, 4.4% of all its extracted actions) and `Waste` (1.31, 3.4%) has a positive impact. In contrast, disclosing fewer `Energy`-related actions than the average (−0.98, not among its
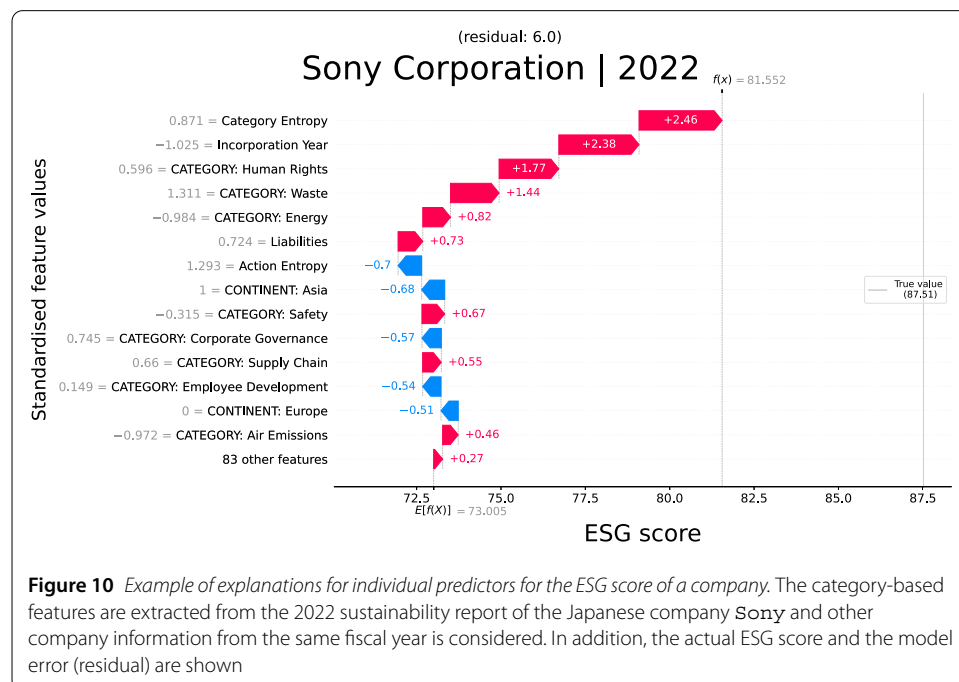

Figure 10 *Example of explanations for individual predictors for the ESG score of a company.* The category-based features are extracted from the 2022 sustainability report of the Japanese company `Sony` and other company information from the same fiscal year is considered. In addition, the actual ESG score and the model error (residual) are shown

top ten disclosed topics) positively affects its scores. Furthermore, being an Asian company (`CONTINENT:Asia` and `CONTINENT:Europe`) slightly hurts its ESG score.

Region-based interpretability analysis reveals common patterns   We also conduct a more granular analysis by exploring the ESG score interpretability of a company cluster. Coherently with the example company, we select Asian companies encompassing five Chinese companies (e.g., Alibaba), five Japanese companies (e.g., Sony), Aramco (Saudi Arabia) and Greely Automobile (Hong Kong). The `Incorporation Year` strongly affects their ESG scores, with an average SHAP value of 2.7, in line with the global interpretation. This is further validated by a strong negative Kendall correlation of –0.61 (p-value of 0.7%). All companies established in the 20th century, such as Toshiba (1904, score of 93.6), Toyota (1937, score of 84.5) and Geely Automobile (1946, score of 75.4), exhibit ESG scores above 66. Whereas, those established in the current century have all lower scores such as Baidu (2000, score of 53.5), China Evergrande (2006, score of 52.8) and Aramco (2018, score of 42.9). From the geographical point of view, being a Chinese company (`COUNTRY:China`) hurts ESG scores with an average SHAP value of 0.6. This is further emphasised by the observation that all the Chinese companies consistently have ESG scores below 62.5, whereas both Hong Kong-based and Japanese companies consistently display higher scores. Furthermore, this analysis confirms that disclosing several actions related to `Human Rights` (e.g., Toshiba and Tokyo Gas) and `Waste` (e.g., Toyota and Sony) positively impact ESG scores, whereas their absence negatively affects them (e.g., Baidu and Daikin Industries). The ESG scores group by region, a details list of the considered Asian companies and the bee-swarm graph of the latter analysis are reported in the Supplementary Material (see Sects. SM1.5, SM3.8 and SM3.9).

5 Discussion

Now, we address the practical implications of our findings (Sect. 5.1) as well as the methodological implications of our proposed approach (Sect. 5.2). Lastly, we discuss some potential limitations of our work in Sect. 5.3.

5.1 Practical implications

High action variety in addressing ESG topics   As highlighted in Sect. 4.1 and 4.2, companies address ESG topics from many perspectives, ranging from recognition and commitments to developments, partnerships and compliance. This foregrounds the complexity and joint efforts needed to address ESG-related aspects and the involved external subjects such as ESG rating agencies. Our analysis unveils that the same action is disclosed, on average, by only 2% of the companies, and by only 15% of all the company sectors, confirming a lack of a common approach across companies and different sectors. However, some ESG topics are addressed through a common strategy by the majority of the considered companies: the actions `Air Emission:Reduction of` and `Energy:Reduction of` are disclosed by 70% and 61% of the companies (Sect. 4.1).

The ESG phenomenon has blurred boundaries and includes plenty of socially responsible topics   Concerning the ESG topics disclosed by companies, our methodology extracts 542 distinct ESG topics from companies' sustainability reports, representing an eight-times greater set of topics originally included in the ESG categorization exploited as a

semantic reference in this work (sixty-four categories, Sect. 3.1.2). Firstly, this unveils the broad scope of the ESG phenomenon involving socially responsible topics ranging from `Waste Management` and `Supply Chain` to `Employee Safety` and `Tax Compliance`. Secondly, this highlights the presence of widely disclosed topics, such as `Supply Chain`, and sector-focused topics such as `Packaging` and `Water` for the Consumer Staples sector. The diversity analysis reported in Sect. 4.2 confirms this sector-based importance for certain topics. For instance, the ESG topic `Water` is more stressed by water-intensive companies, while `Packaging` is more emphasised by companies producing consumer staples (Table 7). This data-driven insight is also validated by the ongoing discussions in the ESG literature concerning ESG Industry Materiality (Sect. 2.3). Furthermore, by comparing with sector- and industry-level ESG materiality matrices (Sect. 4.2), we can assess companies' disclosures against the expected sustainability issues in their reports, highlighting variations in topic coverage.

Exogenous factors might influence companies' non-financial disclosures   The findings reported in Sect. 4.3 emphasise company similarities based on their sectors, confirming indirectly a relatively high presence of common strategies among companies from the same sector. However, the most impacting factor in grouping companies based on their disclosures is their geographical region as shown in Sect. 4.3 and Sect. 4.4. This represents an interesting finding from our data-driven work, validating the ongoing discussions in the literature concerning the influence of the company's geographical origins on their non-financial disclosures (Sect. 2.3). For example, other studies have analogously unveiled the impact of exogenous factors on these disclosures: encompassing region-specific regulations [62] to the political, labour, and cultural environment of the company's nation [32, 33, 63].

Companies' social and environmental performance hold greater importance than the governance performance in the combined ESG scores   The bivariate correlation analysis reported in Sect. 4.4 shows that similarities in ESG scores are neither associated with similarities in disclosed actions nor other financial or company characteristics, representing a noteworthy finding of our work. It however unveils strong monotonic correlations between similarities in the companies' region and country as well as between the ESG score and the social and environmental pillar score (Sect. 4.4). These two appear fairly trivial associations: first, the ESG score is a weighted score combining the scores of the three E/S/G pillars (Sect. 3.1.3); second, the region and country have a natural geographical relation. However, the monotonic associations of ESG scores could be exploited to roughly infer the average influence, and thus the importance, of the E/S/G pillar scores towards the combined score. For instance, a weak or zero monotonic correlation suggests that (dis)similarities among scores of one specific pillar are not associated with (dis)similarities in the combined score. This could imply a particular pillar holds relatively less importance, or weight, in determining the combined ESG scores. Conversely, when a significant pillar is present, its (dis)similarities reflect the (dis)similarities of the combined ESG scores. Hence, based on the monotonic associations of ESG scores, it could be inferred that, on average, the social pillar (0.5) holds slightly greater importance compared to the environmental pillar (0.4). In contrast, corporate governance bears minimal importance in ESG scores

(0.2). Although the company sample and the fiscal years considered may influence the inference of the rating agency's methodology, insights on E/S/G weightings can be helpful to validate the impacting factors of ESG scores unveiled in Sect. 4.5.

Exogenous factors can influence the quality of companies' non-financial disclosures, indirectly impacting their ESG performance assessments   The interpretability analysis of ESG scores (Sect. 4.5) highlights that the company's disclosures impact ESG scores more than other financial aspects or company characteristics. Disclosing several ESG topics (`category entropy`) positively affects ESG scores, whereas fewer disclosed ESG topics hurt scores. This data-driven insight accordingly confirms that transparency on non-financial information rewards companies' ESG assessment [63, 66]. The analysis of Sect. 4.5 also confirms the negligible impact of governance-related topics towards ESG scores in comparison to social- and environmental-related topics such as `Human Rights`, `Energy` and `Waste`. The findings of Sect. 4.5 also unveil that disclosing many actions related to the ESG topics `Employee Development` and `Energy` negatively impact ESG scores. One hypothesis is that their high presence in a company's sustainability report might reveal a poor coverage of other important ESG topics [76]. For example, in the local interpretation of Sony's ESG score, a low disclosing percentage of `Energy`-related actions and a high disclosing percentage of `Waste`-related actions positively impact its score. Another noteworthy finding of this analysis is the region's impact on ESG scores: being a European company positively impacts companies' ESG scores, whereas being a Chinese company hurts them (Sect. 4.5). The Chinese penalization factor might be reinforced by the fact that all the considered Chinese companies were funded between 1999 and 2006, likely due to the remarkable economic development of this region starting in the 2000s, and thus associated with the negative impact of being relatively young companies (Sect. 4.5). Indeed, our interpretability analysis also uncovers a beneficial effect associated with earlier incorporation years. However, additional investigation is necessary to exclude spurious correlations. The impact of the company's region on ESG scores is however coherent with the region-based disclosing similarity previously highlighted and validated by ongoing discussions in the ESG literature concerning a European bias (Sect. 2.3). Further validation of the positive impact of being a European can be found by delving into the environmental performance of European companies in our sample: they demonstrate the highest average environmental score of 81.8, surpassing Asian companies with an average score of 69 and American companies with the lowest average score of 65.6 (see Sect. SM1.5 in the Supplementary Material document).

On the other hand, other studies suggest that companies' ESG assessments might indirectly suffer from biases based on company size (Sect. 2.3). The interpretability analysis in Sect. 4.5 unveils negligible evidence of this company size bias: a greater number of full-time employees has a positive, yet marginal, impact on ESG scores (average SHAP value of 0.24), whereas the company's market capitalization has no impact on interpreting companies' ESG scores. In addition, the Kendall pairwise correlations of company similarities (Sect. 4.4) concerning these two company variables and similarities in ESG scores are not statistically significant: a monotonic correlation of 0.1 for the number of employees (p-value of 71%), and a zero correlation for the market capitalization (p-value of 78%).

5.2 Methodological implications

As mentioned in Sect. 3.3, generative LLMs provide us with the semantic understanding and flexibility needed to overcome the limitation of traditional OIE approaches which rely only on the syntactical sentence structure. However, employing a 7-billion LLM [16] for information extraction (see Retrieval-Augmented Triple Generator in Fig. 1) leads to a higher computational load. Nonetheless, this allows us to generate semantically aware and ESG-focused triples instead of traditional SPO ones. This is pivotal in generating all the meaningful ESG-related insights of our work.

Utilising generative LLMs and the ESG categorization for semantic guidance enhances retrieving more comprehensive data-driven insights   The flexibility and generative abilities of these generative language models also allow us to highlight, and overcome some limitations in the data sources such as those of the ESG categorization used in our work (Sect. 3.1.2). This categorization extrapolates a concise set of ESG topics by categorising several ESG-related indicators shared among ESG rating providers. It accordingly derives the scope of the ESG phenomenon from the perspective of rating agencies, a different viewpoint compared to companies' disclosures analysed in our work. However, these different points of view, and the methodology based on generative LLMs, help us to unveil differences among the ESG topics considered by rating agencies and disclosed by companies. We indeed extract a set of ESG topics/categories disclosed in companies' sustainability reports that is eight times larger than the original list of categories of this classification (542 versus 64, Sect. 5.1). For instance, our methodology unveils "`Education`" as a pivotal ESG topic disclosed by more than two-thirds of the selected companies. This topic is not explicitly included in the categorization, although it could be framed within three of its categories: Access to Basic Services, Human rights (Art. 26) or Philanthropy. Additional examples encompass the extracted ESG category "`Circular Economy`" which might fall under the categorization categories of "Waste" or "Resource Efficiency" as well as the ESG category "`Air Quality`" which could be framed within "Green Buildings" or "Health and Safety".

  Accordingly, the aforementioned ESG categorization encompasses critical topics hidden within vague categories or potentially overlooks them altogether, resulting in a reduction of substantial significance in subsequent analyses. Our approach addresses this limitation by leveraging a generative LLM in conjunction with the ICL technique and the RAG paradigm. This allows us to jointly simulate the outputs of a supervised text classifier, whose labels are the categories of the ESG categorization, and semantically generalize those labels. The ESG categorization is consequently leveraged as semantic guidance by the LLM, helping itself to extract more suitable topics while keeping the domain and semantics of the original ESG categorization. This semantic generalization can also diminish the reliance on specific ESG categorizations when classifying sentences since the ESG categories are exploited as semantic references rather than fixed labels. Nevertheless, this could also lead to the undesirable phenomenon of over-specialization which was tackled using semantic clustering (Sect. 3.4). We also used this ESG categorization, in the data preparation phase, to filter the report sentences using the text embeddings (Sect. 3.2.1). This semantic-oriented filtering approach allows us to further move towards a taxonomy-agnostic methodology as filtering is based on semantics rather than single keywords.

Extracting insights from companies' sustainability reports using generative LLMs, and graph representations   Lastly, our methodology differs from other recent ESG-focused and LLM-based tools (Sect. 2) such as ChatClimate [58] and ChatReport [60] by employing the paradigm of Retrieval-Augmented Generation (RAG), alongside In-Context Learning, for Knowledge Graph generation. This methodology, in combination with bipartite graph representation, allows us to report meaningful insights concerning the actions disclosed in companies' sustainability reports. In comparison, ChatClimate adopts the RAG paradigm to augment ESG-related questions for question-answering, whereas ChatReport leverages this paradigm to operationalise the compliance assessment of sustainability reports towards the recommendation guidelines of the Task Force on Climate-related Financial Disclosures (TCFD).

5.3 Limitations

Because of a significant computational workload, we endeavour to present insights concerning a sample of companies encompassing several sectors, regions and sizes. However, a greater subset - such as 1000 companies' reports - would further validate our findings and enable a more substantive analysis.

Our data preparation NLP pipeline relies on a PDF parser [78] to extract texts from sustainability reports. This parser extracts all texts including those from infographics and tables. This might yield some sentences without a proper syntactic structure, making extracting semantic meaning from them difficult or even impossible. Table 3 in Sect. 3.2.2 exhibits an example of this issue in the first sentence retrieved for the ESG category `Waste`. Although this sentence contains some details regarding this topic, it lacks a coherent message. However, the semantic understanding of LLMs in combination with the In-Context Learning technique and the paradigm of RAG could implicitly address this issue. Indeed, the sentence coverage of our retrieval-augmented triple generation (Sect. 3.3) is equal to 68.1 %, meaning that the language model acts as an implicit filtering layer and avoids generating triples for just about 30% of all the processed input sentences. The aforementioned example is within this set of ignored sentences. Although an end-to-end approach might be desired, discarding such meaningless sentences beforehand could help avoid an unnecessary computational workload. For instance, future works could tackle this issue by enhancing document parsing (e.g., by preserving the original layout) or adding a further, yet lightweight, filtering component. The latter might filter sentences according to their syntactical correctness or meaningfulness.

Another potential limitation concerns the interpretability of ESG scores using SHAP values. The SHAP framework is used to roughly interpret the impact of predictors on individual predictions. Global interpretability is derived using simple aggregating statistics such as mean/median SHAP values. Nevertheless, this aggregating approach for global interpretability might result in a mixed global interpretation in the presence of high diversity in the observations as can be a set of companies from worldwide nations covering eleven distinct sectors. The global impact of some predictors could still be accurate, yet some might be a mix of sector-dependent relationships or caused by the diversity of cause-effect connections. Indeed, the region-based interpretability analysis (Sect. 4.5) unveils more impacting factors or relationships for a specific company cluster in comparison to the global interpretability (Sect. 5.1). However, future works might conduct a further subset-based interpretability analysis by adopting a bottom-up approach and letting company groups emerge by themselves.

Lastly, the data provider for ESG scores used for our work might be a limitation worth highlighting. We rely on the ESG scores from the Refinitiv platform (Sect. 3.1.3), but, as highlighted in the Introduction Section, rating agencies have their assessment methodologies which could result in divergences in companies' ESG scores. Consequently, the findings relying on ESG scores (Sect. 4.4 and 4.5) might vary using ESG scores from other rating agencies such as Sustainalytics which adopts a risk-based assessment [116]. In addition, future works could integrate further ESG-related attributes from these rating agencies such as quantifying companies' water withdrawal, hazardous waste, gender pay gap and employee turnover.

6 Conclusions

In this work, we proposed a data-driven methodology based on generative LLMs to systematically evaluate the context in which ESG topics are disclosed by companies in their sustainability reports. The objective of this work was to contribute to the emerging field of automatic information extraction from companies' sustainability reports by implementing the best NLP pipeline to extract structured insights from lengthy and visually rich PDF documents. This generative LLM-based approach allowed us to directly investigate the companies' perspective concerning the ESG phenomenon.

Large Language Models (LLMs) can be versatile tools to accomplish diverse NLP-related tasks including also extracting structured information from textual data. We further explored this promising research direction by adopting the Retrieved-Augmented Generation (RAG) paradigm, alongside the In-Context Learning (ICL) technique, to extract ESG-related information as semantically structured triples. We then adopted a graph representation (bipartite graphs) to extract non-trivial statistics and conduct meaningful analyses concerning companies' disclosed actions. We employed a pre-trained language model from the open-source community, distinguishing us from other recent publications as far as we know. Furthermore, our LLM-based methodology overcomes important limitations related to traditional OIE techniques and the ESG categorization, allowing us to generate both semantically-aware and ESG-oriented triples instead of traditional subject-predicate-object (SPO) triples. This helped us to report meaningful findings such as statistical, similarity and correlation analyses on the ESG topics and actions extrapolated from companies' sustainability reports as well as conduct an interpretability analysis of ESG scores. Future works might integrate further data sources, such as ESG-related news, to analyse possible inconsistencies in companies' claims and actions. Another interesting research direction might be to integrate Semantic Role Labelling (SRL) to enhance the extracted structured information with semantic roles, such as the agent, manner, and purpose of an action, as well as other contextual information, such as time and location.

Appendix A: Qualitative analysis of the generated triples

We evaluated the generated triples by prompting the same LLM (WizardLM, [16]) to evaluate triple quality. We prompted the model to evaluate the coherence and alignment between the structured information (output) and the sentence (input), considering also the coherence of each triple attribute (`cat`, `pred`, and `obj`). The model was prompted to evaluate each, leveraging also its ICL abilities, using numerical scores on a scale from 0 to 3. The full model instruction used for this evaluation is exhibited in Sect. 2.3.1 in the

Supplementary Material (SM) document. We specifically analysed a random sample of 1000 triples from the total of forty thousand triples generated from all companies' sustainability reports. The triple sample has an average score of 2.65 (std: 0.44), showing a fairly high quality of the structured information extracted. Specifically, the ESG category (`cat`) and object (`obj`) show high average performance (2.76 and 2.78) in comparison to the generated predicate (`pred`, 2.5). The distributions are visually exhibited in Sect. 2.3.2 in the Supplementary Material (SM) document. Table 9 showcases five generated triples with their evaluation. The ESG topic mentioned directly or indirectly in the sentence is semantically captured fairly well, whereas extracting the predicate affecting it might be more challenging. For example, the predicate of the last triple in Table 9 displays low quality in terms of syntax and coherence with the action of the original sentence. For instance, the predicate "`Creation of`" could have been more suitable for this sentence. However, this triple notably captures the actual main ESG topic of this tricky sentence (`Supply Chain`), despite the mention of another potential, yet secondary ESG aspect (employment opportunities). Furthermore, the second triple in Table 9 displays another phenomenon observed in some triples: lengthy predicate attributes enveloping direct objects. This happens especially when the disclosed actions involve transitive verbs as in this example (to meet something). Thus, despite its semantic correctness and coherence with the disclosed action, this might make our effort in generalising challenging.

Appendix B: Ablation study on the model instruction

Furthermore, since the model instruction has a great impact on the quality of the generated text [117], we conduct an ablation study to compare qualitatively the triples generated through different prompt templates: with(out) In-Context Learning (`EXAMPLES` in Fig. 3) and with(out) the semantic output schema (`DATASCHEMA`).

We find that including `examples` in the prompt, and thus exploiting the In-Context Learning capabilities of the model, generates better triples in terms of both information completeness and semantic representation, observable respectively in the first two comparisons and the third one in Table 10. Furthermore, that helps the model compose its response by adhering to a specific output format. Indeed, although the model was already prompted to generate text as a valid JSON object, it outputs texts in a valid format only after adding In-Context Learning.

Moreover, adding a semantic output schema in the prompt (`DATASCHEMA`) helps the generative language model to better focus on ESG-related information as exhibited in Table 11. The semantic schema provides the language model with detailed semantic descriptions concerning the types of information to extract: "an issue related to an ESG aspect" (`cat`), "a nominalised verb affecting that aspect" `pred`, and "an entity undergoing the predicate" (`obj`). This could drastically affect the structured information extracted from a sentence, especially in those with multiple causes, such as in the second comparison shown in Table 11. Furthermore, defining a semantic schema allows us to incorporate a list of ESG topics (ESG categorization, Sect. 3.1.2) into the semantic description of the ESG category attribute (`cat`). This enhanced the model's ability to extract an ESG topic (`cat`) mentioned indirectly in a sentence by jointly leveraging this list of ESG categories as semantic references, its semantic understanding, and its generative abilities. An example of this enhancement can be observed in the third comparison in Table 11, while the

Table 9 *Evaluation of a sample of five generated triples and their original sentences.* The triples were automatically evaluated by an LLM outputting a 0-3 numerical score for each triple attribute

Sentence	Generated triple	cat	pred	obj	AVG
We evaluate products and materials against HSE compliance, safety and sustainability metrics	{'esg_category': 'Product Safety', 'predicate': 'Evaluation of', 'object': 'Products and materials against HSE compliance, safety and sustainability metrics'}	3	3	3	3
With increasing digitalisation, the ability to provide a data environment meeting the highest security and regulatory standards, such as GDPR, is critical	{'esg_category': 'Data Privacy', 'predicate': 'Meeting the highest security and regulatory standards' 'object': 'The ability to provide a data environment'}	3	3	3	3
We partner with trusted, leading non-profit organizations committed to impacting the areas closest to our homes and our hearts, with a particular focus on Los Angeles and New York City, where we have our largest footprint	{'esg_category': 'Philanthropy', 'predicate': 'Partnership with' 'object': 'Trusted, leading non-profit organizations'}	3	2	3	2.67
That data will help us refine operations, modify purchasing decisions and meet our emission reduction goals	{'esg_category': 'Environmental', 'predicate': 'Use of', 'object': 'Data'}	2	2	2	2
We also indirectly create employment opportunities for many young people in our supply chain and retail distribution networks	{'esg_category': 'Supply Chain', 'predicate': 'Indirect' 'object': 'Employment opportunities for many young people'}	2	1	2	1.67

Table 10 Three comparing examples of the triples generated with/without using In-Context Learning in the model instruction

	ESG category (`cat`)	Predicate (`pred`)	Object (`obj`)
`EXAMPLES`	Work-Life Balance	Provision of	A policy providing clarity around flexible work options
without `EXAMPLES`	Family Friendly Policies	Provides	Flexible work option
`EXAMPLES`	Health and Safety	Creation of	COVID-19 safety tips for ridesharing
without `EXAMPLES`	Public Health	Created	COVID-19 safety tips
`EXAMPLES`	Climate Risk Management	Partnership with	CoGo to provide personalised carbon footprints
without `EXAMPLES`	Climate Risk Management	Announced	Partnership

Table 11 Three comparing examples of triples generated with/without the semantic output data schema in the model instruction

	ESG category (`cat`)	Predicate (`pred`)	Object (`obj`)
`DATASCHEMA`	Work-Life Balance	Provision of	A policy providing clarity around flexible work options
without `DATASCHEMA`	Work-Life Balance	Providing clarity around	Flexible work options available to parents and caregivers
`DATASCHEMA`	Philanthropy	Partnership with	The Bill and Melinda Gates Foundation and the Western Cape Department of Health
without `DATASCHEMA`	Health Care	Delivery of	Over 1.4 million prescriptions
`DATASCHEMA`	Climate Risk Management	Partnership with	CoGo to provide personalised carbon footprints
without `DATASCHEMA`	Carbon Footprint	Partnership with	To provide personalised carbon footprints

complete semantic schema can be seen in the Supplementary Material document within the full model instruction (see Sect. SM2.1).

Appendix C: Hyper-parameter choice of the large language model

Large Language models have some hyper-parameters for tuning their textual outputs, and consequently, some choices should be made to address these further degrees of freedom. The temperature parameter controls the randomness of the model responses by influencing the model's confidence in its most likely output [118]. During the decoding phase, this parameter alters the model output by scaling the logits before applying the softmax function. A high temperature makes the model output more diverse and creative but also more unpredictable. Conversely, a lower temperature makes the model output more deterministic and focused. Setting a temperature equal to zero corresponds to greedy decoding [118]. Accordingly, we opt for greedy decoding to ensure deterministic outputs and to make the generation process adhere to instructions as much as possible [60].

Another hyper-parameter that affects text generation is the beam number representing the number of tokens considered during the beam search algorithm [118]. Beam search is a sampling decoding algorithm that improves the output of LLMs by pruning off bad thinking patterns at generation time. This algorithm works by iteratively generating a sequence of b_{dim} tokens, and then outputting the sequence with the highest probability [119]. We found through extensive experiments that the beam number (b_{dim}) ranging from 4 to

6 strikes a good balance between semantic representation and computational workload. We accordingly adopt a beam number equal to 6.

Appendix D: Empirical experiments with different large language models

We conducted empirical experiments using various instruction-tuned Large Language Models (LLMs): Google's `Flan-T5-Large`, `Alpaca-LoRA-7B`, `WizardLM-7B` and OpenAI's `ChatGPT` in its version based on GPT-3.5. These experiments aimed to assess the performance of different LLMs and identify a suitable, cost-free generative LLM for constructing an NLP pipeline to extract structured insights from ESG-related textual documents. Additionally, we evaluated OpenAI's commercial ChatGPT to compare open-sourced LLMs against the state-of-the-art, general-purpose and paid Large Language Model. To maintain consistency, we employed the same model instruction for all LLMs which was also the one adopted for this work (Sect. SM2.1 of the Supplementary Material (SM) document).

Table 12 presents three comparative analyses of textual outputs generated by different LLMs, using a sample of sentences extracted from companies' sustainability reports. OpenAI's `ChatGPT` demonstrated exceptional performance, nearly surpassing all open-sourced LLMs in generating ESG-oriented triples and serving as the de facto golden standard. However, in the third comparison (Table 12), `ChatGPT`'s output was subjectively inferior to that of WizardLM. Although its extracted triple conveys the meaning of the company's disclosure, we argue that it did not focus on the main action disclosed in the sentence. It initially identified "green steel" as the main ESG-related issue ("Green Products") rather than the investment in a platform for customers; misclassifying the ESG topic of the disclosure. `WizardLM`, the LLaMA-based LLM chosen for this work, exhibited remarkable performance in understanding the NLP task and extracting ESG-focused structured information, closely matching the outputs from the paid OpenAI's model (Table 12). `Alpaca`, another LLaMA-based LLM, demonstrated a good understanding of the NLP task but lagged in the quality of structured information extraction across all comparisons, especially for predicates and objects (Table 12). Lastly, `Flan-T5`, Google's instruction fine-tuned version of the Text-to-Text Transfer Transformer (T5) Language Model, generated nonsensical texts, indicating a lack of understanding of the NLP task.

Appendix E: Performance of the OLS model

We here report the performance of the OLS regression through different metrics (see also Sect. SM2.5.2 in the Supplementary Material document). Firstly, the Coefficient of Determination (R^2, [120]) measures the proportion of variation in the dependent variable explained by the model predictors, representing the goodness of the inference ability of the model. Low coefficients express a little variation proportion explained by the model predictors, resulting in poor performance on the inference of the dependent variable (ESG scores). In contrast, in the presence of a high variation proportion explained, the model predicts the dependent variable with small errors. Our OLS model achieves a R^2 of 0.71 using the optimal alpha, demonstrating a broad variation explained by our features to infer ESG scores. On the other hand, the Root Mean Square Error (RMSE, [121]) is a quadratic score, in the same units of the dependent variable, in which the average error in the model predictions is computed by averaging the squared individual errors. Our regression model

Table 12 *Three comparisons of the outputs generated by different Large Language Models.* The original sentence is exhibited before the different model outputs. All the LLMs are prompted with the same model instruction adopted in this work. Model outputs are reported in their raw form

LLM	Original sentence/Model output
"In 2022, we will be completing a review of our supply chain responsibility and risk assessment to identify opportunities to engage with our suppliers and build capacity."	
Flan-T5-Large	`json>src:src.getres`
Alpaca-LoRA-7B	`<json>{"esg_actions": [{"esg_category": "Supply Chain", "predicate": "Review of our supply chain responsibility and risk assessment", "object": "Opportunities to engage with our suppliers and build capacity"}]}</json>`
WizardLM-7B	`<json>{"esg actions": [{"esg_category": "Supply Chain", "predicate": "Review of", "object": "Our supply chain responsibility and risk assessment"}]}</json>`
ChatGPT (GPT-3.5)	`<json>{"esg actions": [{"esg category": "Supply Chain", "predicate": "Completion of", "object": "A review of our supply chain responsibility and risk assessment"}]}</json>`
"Many of our buildings also capture rainwater that is used when irrigation is required."	
Flan-T5-Large	`json>separateby:entity, se`
Alpaca-LoRA-7B	`<json>{"esg_actions": [{"esg_category": "Water", "predicate": "Capture of rainwater", "object": "Used when irrigation is required"}]}</json>`
WizardLM-7B	`<json>{"esg_actions": [{"esg_category": "Water", "predicate": "Capture of", "object": "Rainwater that is used for irrigation"}]}</json>`
ChatGPT (GPT-3.5)	`<json>{"esg actions": [{"esg category": "Water", "predicate": "Capture of", "object": "Rainwater for irrigation"}]}</json>`
"This investment is a platform to provide our customers with more of the green steel they expect from like-minded partners like U. S. Steel."	
Flan-T5-Large	`json>str>Investment:str>`
Alpaca-LoRA-7B	`<json>{"esg_actions": [{"esg_category": "Energy", "predicate": "This investment", "object": "Green steel"}]}</json>`
WizardLM-7B	`<json>{"esg actions": [{"esg_category": "Environmental Management System", "predicate": "Investment in", "object": "A platform to provide customers with more green steel"}]}</json>`
ChatGPT (GPT-3.5)	`<json>{"esg actions": [{"esg category": "Green Products", "predicate": "Platform to provide", "object": "Customers with more of the green steel"}]}</json>`

achieves an RMSE equal to 7.76, representing the average difference between the actual ESG score and the inferred one. Lastly, we report the model performance (7.9 %) using the Weighted Mean Absolute Percentage Error (wMAPE, [113, 122]), a scale-independent score that measures the average of absolute percentage errors.

To conclude the review of the regression model performance, we conduct a residual analysis to check the linear assumptions required to properly shape the problem as a linear model. The assumption of normal distribution of the residuals ($E_i \sim N(0, \sigma^2)$) is confirmed by the Anderson-Darling test [123] with a *p*-value equal to 6.6 % as well as through the QQ plot of residuals versus Normal distribution showing points lie on a line. Concerning homoscedasticity, a condition in which the residual variance is constant across all the model predictions, there are no visible patterns in the scatter plot of residuals versus predicted ESG scores. The same condition is confirmed by the scatter plot of the pre-

dicted ESG scores versus the actual scores. However, a slight overestimation trend might be spotted for ESG scores below 50, showing a limit of our predictors for interpreting these low scores. A graphical panel with all the graphical residual analyses is shown in the Supplementary Material document (see Sect. SM2.5.4).

Abbreviations
CSR, Corporate Social Responsibility; EBITDA, Earnings Before Interest, Taxes, Depreciation, and Amortization; ESG, Environmental, Social, and Corporate Governance; GICS, Global Industry Classification Standard; ICL, In-Context Learning; KG(s), Knowledge Graph(s); LASSO, Least Absolute Shrinkage and Selection Operator; LLM(s), Large Language Model(s); NLP, Natural Language Processing; OIE, Open Information Extraction; OLS, Ordinary Least Squares; RAG, Retrieved Augmented Generation; SASB, Sustainability Accounting Standard Boards; SDG(s), Sustainable Development Goal(s); SHAP, SHapley Additive exPlanations; SM, Supplementary Material; SPO, Subject-Predicate-Object.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00481-2.

> **Additional file 1.** Further and extensive tabular and graphical views are exhibited in the Supplementary Material (SM) document (PDF 2.1 MB)

Author contributions
MB collected, prepared and processed the data, implemented the proposed approach, wrote the paper and interpreted the findings. MB, CN and JS conceptualised and designed the proposed approach. CN, JS, BL and AP supervised the research direction of this study. CN, JS and BL supervised the writing process. All authors read and approved the final manuscript.

Data availability
The sustainability reports used, processed and analysed during the current study can be publicly retrieved from the following websites: sasb.org [73] and responsibilityreports.com [74]. The complete list of the companies we considered can be found in Sect. SM1.2 of the supplementary material document. The ESG categorization exploited in our work was extrapolated from Table IV (4) of the work [8] by Berg *et al.*. It is also exhibited in Sect. SM1.1 of the supplementary material document. The sector-level materiality was extracted from the table in Appendix C of the work [35] by Khan *et al.* The ESG scores and other financial information used in this study are available from the Refinitiv platform, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.
All the datasets used, processed and analysed and the extracted companies triples can be found in this repository: github.com/saturnMars/derivingStructuredInsightsFromSustainabilityReportsViaLargeLanguageModels. Further data and findings are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]University of Trento, Trento, Italy. [2]Ipazia S.p.A., Milan, Italy. [3]Fondazione Bruno Kessler (FBK), Trento, Italy.

References
1. United Nations: the sustainable development agenda. https://www.un.org/sustainabledevelopment/development-agenda. Accessed 22-09-2023
2. European Union: non-financial reporting directive (NFRD). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014L0095. Accessed 2023-07-04
3. European Union: corporate sustainability reporting directive (CSRD). https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en. Accessed 2023-07-04
4. Wong C, Petroy E (2020) Rate the raters 2020: investor survey and interview results. Survey report, SustainAbility Institute by ERM. https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sustainability-ratetheraters2020-report.pdf

5. Chatterji AK, Durand R, Levine DI, Touboul S (2016) Do ratings of firms converge? Implications for managers, investors and strategy researchers. Strateg Manag J 37(8):1597–1614

6. Abhayawansa S, Tyagi S (2021) Sustainable investing: the black box of environmental, social, and governance (ESG) ratings. J Wealth Manag 24(1):49–54

7. Billio M, Costola M, Hristova I, Latino C, Pelizzon L (2021) Inside the esg ratings: (dis)agreement and performance. Corp Soc-Responsib Environ Manag 28(5):1426–1445

8. Berg F, Koelbel JF, Rigobon R (2022) Aggregate confusion: the divergence of ESG ratings. Rev Finance 26(6):1315–1344

9. Ehlers T, Elsenhuber U, Jegarasasingam K, Jondeau E (2023) Deconstructing ESG scores: how to invest with your own criteria? IMF Work Pap 2023(057):001. https://doi.org/10.5089/9798400235283.001.A001

10. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

11. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV (2021) Finetuned language models are zero-shot learners. arXiv preprint. arXiv:2109.01652

12. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi EH, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W (2022) Emergent abilities of large language models. arXiv:2206.07682

13. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Red Hook, pp 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

14. Asratian AS, Denley TM, Häggkvist R (1998) Bipartite graphs and their applications, vol 131. Cambridge University Press, Cambridge

15. Guillaume J-L, Latapy M (2006) Bipartite graphs as models of complex networks. Phys A, Stat Mech Appl 371(2):795–813

16. Xu C, Sun Q, Zheng K, Geng X, Zhao P, Feng J, Tao C, Jiang D (2023) WizardLM: empowering large language models to follow complex instructions. arXiv:2304.12244

17. Dong Q, Li L, Dai D, Zheng C, Wu Z, Chang B, Sun X, Xu J, Li L, Sui Z (2023) A survey on in-context learning. arXiv:2301.00234

18. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-T, Rocktäschel T et al (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst 33:9459–9474

19. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners. Adv Neural Inf Process Syst 35:22199–22213

20. Trott S, Jones C, Chang T, Michaelov J, Bergen B (2023) Do large language models know what humans know? Cogn Sci 47(7):13309

21. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z et al (2023) A survey of large language models. arXiv preprint. arXiv:2303.18223

22. Reynolds L, McDonell K (2021) Prompt programming for large language models: beyond the few-shot paradigm. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems, pp 1–7

23. Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, Liu Q (2023) Aligning large language models with human: a survey. arXiv:2307.12966

24. Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, Li J, Hu R, Zhang T, Wu F et al (2023) Instruction tuning for large language models: a survey. arXiv preprint. arXiv:2308.10792

25. Carta S, Giuliani A, Piano L, Podda AS, Pompianu L, Tiddia SG (2023) Iterative zero-shot LLM prompting for knowledge graph construction. arXiv:2307.01128

26. Meyer L-P, Stadler C, Frey J, Radtke N, Junghanns K, Meissner R, Dziwis G, Bulert K, Martin M (2023) Llm-assisted knowledge graph engineering: experiments with chatgpt. arXiv preprint. arXiv:2307.06917

27. Trajanoska M, Stojanov R, Trajanov D (2023) Enhancing knowledge graph construction using large language models. arXiv:2305.04676

28. Zhu Y, Wang X, Chen J, Qiao S, Ou Y, Yao Y, Deng S, Chen H, Zhang N (2023) LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. arXiv:2305.13168

29. Reinanda R, Meij E, de Rijke M et al (2020) Knowledge graphs: an information retrieval perspective. Found Trends Inf Retr 14(4):289–444

30. Fensel D, Şimşek U, Angele K, Huaman E, Kärle E, Panasiuk O, Toma I, Umbrich J, Wahler A, Fensel D et al (2020) Introduction: what is a knowledge graph? In: Knowledge graphs: methodology, tools and selected use cases, pp 1–10

31. Hogan A, Blomqvist E, Cochez M, d'Amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JEL, Navigli R, Neumaier S et al (2021) Knowledge graphs. ACM Comput Surv 54(4):1–37

32. Yu EP-Y, Van Luu B (2021) International variations in esg disclosure—do cross-listed companies care more? Int Rev Financ Anal 75:101731

33. Baldini M, Maso LD, Liberatore G, Mazzi F, Terzani S (2018) Role of country-and firm-level determinants in environmental, social, and governance disclosure. J Bus Ethics 150:79–98

34. Eccles RG, Krzus MP, Rogers J, Serafeim G (2012) The need for sector-specific materiality and sustainability reporting standards. J Appl Corp Finance 24(2):65–71

35. Khan M, Serafeim G, Yoon A (2016) Corporate sustainability: first evidence on materiality. Account Rev 91(6):1697–1724

36. Busco C, Consolandi C, Eccles RG, Sofra E (2020) A preliminary analysis of SASB reporting: disclosure topics, financial relevance, and the financial intensity of ESG materiality. J Appl Corp Finance 32(2):117–125

37. Zou X (2020) A survey on application of knowledge graph. J Phys Conf Ser 1487:012016

38. Ji S, Pan S, Cambria E, Marttinen P, Philip SY (2021) A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst 33(2):494–514

39. Jia Z, Pramanik S, Saha Roy R, Weikum G (2021) Complex temporal question answering on knowledge graphs. In: Proceedings of the 30th ACM international conference on information & knowledge management, pp 792–802

40. Shao B, Li X, Bian G (2021) A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph. Expert Syst Appl 165:113764
41. Fensel D, Simsek U, Angele K, Huaman E, Kärle E, Panasiuk O, Toma I, Umbrich J, Wahler A (2020) Knowledge graphs. Springer, Cham
42. Yan J, Wang C, Cheng W, Gao M, Zhou A (2018) A retrospective of knowledge graphs. Front Comput Sci 12:55–74
43. Katiyar A, Cardie C (2017) Going out on a limb: joint extraction of entity mentions and relations without dependency trees. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), pp 917–928
44. Wang Y, Yu B, Zhang Y, Liu T, Zhu H, Sun L (2020) Tplinker: single-stage joint extraction of entities and relations through token pair linking. arXiv preprint. arXiv:2010.13415
45. Niklaus C, Cetto M, Freitas A, Handschuh S (2018) A survey on open information extraction. arXiv preprint. arXiv:1806.05599
46. Ro Y, Lee Y, Kang P (2020) Multi^{2}OIE: multilingual open information extraction based on multi-head attention with BERT. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 1107–1117. https://doi.org/10.18653/v1/2020.findings-emnlp.99. https://aclanthology.org/2020.findings-emnlp.99
47. Wang C, Liu X, Chen Z, Hong H, Tang J, Song D (2021) Zero-shot information extraction as a unified text-to-triple translation. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 1225–1238. https://doi.org/10.18653/v1/2021.emnlp-main.94. https://aclanthology.org/2021.emnlp-main.94
48. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805
49. Liu P, Gao W, Dong W, Huang S, Zhang Y (2022) Open information extraction from 2007 to 2022 – a survey. arXiv:2208.08690
50. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X (2023) Unifying large language models and knowledge graphs: a roadmap. arXiv preprint. arXiv:2306.08302
51. Chen J, Ma L, Li X, Thakurdesai N, Xu J, Cho JHD, Nag K, Korpeoglu E, Kumar S, Achan K (2023) Knowledge graph completion models are few-shot learners: an empirical study of relation labeling in e-commerce with LLMs. arXiv:2305.09858
52. Axelsson A, Skantze G (2023) Using large language models for zero-shot natural language generation from knowledge graphs. arXiv:2307.07312
53. Chou C, Clark R, Kimbrough SO (2023) What do firms say in reporting on impacts of climate change? An approach to monitoring ESG actions and environmental policy. In: Corporate social responsibility and environmental management
54. Raghupathi V, Ren J, Raghupathi W (2020) Identifying corporate sustainability issues by analyzing shareholder resolutions: a machine-learning text analytics approach. Sustainability 12(11):4753
55. Marodon R, Jacouton J-B, Laulanie A (2022) The proof is in the pudding. revealing the SDGs with artificial intelligence. Working paper 85f81dba-c8e2-4255-878a-0, Agence Française de Développement. https://ideas.repec.org/p/avg/wpaper/en14520.html
56. SDG Prospector – artificial intelligence serving the SDGs. https://www.sdgprospector.org. Accessed 22-09-2023
57. Webersinke N, Kraus M, Bingler JA, Leippold M (2021) Climatebert: a pretrained language model for climate-related text. arXiv preprint. arXiv:2110.12010
58. Vaghefi SA, Wang Q, Muccione V, Ni J, Kraus M, Bingler J, Schimanski T, Colesanti-Senni C, Stammbach D, Webersinke N et al (2023). Chatclimate: grounding conversational ai in climate science
59. ChatClimate grounded on the latest IPCC report. https://www.chatclimate.ai. Accessed 22-09-2023
60. Ni J, Bingler J, Colesanti-Senni C, Kraus M, Gostlow G, Schimanski T, Stammbach D, Vaghefi SA, Wang Q, Webersinke N et al (2023) Paradigm shift in sustainability disclosure analysis: empowering stakeholders with chatreport, a language model-based tool. arXiv preprint. arXiv:2306.15518
61. TCFD (2017) Recommendations of the task force on climate-related financial disclosures. Task force on climate-related financial disclosures. https://assets.bbhub.io/company/sites/60/2021/10/FINAL-2017-TCFD-Report.pdf
62. Campbell JL (2007) Why would corporations behave in socially responsible ways? An institutional theory of corporate social responsibility. Acad Manag Rev 32(3):946–967
63. Drempetic S, Klein C, Zwergel B (2020) The influence of firm size on the ESG score: corporate sustainability ratings under review. J Bus Ethics 167:333–360
64. European Union: EU emissions trading system (EU ETS). https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets_en. Accessed 2023-09-25
65. European Union: fit for 55. https://www.consilium.europa.eu/en/policies/green-deal/fit-for-55-the-eu-plan-for-a-green-transition/. Accessed 2023-09-25
66. LaBella MJ, Sullivan L, Russell J, Novikov D (2019) The devil is in the details: the divergence in esg data and implications for responsible investing. QS Investors, New York
67. Dobrick J, Klein C, Zwergel B (2023) Size bias in refinitiv esg data. Finance Res Lett 55:104014
68. Admin@Evo (2023) Why impact materiality is critical to double materiality assessments. Section: blog. https://gistimpact.com/news-insights/why-impact-materiality-is-critical-to-double-materiality-assessments/. Accessed 2024-04-19
69. Working paper: balancing your materiality assessment. Deloitte (2022). https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/risk/deloitte-nl-risk-double-materiality.pdf
70. Doyle TM (2018) Ratings that don't rate: the subjective world of esg ratings agencies. American Council for Capital Formation, 65–71
71. OECD (2020) OECD business and finance outlook 2020: sustainable and resilient finance. OECD business and finance outlook, vol 6. OECD, Paris. https://doi.org/10.1787/eb61fd29-en.

72. Aliakbari E, Globerman S (2023) The impracticality of standardizing ESG reporting (ESG: myths and realities)
73. SSAB: SASB reporters. https://sasb.org/company-use/sasb-reporters. Accessed 2022-04-07
74. IR Solutions: ResponsibilityReports. https://www.responsibilityreports.com. Accessed 2022-04-07
75. Hugging Face: statistics on the number of monolingual models by language hosted on the Hugging Face platform. https://huggingface.co/languages. Accessed 2023-07-04
76. Refinitiv: environmental, social and governance (ESG) scores from Refinitiv – May 2022. https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf. Accessed 2023-07-10
77. Sahin Ö, Bax K, Czado C, Paterlini S (2022) Environmental, social, governance scores and the missing pillar—why does missing information matter? Corp Soc-Responsib Environ Manag 29(5):1782–1798
78. Artifex: PyMuPDF – pypi.org. https://pypi.org/project/PyMuPDF. Accessed 22-09-2023
79. Sadvilkar N, Neumann M (2020) PySBD: pragmatic sentence boundary disambiguation. In: Proceedings of second workshop for NLP open source software (NLP-OSS). Association for Computational Linguistics, Online, pp 110–114. https://www.aclweb.org/anthology/2020.nlposs-1.15
80. Sadvilkar N PySBD – pypi.org. pypi.org/project/pysbd. Accessed 22-09-2023
81. Croft WB, Metzler D, Strohman T (2010) Search engines: information retrieval in practice, vol 520. Addison-Wesley, Reading
82. Bast H, Buchhold B, Haussmann E (2016) Semantic search on text and knowledge bases. Found Trends Inf Retr 10(2–3):119–271
83. Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. arXiv preprint. arXiv:1908.10084
84. Muennighoff N (2022) SGPT: GPT sentence embeddings for semantic search. arXiv:2202.08904
85. Guo J, Cai Y, Fan Y, Sun F, Zhang R, Cheng X (2022) Semantic models for the first-stage retrieval: a comprehensive review. ACM Trans Inf Syst 40(4):1–42
86. Buttcher S, Clarke CL, Cormack GV (2016) Information retrieval: implementing and evaluating search engines. Mit Press, Cambridge
87. Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, Song X, Ward R (2016) Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Trans Audio Speech Lang Process 24(4):694–707
88. Yang W, Zhang H, Lin J (2019) Simple applications of bert for ad hoc document retrieval. arXiv preprint. arXiv:1903.10972
89. NLP Group of The University of Hong Kong: Instructor-xl · Hugging Face – huggingface.co. https://huggingface.co/hkunlp/instructor-xl. Accessed 25-09-2023
90. Su H, Kasai J, Wang Y, Hu Y, Ostendorf M, Yih W-T, Smith NA, Zettlemoyer L, Yu T et al (2022) One embedder, any task: Instruction-finetuned text embeddings. arXiv preprint. arXiv:2212.09741
91. Allen Institute for AI: open information extraction – demo. https://demo.allennlp.org/open-information-extraction. Accessed 2023-07-10
92. Yurtsev E Kor – pypi.org. https://pypi.org/project/kor. Accessed 25-09-2023
93. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li E, Wang X, Dehghani M, Brahma S et al (2022) Scaling instruction-finetuned language models. arXiv preprint. arXiv:2210.11416
94. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G (2023) LLaMA: open and efficient foundation language models. arXiv:2302.13971
95. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, Hashimoto TB (2023) Stanford alpaca: an instruction-following llama model
96. Jobbins T TheBloke/wizardLM-7B-HF · Hugging Face. https://huggingface.co/TheBloke/wizardLM-7B-HF. Accessed 25-09-2023
97. Estrada E (2011) The structure of complex networks: theory and applications. Oxford University Press, London. https://doi.org/10.1093/acprof:oso/9780199591756.001.0001
98. Newman M (2010) Networks: an introduction. Oxford University Press, London
99. Brandes U (2005) Network analysis: methodological foundations, vol 3418. Springer, Berlin
100. Zhang J, Luo Y (2017) Degree centrality, betweenness centrality, and closeness centrality in social network. In: 2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017). Atlantis Press, pp 300–303
101. Faust K (1997) Centrality in affiliation networks. Soc Netw 19(2):157–191
102. Barthelemy M (2004) Betweenness centrality in large complex networks. Eur Phys J B 38(2):163–168
103. Cover TM (1999) Elements of information theory. Wiley, New York
104. Costa LdF (2021) Further generalizations of the Jaccard index. arXiv preprint. arXiv:2110.09619
105. Gotelli JN, Ulrich W (2012) Statistical challenges in null model analysis. Oikos 121(2):171–180
106. Zheng A, Casari A (2018) Feature engineering for machine learning: principles and techniques for data scientists. "O'Reilly Media, Inc.", Newton
107. Abdi H (2007) The Kendall rank correlation coefficient. In: Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 508–510
108. Cohen I, Huang Y, Chen J, Benesty J, Benesty J, Chen J, Huang Y, Cohen I (2009) Pearson correlation coefficient. In: Noise reduction in speech processing, pp 1–4
109. Ott RL, Longnecker MT (2015) An introduction to statistical methods and data analysis. Cengage Learning. Boston
110. Draper NR, Smith H (1998) Applied regression analysis, vol 326. Wiley, New York
111. Dismuke C, Lindrooth R (2006) Ordinary least squares. Methods Des Outcomes Res 93(1):93–104
112. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc, Ser B, Stat Methodol 67(2):301–320
113. Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, Berlin
114. Berrar D (2018) Cross-validation. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

115. Find industry topics. SASB. https://sasb.org/standards/materiality-finder/find/?company[0]=IT0005239360. Accessed 2023-12-21
116. Sustainalytics: ESG risk ratings – methodology abstract, version 2.1. https://connect.sustainalytics.com/hubfs/INV/Methodology/Sustainalytics_ESG%20Ratings_Methodology%20Abstract.pdf. Accessed 2023-07-10
117. Zhao Z, Wallace E, Feng S, Klein D, Singh S (2021) Calibrate before use: improving few-shot performance of language models. In: International conference on machine learning. PMLR, pp 12697–12706
118. Tunstall L, Von Werra L, Wolf T (2022) Natural language processing with transformers. "O'Reilly Media, Inc.", Newton
119. Hewitt J, Manning CD, Liang P (2022) Truncation sampling as language model desmoothing. arXiv preprint. arXiv:2210.15191
120. Zhang D (2017) A coefficient of determination for generalized linear models. Am Stat 71(4):310–316
121. Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Clim Res 30(1):79–82
122. Kim S, Kim H (2016) A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast 32(3):669–679
123. Nelson LS (1998) The Anderson-Darling test for normality. J Qual Technol 30(3):298

Publisher's Note