
GLITTER OR GOLD? DERIVING STRUCTURED INSIGHTS FROM SUSTAINABILITY REPORTS VIA LARGE LANGUAGE MODELS

Marco Bronzini

University of Trento, Ipazia S.p.A.
Trento, Milano (Italy)
marco.bronzini-1@unitn.it

Carlo Nicolini

Ipazia S.p.A.
Milano (Italy)
c.nicolini@ipazia.com

Bruno Lepri

Fondazione Bruno Kessler (FBK), Ipazia S.p.A.
Trento, Milano (Italy)
lepri@fbk.eu

Andrea Passerini

University of Trento
Trento (Italy)
andrea.passerini@unitn.it

Jacopo Staiano

University of Trento
Trento (Italy)
jacopo.staiano@unitn.it

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. Such information is publicly released in a variety of non-structured and multi-modal documentation. Hence, it is not straightforward to aggregate and consolidate such data in a cohesive framework to further derive insights about sustainability practices across companies and markets. Thus, it is natural to resort to Information Extraction (IE) techniques to provide concise, informative and actionable data to the stakeholders. Moving beyond traditional text processing techniques, in this work we leverage Large Language Models (LLMs), along with prominent approaches such as Retrieved Augmented Generation and in-context learning, to extract semantically structured information from sustainability reports. We then adopt graph-based representations to generate meaningful statistical, similarity and correlation analyses concerning the obtained findings, highlighting the prominent sustainability actions undertaken across industries and discussing emerging similarity and disclosing patterns at company, sector and region levels. Lastly, we investigate which factual aspects impact the most on companies' ESG scores using our findings and other company information.

Keywords ESG Dimensions · Non-financial Disclosures · Information Extraction · 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 broken down into quantitative and qualitative indicators, such as waste management, emissions, labour rights, and diversity, that define and evaluate the extent 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.

In 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]), enforced since 2014, which requires all public-interest companies with more than 500 employees to disclose non-financial information periodically. Furthermore, starting in 2024, large companies must comply with new demanding rules in the European Union’s Corporate Sustainability Reporting Directive (CSRD, [3]), enforced in January 2023. This directive also enlarged the pool of companies concerned by a factor of 4, from roughly 12 thousand to 50 thousand companies [4].

These non-financial disclosures are typically reported in sustainability reports, web pages, social media posts, news, press releases or earning calls transcripts. To manage 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* [5]. Currently, these rating agencies rely on proprietary assessment methodologies, with different perspectives on the measurement, scope and weight of ESG aspects, resulting in divergences in companies’ evaluations across agencies and thus unsatisfactory degrees of explainability, transparency or fairness [6, 7, 8, 9]. Stakeholders might overcome this issue by directly accessing non-financial information and imposing their own scope and weight to assess corporates’ ESG performance [5, 10]. However, it can be challenging and laborious to extract meaningful insights from ESG-related data sources which often include lengthy documents.

In this scenario, data-driven approaches, and specifically Natural Language Processing (NLP) techniques, can automatically extract insights from textual data, providing stakeholders with a good trade-off between relying on opaque assessments and reading verbose textual content. These extracted insights can provide them with a means for comparing companies’ ESG-related assertions.

In this work, we propose a methodology based on Large Language Models (LLMs, [11, 12, 13]) to extract insights from non-financial disclosures. LLMs have consistently been shown to hold semantic understanding and store factual knowledge [14, 15, 16, 17]. The instruction-tuning approach (that is, providing task descriptions via natural language instructions) can further enhance their generalization capabilities to successfully tackle downstream NLP tasks like information extraction [11, 18, 19, 20, 21]. Generating model instructions, a task known as prompt engineering, can consequently make LLMs even more powerful and flexible tools. Furthermore, these models have also demonstrated a remarkable ability as few-shot learners using in-context learning [22], a technique that relies on providing a few input-output samples within the textual prompt itself.

Our approach relies on an instruction-finetuned generative LLM, alongside the technique of in-context learning and the paradigm of Retrieval Augmented Generation (RAG, [14]), to extract structured information related to ESG aspects from companies’ non-financial disclosures. Specifically, we adopted the LLaMA-based “WizardLM” model [20] which empowers the LLaMA foundational model [23] to follow complex instructions. The authors [20] enhanced the foundational model through fine-tuning, employing open-domain instruction data of different complexities, which were generated automatically using language models instead of manually crafted instructions.

We used this generative language model to extract structured information as triples in the form of node-edge-node, and accordingly generate a knowledge graph. Graph representation offers a versatile method of storing structured information [24] since concepts and their relationships can be represented using nodes and edges [25, 26]. Knowledge graphs can be constructed using several approaches ranging from manually crafted knowledge bases to advanced NLP-based pipelines [27]. Here, we propose an NLP pipeline relying on a generative LLM to automatically generate triples and populate a knowledge graph from sustainability reports. This approach for knowledge graph generation has also been explored by other recent research works [28, 29, 30, 31] based on OpenAI’s commercial LLMs. We followed the same research direction but relied instead on an open-sourced LLM [20] and introduced some semantic constraints to extract ESG-focused triples.

Starting from the generated knowledge graph, we created three distinct bipartite graphs [32, 33] to extrapolate meaningful insights. 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-related issues from several perspectives ranging from recognition to partnerships, highlighting the complexity and joint efforts needed to address ESG-related matters. Nevertheless, many companies adopt a shared approach to address certain ESG concerns, such as reducing energy and air emissions. In addition, common strategies emerged among companies from the same geographical region or the same sector. Lastly, our work investigated which factual aspects impact the most on companies’ ESG scores using features based on our findings (actions disclosed) and other company information. This interpretability analysis highlighted that the company’s disclosures impact more than other company or financial aspects on ESG scores. In addition, it unveiled that transparency rewards since disclosing a lot of non-financial information affects ESG scores positively, whereas reporting fewer ESG-related issues hurts scores.

2 Related work

In this section, we first discuss the state-of-the-art approaches for creating Knowledge Graphs (KGs, Section 2.1), encompassing a spectrum ranging from conventional NLP pipelines to the exploitation of Large Language Models (LLMs). In Section 2.2, we then summarise the main studies focused on ESG-related textual information.

2.1 Knowledge Graphs generation

Knowledge Graphs (KGs, [25, 26]) offer a versatile method of representing knowledge that can be leveraged in various use cases and domains [34, 35]. They can be applied to question-answering [36], recommendation systems [37], and information retrieval [24]. KGs bear a closer resemblance to an abstract framework rather than a mathematical structure. Knowledge representation applications vary significantly in their construction, ranging from entirely manually crafted knowledge bases to KGs that are automatically extracted and processed [27, 35]. 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 [35, 38, 39]. Traditional approaches for knowledge acquisition include several entity-oriented tasks and relation-oriented tasks. Named-Entity Recognition (NER), Named-Entity Linking (NEL) and Entity Alignment (EA) are examples of the former. Relation-oriented tasks instead involve extracting, classifying or predicting relations between entities [35, 39]. Thus, KG generation traditionally involves several NLP tasks which are generally disjointly learned through supervised- or semi-supervised learning processes. To overcome error accumulation, new one-stage NLP pipelines have been proposed to extract jointly both entities and relations [40, 41].

In this context, the NLP task of Open Information Extraction (OIE, [42]) could also be exploited to extract structured information from texts. This NLP task aims to extract relationships between two entities from text, without relying on a predefined template or a specific domain. OIE techniques typically extract structured information in the form of subject-predicate-object (SPO) triples, relying on the intrinsic syntactical sentence structure [42]. 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.,

Multi²OIE [43] and DeepEx [44]) have employed transformer-based LLMs (e.g., BERT [45]) to extract both syntactic and semantic information [46].

LLMs trained on large-scale corpora have indeed demonstrated significant potential across diverse NLP tasks [47]. The task adaptability of LLMs mainly stems from the incorporation of the supervised technique of instruction tuning. This technique enables LLMs to handle a wide array of NLP tasks via natural language instructions called prompts. Prompt engineering represents an innovative field focused on creating and refining model instructions to maximise the effectiveness of LLMs across various applications and research areas [47]. A model instruction, or prompt, is a sequence of natural language inputs for LLMs specified for an NLP-related task; it generally consists of three main components: (1) a task instruction, (2) a context, and (3) an input text [47]. Researchers can harness the generalisability of LLMs and the versatility of prompts to address tasks related to knowledge acquisition [47]. For instance, [48] proposed to use LLMs to extract relations and entities from a text corpus for KG generation (i.e., LLM-augmented KG construction). Many researchers [28, 29, 30, 31] have currently been demonstrating 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 [49]), (3) produce natural-language descriptions of KG facts (i.e., KG-to-text [50]) and (4) answer natural-language questions (i.e., question-answering [47]).

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 LLMs, alongside cutting-edge approaches such as the Retrieval Augmented Generation (RAG) paradigm [14] and in-context learning, to address the primary limitation of traditional OIE methods in our context: namely, not exploiting a predefined template for extracting information. An LLM-based information extraction approach allows us to generate semantically-aware and ESG-focused triples rather than traditional SPO triples. This represents a foundation block to generate an ESG-oriented KG.

2.2 Text Analytics on ESG-related information

Text Analytics is an NLP process that combines information extraction, data mining, machine learning, and computational linguistics to extract knowledge from written resources automatically and on scale.

In an ESG-related domain, Text Analytics can be used to process companies’ non-financial textual information and extract meaningful insights concerning statements, facts and actions disclosed by companies. For example, Chou *et al.* [51] 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. Differently, Raghupathi *et al.* [52] conducted an exploratory study to gain insights into the shareholders’ perspectives and objectives regarding sustainability and climate change. The authors [52] processed climate and sustainability shareholder resolutions through traditional approaches ranging from trigram co-occurrences to clustering and classification.

Furthermore, the semantic understanding and flexibility of LLMs can be leveraged for Text Analytics in the ESG domain. Starting from the comprehensive factual knowledge of LLMs [14], researchers have recently augmented it through the incorporation of domain-specific information to enhance downstream tasks. Jacouton *et al.* [53] introduced *SDG Prospector*, a tool exploiting LLMs (DistilRoBERTa [54]) to identify paragraphs in Public Development Banks’ sustainability reports that address SDGs. This is achieved through transfer learning, which allows the LLM to further be trained on a large corpus of text labelled with SDGs. In addition, the authors [53] recently released a public version of the tool [55] enabling the detection and quantification of SDG-related disclosures in any document. Similarly, Webersinke *et al.* [56] introduced *ClimateBert*, a transformer-based language model that is further trained on over 2 million paragraphs of climate-related texts. Afterwards, the authors [56] evaluated its performance by applying it to climate-related classification tasks to evaluate text climate relatedness, text sentiment as well as fact check claims [56]. LLMs can also accomplish question-answering tasks, although they might face two major challenges when answering: relying on outdated information and generating inaccurate information (i.e., hallucinations). The work of Vaghefi *et al.* [57] addressed these challenges by adopting the paradigm of

RAG [14]. They provided GPT-4 with access to external and scientifically accurate sources related to climate. Specifically, they augmented the posed questions by integrating relevant contextual information retrieved from the Sixth Assessment Report, released by the United Nations’ Intergovernmental Panel on Climate Change (IPCC). This approach adopts semantic search to find paragraphs relevant to a question and integrates them into the model query (posed question) via in-context learning. The authors also released a conversational agent [58] based on their proposed approach. Likewise, in the work conducted by Ni *et al.* [59], a tool named *ChatReport* was introduced. This tool adopted an LLM-based approach to evaluate companies’ sustainability reports according to the eleven recommendations [60] provided by the Task Force for Climate-Related Financial Disclosures (TCFD). The authors [59] first employed semantic search on a segmented report to identify text chunks that are pertinent to each recommendation. Afterwards, they prompted an LLM (i.e., OpenAI’s ChatGPT) to summarise the text chunks related to each recommendation, and then benchmark the summarised text against the TCFD guidelines, generating a compliance score ranging from 0 to 100.

Our work differs from other ESG-focused and LLM-based works by jointly (1) leveraging generative LLMs for Knowledge Graph (KG) generation, (2) enhancing in-context learning, (3) adopting an open-sourced generative LLM and (4) relying directly on companies’ sustainability reports. This methodology, alongside the exploitation of bipartite graph representation, allows us to conduct non-trivial analyses concerning ESG-related issues and actions disclosed by companies.

3 Materials and Methods

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

3.1 Data sources

Here, we describe the three main data sources used in our work, which include (i) sustainability reports (Section 3.1.1), (ii) an ESG taxonomy (Section 3.1.2), and (iii) ESG rating scores and other companies’ information (Section 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 [4]. However, ESG reporting can be subjective and opaque due to the complexity of reporting qualitative aspects, particularly those related to social and governance issues [61]. Furthermore, the lack of a standardised framework for ESG reporting makes quantitative/comparative analyses difficult [62, 63].

For this study, we collect 6,456 sustainability reports from 4,222 different companies using the report URLs available on two public websites [64, 65]. 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 work, 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 [66]).

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 over six thousand sustainability reports from four thousand diverse companies poses a significant computational workload. Hence, we endeavour to choose a representative subset of companies striking a balance between sector representation and company prominence. This subset comprises approximately one hundred and twenty-four companies spanning eleven sectors (Table 1).

Sector	Number	Companies
Industrial	20	Airbus, 3M Corporation, Adecco, Daikin Industries, ...
Technology	16	Apple, STMicroelectronics, Intel, NVIDIA, LG Display, ...
Financial Services	13	Bank of America, Deutsche Bank, Visa, Mastercard, ...
Consumer Defensive	11	Coca-Cola, British American Tobacco, Walmart, ...
Consumer Cyclical	11	Amazon, Adidas, Toyota Motor, Alibaba, Tesla, ...
Communication Services	10	Fox, Netflix, Walt Disney, Activision Blizzard, Meta, ...
Healthcare	10	Amplifon, AstraZeneca, Bayer, Johnson&Johnson, ...
Basic 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	

Table 1: 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.

Further details concerning the original dataset, the full list of the selected companies and the fiscal years of the considered reports can be found in the Supplementary Material (SM) document (see Sections SM20, SM21 and SM22).

3.1.2 ESG taxonomy

We adopt the ESG classification taxonomy proposed by Berg *et al.* in [9]. 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 taxonomy comprising sixty-four ESG categories. These categories encompass environmental, social and corporate governance issues such as employee development, supply chain, climate risk management, energy, financial inclusion, biodiversity, customer relationship, access to basic services, board diversity, etc. A complete list of all the ESG categories can be found in the Supplementary Material file (see Section SM19).

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 issues.

Regarding Refinitiv, the data provider used in this work, individual scores represent percentages [67], where lower values (0-25) indicate poor relative ESG performance and a lack of transparency in publicly reporting material on ESG data (i.e., laggard companies). Conversely, higher values (75-100) indicate excellent relative 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 [10, 68].

The company’s ESG performance is assessed relative to industry peers due to industry-specific ESG concerns. For instance, packaging might be more relevant for companies operating in the consumer-defensive sector (e.g., Pepsi), while basic material companies (e.g., DuPont) should stress issues related to employee safety. The combined ESG score is a cumulative measure of E/S/G pillars’ weights, which differ by indus-

try for the first two pillars (E and S), while the weight of the third pillar (G) remains consistent across all industries [67].

For each company, we collect 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 gather the combined ESG scores, individual scores for the E/S/G pillars, company sector, industry, country, region and continent as well as the number of employees, market capitalization, EBITDA (Earnings Before Interest, Taxes, Depreciation, and Amortization), and total liabilities.

3.2 Data preparation

Our NLP pipeline (Figure 1) consists of several components, 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 (Section 3.2.1) as well as using semantic search to select only sentences related to ESG topics (Section 3.2.2). The latter relies on the ESG taxonomy introduced in Section 3.1.2.

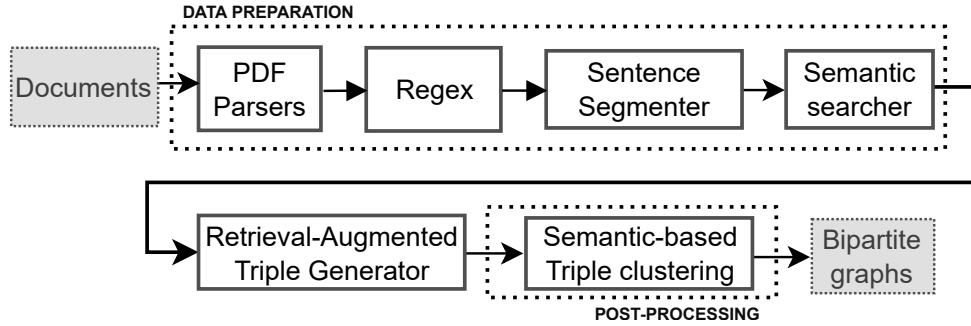


Figure 1: Our proposed approach and its components. Given a collection of textual documents as inputs and preprocessed using different NLP methods, semantically structured insights are extracted by the retrieval-augmented triple generator. Bipartite graphs are created after performing a semantic-based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

3.2.1 NLP pipeline

Sustainability reports are lengthy and visually rich PDF files that generally span more than one hundred pages [69]: our reports’ collection has a median page value of 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, [70]) 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 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 [71, 72], 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”¹. 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 [73]. We adopt the approach of neural semantic search [74, 75, 76, 77], a vector-based search method, that exploits text embeddings to represent both documents and queries in the same vector space. 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 [74].

Our retrieval task involves discovering sentences related to each of the ESG taxonomy categories (Section 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 (sentences) are not interchangeable as they represent different semantic object types and have different lengths; similarly to the question answering framework [75, 76]. In contrast, symmetric semantic search is adopted when queries and corpus documents are interchangeable such as in similar document retrieval systems [78, 79, 80].

After extensive testing, we determine that “INSTRUCTOR-xl” [81], an instruction-tuned embedding model [82], is the most suitable choice for our specific tasks.

Instruction-tuned embedding models enhance the embedding process by adjusting the model output according to different downstream applications by using task and domain descriptions instead of fine-tuning the entire embedding model [82]. Accordingly, this type of embedding model makes use of natural-language instructions to better represent the input contents semantically; thus requiring prompt engineering to some extent. The authors of the model [82] offer a universal format (“Represent the [domain] [text type] for [task objective]:”) along with an examples’ 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”.

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

Table 2: 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* category, which comes from an infographic or a tabular layout.

After generating the text embeddings for a sentence-segmented sustainability report, we retrieve the most semantically relevant sentences for each ESG category (semantic search, Table 2). We set the similarity cut-off threshold t_{sim} equal to 0.6 to retrieve relevant sentences to a given ESG category. Empirical experiments show an acceptable sentence relevance with a similarity threshold equal to or above 0.4. Thus, we adopt a

¹Extrapolated from page 2 of Daikin’s 2022 sustainability report

cut-off point of 0.6 as a good trade-off between sentence coverage and computational workload. In addition, we 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 categories, and actions taken by companies to address issues related to those categories. To achieve this goal, we need to represent ESG-related sentences as triples in a predefined semantic template: category-predicate-object (cat-pred-obj). Precisely, each triple should consist of an ESG category (cat) taken from the ESG taxonomy (Section 3.1.2), a predicate affecting that category (pred), and an entity (obj) related to the ESG category undergoing the predicate.

Consequently, our goal requires knowing the semantic meaning of words as well as defining a semantic template to generate ESG-focused triples. The latest OIE techniques (Section 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*”², conventional OIE techniques [83] would identify and generate a traditional SPO triple as the following: (Microsoft, Invested, 125 million). Although the above 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 involves both information extraction and semantic classification tasks. Secondly, our goal requires extracting ESG-related actions rather than generic statements. Hence, triples should envelop predicates and objects related to an ESG category.

As previously highlighted, LLMs have already demonstrated abilities in semantic understanding and handling a broad range of NLP-related tasks [11, 18]. Accordingly, in this work, we employ instruction-tuned LLMs, the in-context learning technique and the prominent RAG paradigm [14] to address this challenge. RAG is an emerging text generation approach that combines deep learning with traditional retrieval techniques [84] via in-context learning. It stems its power from the ability to combine factual knowledge (stored in model parameters) with external knowledge provided in the model instruction (in-context learning) to generate text that is consistent with both knowledge sources. Our work exploits this text generation paradigm to provide an LLM with an input (ESG-related sentence) and an external context (examples and a semantic schema) to extract structured information from the sentence.

We choose the Kor library [85] 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. Each element of the triple has a unique name and a description conveying the semantic meaning of the associated entity. It then uses this information to generate a textual instruction that is used to prompt an LLM using a predefined template. Its precise content is shown in Figure 2. The full model instruction is reported in the Supplementary Material file (Section SM2).

Our work tested different instruction-tuned LLMs such as Google’s Flan-T5 [86] and LLaMA-based models [23]. We empirically found that LLaMA-based models (e.g., Alpaca [87]) generate better results when prompted to extract structured information. Among the LLaMA-based models, WizardLM-7B [20, 88] is found to produce higher-quality results when prompted to extract structured data through the aforementioned instruction which exploits in-context learning.

Another degree of freedom is the choice of hyper-parameters for text generation such as the temperature and the number of beams. The temperature parameter controls the randomness of the model responses by influencing the model’s confidence in its most likely output [89]. During the decoding phase, this parameter alters the model output by scaling the logits before applying the softmax function. A high temperature

²This sentence is created for explanatory purposes. It does not represent a real fact.

```

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 2: 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.

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 [89]. Accordingly, we opt for greedy decoding to ensure deterministic outputs and to make the generation process adhere to instructions as much as possible [59]. As previously said, another hyper-parameter that affects text generation is the beam number ($b_{dim} = 6$) representing the number of tokens considered during the beam search algorithm [89]. 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 [90]. We found through extensive experiments that the beam number ranging from 4 to 6 strikes a good balance between semantic representation and computational workload.

3.4 Knowledge Graph generation

Before constructing a 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 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 [82] 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 conduct 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*. Other cluster examples and replacing examples are reported in the Supplementary Material file (Section SM1).

3.5 Approaches for statistical analyses

In the Results Section (Section 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 [32, 91, 92]. It can be described through its binary *bi-adjacency* matrix \mathbf{B} , a $\{0, 1\}^{n \times m}$ matrix where n and m are the numbers of nodes in the two partitions.

Specifically, we use three distinct bipartite graphs for our analyses which are obtained through node and edge filtering of the comprehensive KG. In our setting, the graph edges are context-dependent and thus change based on the perspective used to generate the bipartite graph. From this point forward, they are referred to as (1) the category-predicate graph $\mathbb{B}_{\text{catpred}}$, (2) the company-category graph \mathbb{B}_{ccat} , and (3) the company-action graph $\mathbb{B}_{\text{coact}}$. We define an action *act* as the concatenation between the category *cat* and the predicate *pred* of a *cat-pred-obj* triple. For instance, given the ideal triple mentioned in Section 3.3, the action *act* is defined as the category *Waste* concatenated with the predicate *Investment in*, resulting in the action “Waste:Investment in”.

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 Section SM4).

3.5.1 Bipartite graph statistics

Most of the unipartite graph metrics can be extended to the bipartite case [32, 92]. Specifically, here we compute the bipartite variants of network statistics such as degree centrality, closeness centrality and betweenness centrality [93, 92].

The degree centrality of a partition node is the fraction of the nodes of the other partition connected to it [94]. The closeness centrality of a node is instead determined by calculating its average shortest path distance to all other nodes reachable within the same partition and from the other partition [92, 91]. It represents the efficiency of a partition node to be connected directly to nodes from the other partition and indirectly to nodes from the same partition [95]. For instance given \mathbb{B}_{ccat} , 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 [96] assesses the level of influence a node holds over information flow within a graph (control power) by counting the number of times a node lies on the shortest path between two other nodes [91]. In the context of bipartite graphs, it is used to identify nodes serving as critical mediators in enabling interactions between the two separate partitions [32].

3.5.2 ESG-related actions’ variability

We leverage information theory to assess the variety of ESG-related actions disclosed by companies. Specifically, we use Shannon’s entropy (Equation 1, [97]) that measures the information gained (expected uncertainty) from the occurrence of an event $x \in \mathcal{X}$, which is inversely related to its occurrence probability $p(x)$. Rare events convey much information, while those frequent have little information.

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

In our context, the events (\mathcal{X}) represent the disclosed ESG-related actions. High entropy denotes variability in the event occurrences, indicating an ESG category associated with many predicates with an almost uniform probability of occurrence. On the other hand, low entropy indicates the predominance of a limited set of ESG categories.

3.5.3 Similarity analysis

We estimate company similarities based on jointly disclosed ESG-related actions through the Jaccard similarity coefficient [98]. This similarity coefficient has been extensively used in several domains as it provides an understandable manner to quantify the similarity between two mathematical structures (e.g., discrete sets, [98]). It is defined as the intersection cardinality of the action sets divided by the union cardinality of them (Equation 2) and has a numerical range between 0 (no common items) and 1 (equal sets).

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

We first generate a binary matrix representing the presence (1) or absence (0) of an action (column) in the triples of a company (row). We then compute pairwise company similarities by computing similarity coefficients between ESG-related action sets.

However, identifying non-random structural patterns can pose statistical challenges since randomness can introduce noise in the pattern distribution [99]. In the context of binary matrices, the predominant method to address the influence of stochastic fluctuations on pattern recognition involves null models [100]. This approach compares the observed structural patterns with those obtained from a set of randomised versions of the original matrix (e.g., null model, [100]). Accordingly, we generate a null model through a bootstrapping technique by computing company similarities on the randomised binary matrix through 1,000 simulations. We subsequently subtract this null model from the observed company similarities.

3.5.4 Correlation analysis

We aim to evaluate whether company similarities in terms of disclosed ESG-related actions are correlated to similarities in ESG scores or other company characteristics such as geographical location or market capitalization. We first measure feature similarities through different strategies, ensuring the same numerical range and monotonicity. Company similarities concerning textual features such as company sector, country and region are embedded using the “INSTRUCTOR-xl” embedding model [82] and their semantic similarities are assessed through the cosine similarity. Instead, the similarities in numerical features, such as ESG scores, market capitalization and the number of employees, are measured by computing the absolute numerical difference normalised using max-min scaling as formulated in Equation 3.

$$\hat{x}_{\text{sim}} := \frac{x_{\Delta} - x_{\Delta_{\max}}}{x_{\Delta_{\min}} - x_{\Delta_{\max}}} \quad x_{\Delta} := |x_2 - x_1| \quad (3)$$

The majority of these features measure company similarities in the range between 0 and 1, in which zero represents the most dissimilar company value and one means equal company values. However, cosine similarity maps to values in the range from -1 to 1, in which negative values indicate opposite words embedded in the vector space. Hence, we re-scale the cosine similarities using min-max scaling to ensure the same zero-to-one range of the other similarity features. A complete list of all the features and measures used can be found in the Supplementary Material (see Section SM9).

Afterwards, we perform a bivariate analysis through a correlation analysis to measure the monotonic association between action similarities and similarities of other company features. The relationships between magnitude changes of variables are typically measured through a correlation coefficient defined between -1 and 1, expressing respectively negative and positive correlated monotonic changes [101].

We rely on Kendall’s τ correlation coefficient (Equation 4, [102]), a non-parametric and rank-based method.

$$\tau = \frac{n_c - n_d}{n_c + n_d} \quad \text{s.t. } n_c = \text{number of concordant pairs} \quad (4)$$

$$n_d = \text{number of discordant pairs}$$

Rank-based correlation methods overcome some limitations of traditional correlation methods such as the well-known Pearson correlation coefficient [103]: they can measure nonlinear monotonic relationships, are more robust against outliers and normality assumption is not required [101]. Kendall’s rank correlation tests the similarities in data ordering when it is ranked by quantities such as company similarities. It measures the strength of monotonic association based on the pattern of concordance and discordance between observation pairs [104]: concordant pairs are ordered in the same way (order consistency), while discordant pairs are ordered differently (order inconsistency) [102]. 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 [102].

3.5.5 Interpretability of ESG scores

Lastly, we investigate the interpretability of ESG scores through a linear regression and the SHAP (SHapley Additive exPlanations) framework [105]. The ESG scores are created by rating agencies and are based on several unknown attributes combined with undisclosed weights. Here, we investigate which factual aspects impact the most on companies' ESG scores by exploiting the interpretability of a first-order linear regression model.

The model predictors are features based on our findings and other company information (Section 3.1.3). For each company, we compute the category and action entropy, representing the diversity, and indirectly the cardinality, of the ESG categories and actions disclosed by each business (Section 3.5.2). In addition, we consider the percentages of the top ten most disclosed categories and compute the proportion of all disclosed categories related to the three E/S/G pillars. Lastly, we consider as predictors nine company features encompassing company details (e.g., sector, region, and incorporation date) and annual financial data (e.g., market cap, and liabilities). Categorical variables, such as the company sector, are turned into binary dummy variables [106], generating a total of 97 features, whereas standardisation is applied to numerical features. Further details and an example observation are shown in the Supplementary Material (see Section SM12).

Then, we perform an Ordinary Least Squares (OLS) regression with Elastic Net Regularization to perform inference on ESG scores. OLS models [107] find the optimal model coefficients by minimizing the residual sum of squares between the observed dependent variable (ESG score) and the output of the linear function of the independent predictors. Regularization methods can be adopted to tackle overfitting by introducing penalties during least square optimization [107, 108]. Although the task is not to predict unseen data, mitigating overfitting still matters since a trade-off should be found between reducing the model errors and interpreting the actual relationship between the dependent variable and the predictors [109]. These penalties work as shrinkage methods for feature selection [108] by shrinking feature coefficients toward zero and each other in Ridge regression (e.g., L_2 norm, squared magnitude), or imposing some feature coefficients to zero in LASSO (Least Absolute Shrinkage and Selection Operator) regression (e.g., L_1 norm, absolute magnitude). Although LASSO regression has shown success in many situations [110], it exhibits some limitations when the number of predictors ($|features| = 97$) is higher than the observations ($|companies| = 89$) as well as in presence of strong pairwise correlations. Elastic Net Regularization [108] can be viewed as a generalization of LASSO that improves the model performance in these scenarios by linearly combining the L_2 and L_1 penalty methods. The Elastic net cost function [110] is defined as:

$$\text{ElasticNetLoss}_{\theta} \left(y, y_{\text{pred}} \right) = \text{MSE}(y, y_{\text{pred}}) + \alpha L_1 \text{ratio} \sum_{i=1}^m |\theta_i| + \alpha(1 - L_1 \text{ratio}) \sum_{i=1}^m |\theta_i| \quad (5)$$

The parameter $L_1 \text{ratio} = 0.5$ determines the proportion between the two penalties, while α is a multiplier factor for both penalty terms and θ are the model parameters.

We employ an eight-fold cross-validation approach [111], with the above cost function, to estimate the optimal α value ($\alpha = 0.49$, see Section SM14). We here report the performance of the OLS regression through different metrics (see also Section SM13 in the Supplementary Material document). Firstly, the Coefficient of Determination (R^2 , [112]) 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, [113]) 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 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, [114, 108]), 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 [115] 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 predicted 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 Section SM15).

Subsequently, we employ the SHAP framework [105] to investigate which predictors impact the most on the inference of ESG scores. SHAP is an additive feature importance measure based on cooperative game theory and enhances the interpretability of machine learning models. This model-agnostic method provides a local interpretation of a model prediction as an additive sum of the directed effects of each predictor. It measures the impact of each predictor towards the model output by evaluating a conditional expectation function. SHAP starts from the prior knowledge of the expected model output $E[f(X)]$. It then evaluates, for each model prediction, the magnitude and direction changes in this expected value when conditioned on each predictor (SHAP values). Features with positive SHAP values affect the expected model output with additive increments, conversely, those with negative values have additive decrement effects. The magnitude quantifies the intensity of this additive effect.

4 Results

Here, we first examine the quality of the triples generated (Section 4.1). We then report the network statistics computed at the node level from the three bipartite graphs in Section 4.2. In addition, Section 4.3 shows some descriptive statistics and insights concerning ESG-related actions. Afterwards, we report company similarities based on jointly disclosed ESG actions (Section 4.4). The follow-up section (4.5) 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 (Section 4.6).

4.1 Comparative analysis of the generated triples

Overall, the triples generated exhibit high-quality semantically structured information, for instance, triples B, C, D and F in Table 3 represent meaningful ESG-related actions. However, there are also some unsatisfactory triples mainly due to information incompleteness. For instance, in row A the lack of a meaningful object (obj) in the triple results in the omission of important information related to the two mentioned cat-

	ESG category (cat)	Predicate (pred)	Object (obj)
A	Remuneration	Compensation	This compensation falls into two categories
B	Employee Development	Providing access to	Career development programs, ongoing inclusion and diversity education, and support throughout their career journey
C	Data Privacy	Establishment of	A corporate network security evaluation system
D	Green Buildings	Equipped with	Dematerialised management methods
E	Air Emissions	Conversion to	Emissions of PFCs and other greenhouse gases
F	Environmental	Alignment with	The SASB 2018 Real Estate Standards and the TCFD

Table 3: Examples of the triples *cat-pred-obj* generated through our approach. The six triples were picked randomly from all the nearly 50 thousand generated triples.

egories. Similarly, the triple in row E has a partial satisfaction level since it does not entirely encompass the meaning conveyed in the original sentence, which involves the conversion of these emissions into CO₂.

Furthermore, since the model instruction has a great impact on the quality of the generated text [116], we conduct an ablation study to compare qualitatively the triples generated through different prompt templates: with(out) in-context learning (EXAMPLES in Figure 2) and with(out) the semantic output schema (DATASHEMA in Figure 2).

	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 4: Three comparing examples of the triples generated with/without using in-context learning in the model instruction.

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 4. 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.

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

Table 5: Three comparing examples of triples generated with/without the semantic output data schema in the model instruction.

Moreover, adding a semantic output schema in the prompt (DATASHEMA) helps the generative language model to better focus on ESG-related information as exhibited in Table 5. 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 change the extracted information, especially in sentences with multiple prepositions, such as in the second comparison shown in Table 5. Furthermore, defining a semantic schema allows us to incorporate the ESG taxonomy (Section 3.1.2) into the semantic description of the item cat. This enhanced the model’s ability to extract an ESG-related issue from a sentence (cat) by emulating a supervised text classifier using the ESG taxonomy categories as semantic labels while leveraging its generative capabilities. An example of this enhancement can be observed in the third comparison in

Table 5, while the complete semantic schema can be seen in the Supplementary Material document within the full model instruction (see Section SM2).

4.2 Bipartite graphs’ analysis

We report the statistics for the three bipartite graphs $\mathbb{B}_{\text{catpred}}$, $\mathbb{B}_{\text{cocat}}$ and $\mathbb{B}_{\text{coact}}$. The distribution of degree centrality of the category-predicate bipartite graph $\mathbb{B}_{\text{catpred}}$ has an average degree centrality less than 1%, indicating a variety of different actions disclosed for ESG categories. However, there are predominant predicate nodes that are associated with more than ninety ESG categories (degree $\geq 16.6\%$) such as *Commitment* and *involvement with* (with 113 categories), *Advisor support for* (102), *Partnership with* (97), and *Establishment of* (94). The closeness centrality values of the prominent nodes of this bipartite graph exhibit higher values apparently in contrast to the relatively low values of the degree centrality. This means that the predicate nodes are not directly connected to a large number of category nodes ($< 20\%$), yet these nodes are still well-connected indirectly to other nodes in their partition ($> 88\%$) through triangular paths. Thus, these predicate nodes are indirectly connected through common category nodes.

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

Table 6: Graph metrics of a sample of all the category nodes of the bipartite graph $\mathbb{B}_{\text{cocat}}$.

The company-category bipartite graph $\mathbb{B}_{\text{cocat}}$ has fewer nodes (124 company nodes and 542 category nodes) and results more connected (graph density of 10.8%). This makes its degree distribution more broaden with an average degree equal to almost 11%. Moreover, a widening area can be observed in the degree distribution for values greater than 80%. This is caused by mainstream and cross-sector ESG categories such as *Climate Risk Management*, *Supply Chain*, *Energy* and *Corporate Governance* (Table 6), which are connected to, and thus disclosed by, almost all the companies (degree $> 92\%$). However, it is also worth highlighting the ESG-related issues that are surprisingly disclosed by a few companies (Table 6): for instance, *Market Responsibility*, *Anti-Discrimination* and *LGBTQ+ Inclusion* are disclosed by less than 5% of all the considered companies.

Lastly, the company-action bipartite graph $\mathbb{B}_{\text{coact}}$ exhibits a rather skewed distribution of degree centrality. However, that is mainly due to a few predominant actions connected to and thus reported by several companies. Examples of such outliers include *Air Emission:Reduction of* and *Energy:Reduction of* that are disclosed respectively by 70% and 60% of all the considered companies. Further statistics and extensive tables for all three bipartite graphs are shown in the Supplementary Material document (see Sections SM4, SM5 and SM6).

4.3 Diversity analysis on disclosing ESG categories

In this section, we analyze the diversity of the ESG categories reported in the sustainability reports. We first assess the variety of the actions disclosed by companies in terms of predicates (pred) for each ESG category (cat) through Shannon’s entropy (Section 3.5).

The predicates associated with each ESG category vary significantly (average entropy equal to 1.5 nats), particularly in umbrella ESG categories such as *Corporate Governance*, *Human Rights*, and *Supply Chain* (Table 7). There is also a significant correlation ($\text{corr} = 0.84$) between the entropy and the number of companies making disclosures on an ESG category, suggesting variations in the way companies address

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: 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-related issues. For example, when addressing Product Safety, companies approach it from different perspectives (Table 7), ranging from developments (2.3%) to regulatory compliance (2.3%) and assessments (4.4%).

Table 8 presents a sample of all the 542 ESG categories extracted from the sustainability reports. The coverage of these categories is reported using three metrics: the percentages of triples including a category as well as the percentages of companies and sectors reporting it. Additionally, the category coverage is better investigated by reporting the percentages of company triples aggregated by sector. A proportion of almost 12% among all these ESG categories is disclosed across all company sectors, encompassing various umbrella aspects such as Climate Risk Management, Supply Chain and Business Ethics. It is noteworthy that certain sectors emphasise specific issues more than others (Table 8). For example, Packaging is more emphasised in Consumer Defensive 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 triples within this sector, while is less stressed in the Consumer Cyclical sector (1.5%) which encompasses companies such as Amazon (4%), The Home Depot (4%), and the Alibaba Group (3%).

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

Table 8: Sample of the ESG categories disclosed by companies in their sustainability reports.

The distribution of the actions disclosed by companies is sparser since ESG-related issues are faced through several approaches. The same action is reported on average by less than 2% of the considered companies, although there are some widely disclosed actions such as those mentioned in the previous section (4.2). Whereas, on average only 15% of the company sectors are engaged in the same action, highlighting different sector priorities. For example, the Assessment of aspects concerning Climate Risk Management is more emphasised by Real Estate companies such as Park Hotels Resorts (2% of all the company triples) and Sun Communities (1%). Conversely, companies in the Basic Materials sector, such as United States Steel

(1%), Yamana Gold (1%) and Aluminum Corporation of China (1%), emphasise instead the Commitment and involvement concerning Employee Safety.

Further ampler tables can be found in the Supplementary Material document (see Section SM3).

4.4 Company similarities based on disclosed ESG-related actions

Here, we evaluate company similarities based on jointly disclosed actions using the Jaccard similarity coefficient (see Section 3.5).

Relatively high similarities appear between companies from the same sectors (Figure 3 and Table 9). For example, five companies out of the ten most similar to Deutsche Bank (Financial) are banks too (Royal Bank of Canada (similarity equal to 7%), Banco Santander (6%) and UniCredit (6%)).

Moreover, relatively high similarities emerge also from companies from the same geographical area (Figure 3). For example, 80% of the ten most similar companies of Sony (Japan, Eastern Asia) are companies from East Asia (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 area (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. More details can be found in the Supplementary Material (see Section SM8).

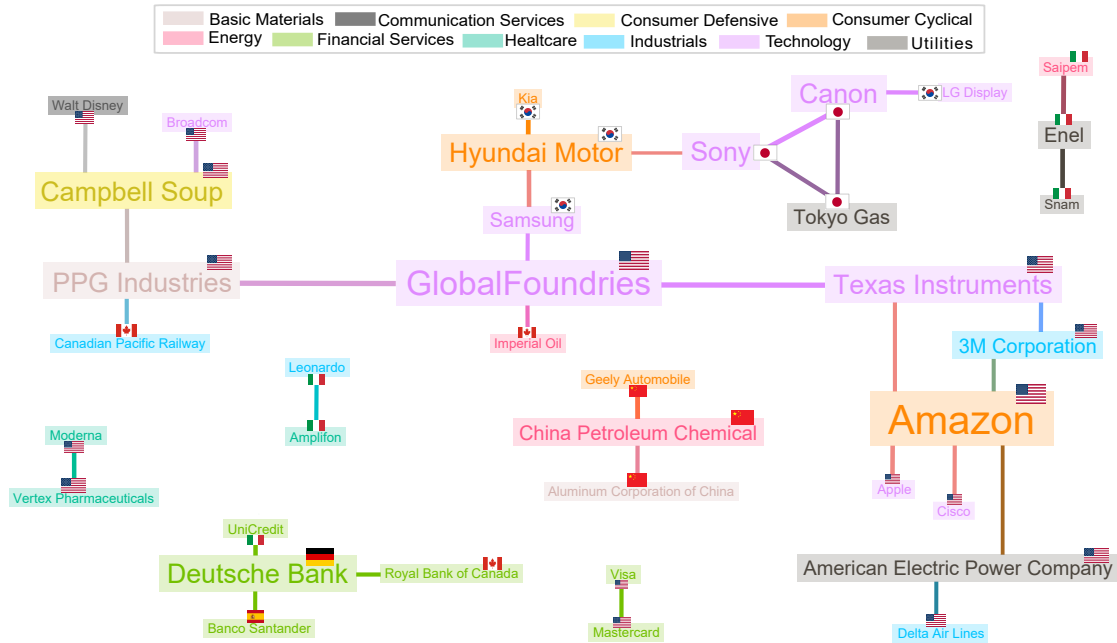


Figure 3: A network diagram linking companies that report similar actions, determined by the Jaccard similarity coefficient. It exhibits only connections between companies with a similarity equal to or greater than 6%. Node colour corresponds to distinct sectors, and node size is proportional to their connectivity. Some connections are noteworthy for linking companies within the same sectors or geographical regions.

4.5 Correlation analysis among company similarities

This section answers the research question concerning whether company similarities in terms of actions disclosed (Section 4.4) are associated with similarities in other company information (Section 3.1.3).

We first create comparable similarity measures for all the features and then perform bivariate correlation analysis, through Kendall's correlation coefficient (Section 3.5). This bivariate analysis assesses the monotonic associations between action similarities and similarities of other company information. We compute

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%)

Table 9: A company sample with the top three most reported actions and the most similar companies for each. The company similarity is assessed by computing the Jaccard similarity on the action set of companies.

correlations for each company with all information available (81%) and report aggregated findings as displayed in Figure 4.

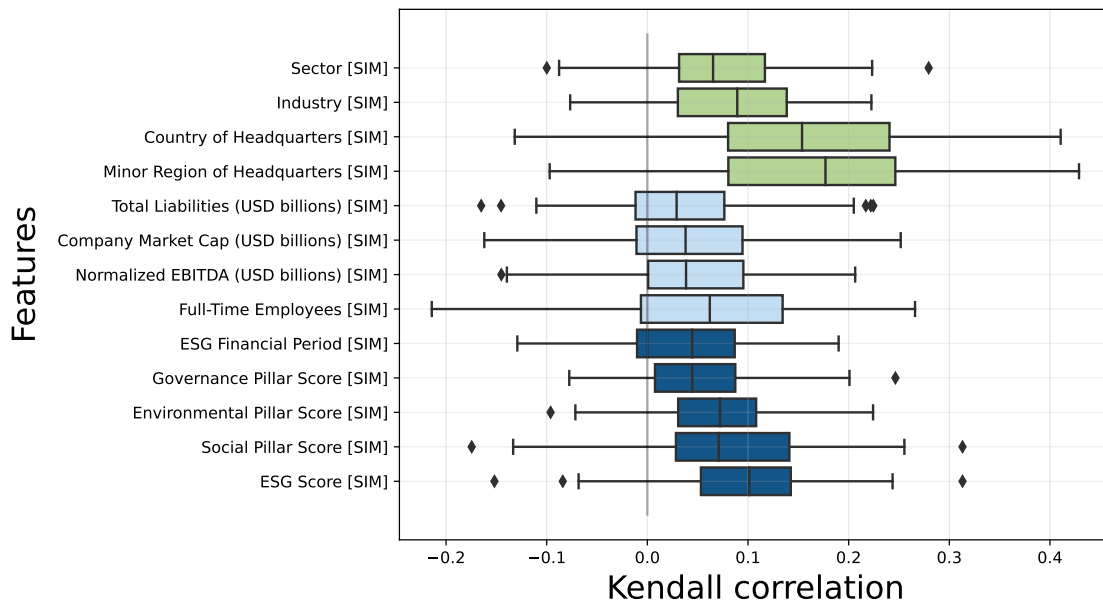


Figure 4: Distributions of the correlation between company similarities in terms of jointly disclosed actions and other company feature similarities. Features are color-grouped according to their type of information and measure type. Light-green features are generic categorical company characteristics, while azure and dark blue features represent numerical features concerning companies' financial and ESG information.

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, respectively. In addition, only these two features demonstrate median p-values, resulting from the null hypothesis test of zero monotonic correlation, below the established accepting threshold of 5%, respectively 1% and 2%. The p-value distributions

for all the features are shown in the Supplementary Material (see Section SM10). These two monotonic associations also confirm the findings of the previous section (4.4) in which company action similarities are empirically associated with the geographical location of the companies. Taking the same example companies of the previous section, Sony and Enel exhibit a relatively high monotonic correlation between company similarities in terms of actions disclosed 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).

Regarding the similarity correlations with other company features, the ESG score and the company Industry exhibit respectively median correlation values equal to 0.1 and 0.09. However, their statistical significance appears relatively weak also due to high median p-values (13% and 15%), which suggest accepting the null hypothesis of zero monotonic correlation.

After analysing company similarities from the action perspective, a pairwise correlation analysis is performed to unveil monotonic associations concerning company similarities among all these company features. Strong monotonic correlations appear between similarities in the company Region and Country (median correlation equal to 0.7) as well as between the ESG score and the Social (0.5) and Environmental Pillar score (0.4). No other monotonic correlations between the ESG score and company information emerge. A graphical representation of all these pairwise correlations can be found in the Supplementary Material (see Section SM11).

4.6 Interpretability of ESG scores

Lastly, we investigate the interpretability of ESG scores by employing a first-order linear regression and the SHAP framework (Section 3.5.5). Specifically, we evaluate how various factual and corporate aspects have the greatest influence on these scores (Figure 5).

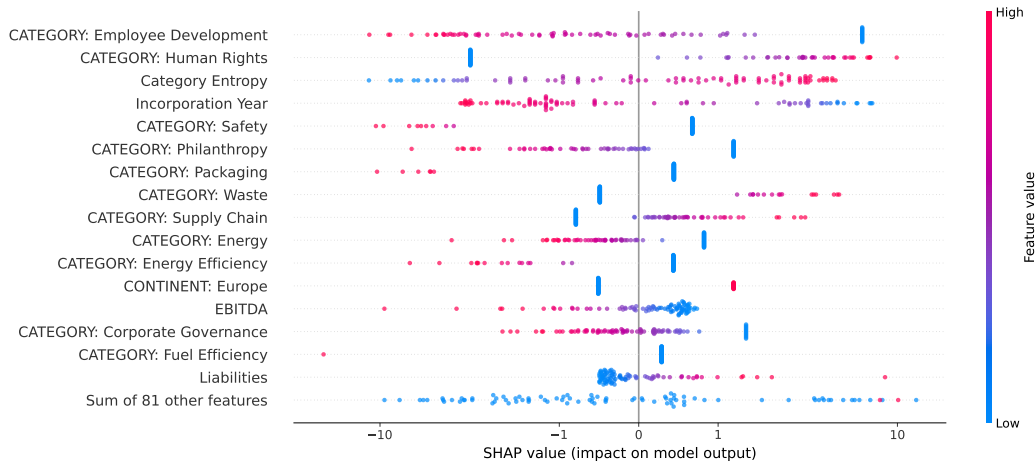


Figure 5: Summary of the top sixteen features impacting the most the inference of ESG score. The features are ordered according to their median shape value. The x-axis represents the degree of a positive and negative impact on model output. Each dot represents a company instance and colours represent the company values of the standardised feature.

On average, the most impacting aspects affecting ESG scores are the percentages of actions related to Human Rights and Employee Development disclosed by companies in their sustainability reports, with a mean SHAP value of 2.6 and 2.7 respectively. High percentages of the former (colour scale in Figure 5) positively impact ESG scores, while the latter has the effect of hurting scores. Similar negative effects are exhibited through high disclosing percentages in issues related to Philanthropy (mean SHAP value of 1.1) and Energy (0.7). Exhibiting a similar average magnitude, but an opposite effect, disclosing several issues related to Waste (0.8) or Supply Chain (0.7) has a positive impact. Moreover, a high diversity, and thus cardinality, of the ESG categories disclosed in sustainability reports by companies (Category Entropy) positively affects ESG scores with a mean SHAP value of 2.1. Sharing a similar magnitude (1.9), being founded

earlier, represented by an older Incorporation Year, positively impacts a company's score. Further noteworthy aspects impacting positively 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 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 Section SM16 in the Supplementary Material document).

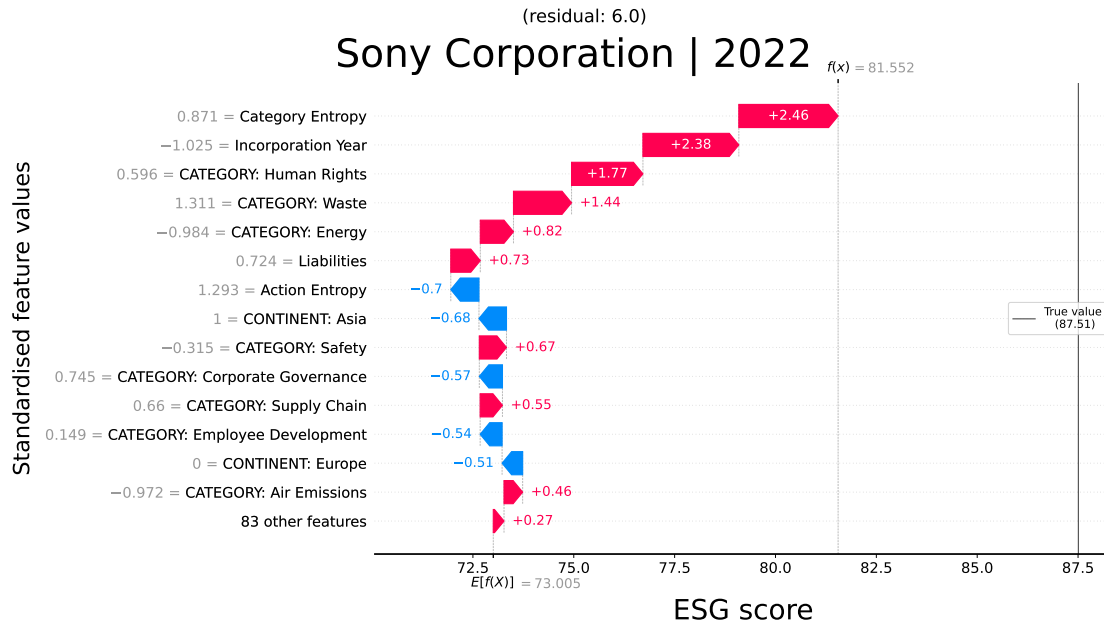


Figure 6: Example of explanations for individual predictors for the ESG score of a company. The category-based features are extracted from the 2022 sustainability report of the Japanese company and other company information from the same fiscal year is considered. In addition, the actual ESG score and the model error (residual) are shown.

Moving from global to local interpretability, we choose Sony as an example company and investigate the most impacting factors for its ESG score (Figure 6). The Incorporation Year and Category Entropy features, respectively far below (standardised value of -1.03, representing the year 1946) and above (0.87) the average of company values, positively affect its score. In addition, disclosing several actions related to Human Rights (0.57) and Waste (1.31) has a positive impact. In contrast, disclosing fewer Energy-related actions than the average (-0.98) positively affects its scores. Interestingly, being an Asian company (CONTINENT:Asia and CONTINENT:Europe) slightly hurts its score.

Lastly, we conduct a more granular analysis by exploring the ESG score interpretability of a company cluster. Following 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). These twelve companies operate in eight different sectors, covering the majority (70%) of the considered sectors. The Incorporation Year strongly affects the ESG scores of Asian companies (average SHAP value of 2.7), in line with the global interpretation. All companies established in the 20th century, such as Toshiba (1904, score of 93.6), Toyota (1937, score of 84.5) and Geely (1946, score of 75.4), exhibit ESG scores above 66. Whereas those established in this 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 Chinese companies consistently have ESG scores below 62.5, whereas both Japanese and Hong Kong-based companies consistently display higher scores. In addition, 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 lack hurts (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 Sections SM16, SM17 and SM18).

5 Discussion

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

5.1 Practical implications

As highlighted in Section 4.2 and 4.3, companies address ESG-related issues 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 regulatory 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 aspects are addressed through a common strategy by the majority of the companies such as reducing air emissions and energy as well as donating and advisory support concerning philanthropic aspects.

Concerning the topics disclosed, our methodology extracts more than five hundred ESG-related issues from companies' sustainability reports, representing an eight-times greater set of issues originally included in the ESG taxonomy used in this work (section 3.1.2). Firstly, this unveils the broad scope of the ESG phenomenon involving socially responsible issues ranging from waste management and supply chain to employee safety and tax compliance. Secondly, this highlights the presence of widely disclosed issues, such as supply chain, and sector-specific issues such as packaging for the consumer defensive sector. The diversity analysis reported in Section 4.3 confirms this sector-based importance for certain topics. For instance, water-related issues are more stressed by companies operating in water-consuming sectors such as consumer defensive (e.g., Coca-Cola) and basic materials (e.g., DuPont) rather than financial services companies (e.g., Goldman Sachs). These analyses confirm the difficulty of defining, framing, and standardising ESG reporting [63] and comparing companies' ESG-related actions.

The findings reported in Section 4.4 emphasise company similarities based on their sectors, confirming indirectly a relatively high presence of common strategies mainly among companies from the same sector. However, the most impacting factor in grouping companies based on their disclosures is their geographical location as shown in Section 4.4 and Section 4.5. This represents an interesting finding of our work emphasising the external pressures on companies to operate in a free market economy considering also their social responsibility, and thus, to be accountable for non-financial aspects too. This spotlights the influence of the company's origins in terms of regulatory compliance and cultural factors, although businesses nowadays operate in a global market. European, American and Asian companies might prioritise socially responsible efforts, investments and disclosures based on the demands coming from their own region. For instance, European companies' climate-related efforts could be driven by more stringent and pioneering climate regulations in Europe, such as the European Union's Emissions Trading System [117] or the ambitious "Fit for 55" plan [118] recently proposed by the European Union to make this region climate-neutral by 2050. This may also be confirmed by averaging the environmental pillar scores by region: European companies have an average score of 81.8, Asian companies have an average score of 69 and American companies exhibit an average score of 65.6 (see Section SM16 in the Supplementary Material document).

The bivariate correlation analysis reported in Section 4.5 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 company region and country (median correlation of 0.7) as well as between the ESG score and the social (0.5) and environmental pillar score (0.4). These two appear fairly trivial associations: first, the ESG score

is a weighted score combining the scores of the three E/S/G pillars (Section 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 might 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).

The interpretability analysis of ESG scores reported in Section 4.6 highlights that the company’s disclosures impact ESG scores more than other financial aspects or company characteristics. Disclosing plenty of non-financial information (category entropy) positively affects ESG scores, whereas fewer disclosures hurt scores. Thus, transparency rewards. This analysis also confirms the negligible impact of governance-related issues towards ESG scores in comparison to social- and environmental-related issues such as human rights, energy and waste. Other noteworthy findings of this analysis are that being a European company and being established earlier as well as having high liabilities have a slight, yet positive, impact on ESG scores, while high annual earnings (EBITDA) slightly hurt scores. Moreover, the region-based interpretability analysis of ESG scores (Section 4.6) spotlights that being a Chinese company hurts ESG scores. This effect might be explained by low Chinese standards concerning ESG-related issues as well as the fact that all the considered Chinese companies were established between 1999 and 2006, likely due to its remarkable economic development starting in the 2000s, and thus associated with the negative impact of being relatively young companies (Incorporation Year). The negative impact of disclosing more actions related to employee development or energy is still unknown and further analyses might better investigate it. One hypothesis is that their high presence might unveil a lack of other ESG-related issues important for the company sector. For instance, in the example of 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.

5.2 Methodological implications

As mentioned in Section 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. This allows us to generate semantically-aware and ESG-focused triples instead of traditional SPO ones. This is pivotal in generating all the meaningful findings and statistics of our work.

In addition, the flexibility and generative abilities of these language models allow us to highlight, and overcome some limitations in the data sources such as those of the ESG taxonomy. This taxonomy extrapolates a concise set of ESG topics by categorising several ESG-related indicators shared among ESG rating providers (Section 3.1.2). However, defining the scope of the ESG phenomenon from the perspective of rating agencies might result in a biased or incomplete set of ESG-related issues. Indeed, our LLM-based methodology extracts, as mentioned earlier, an eight-times greater set of ESG categories than those included in this taxonomy. For instance, our methodology unveils Education as a pivotal ESG-related issue disclosed by more than two-thirds of the selected companies. Education is not explicitly included in the taxonomy, although it could be framed within three taxonomy categories: Access to Basic Services, Human rights (Art. 26) or Philanthropy. Additional examples include Circular Economy which might fall under the taxonomy categories of Waste or Resource Efficiency as well as Air Quality which could be framed within Green Buildings or Health and Safety.

Accordingly, the aforementioned ESG taxonomy encompasses critical issues hidden within vague categories or potentially overlooks them altogether, resulting in a reduction of substantial significance in subsequent analyses. This drawback is overcome by our methodology that exploits generative language models and in-context learning to jointly emulate the output of a supervised text classifier, whose labels are the ESG categories, and generalise those labels semantically. This helps us to extract more suitable topics while keeping the domain and semantics of the original ESG taxonomy. This generative-based and

semantically-aware classification can also be leveraged in several NLP-related scenarios to enhance, using a bottom-up approach, the limited set of known labels. Nevertheless, this could also lead to the undesirable phenomenon of over-specialization which was tackled using semantic clustering (Section 3.4).

Lastly, our methodology differs from other recent ESG-focused and LLM-based tools (Section 2) such as ChatClimate [57] and ChatReport [59] 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

Our data preparation NLP pipeline relies on a PDF parser [70] 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 2 in Section 3.2.2 exhibits an example of this issue in the first sentence retrieved for the Waste category. The sentence contains some details about the ESG-related issue, but 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 (Section 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 (Section 4.6) unveils more impacting factors or relationships for a specific company cluster in comparison to the global interpretability (Section 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 (Section 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 (Section 4.5 and 4.6) might vary using ESG scores from different rating agencies such as Sustainalytics which adopts a risk-based assessment [119]. 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

LLMs can be versatile tools to accomplish diverse NLP-related tasks such as extracting structured information from textual data. We further explored this promising research direction by adopting in-context learning and RAG to extract ESG-related information as semantically structured triples. We then adopted

a graph representation 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 taxonomy, allowing us to generate both semantically-aware and ESG-focused triples instead of traditional subject-predicate-object triples. This helped us to yield meaningful findings such as statistical, similarity and correlation analyses on the ESG-related 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.

Supplementary Material

Further and extensive tabular and graphical views are exhibited in the Supplementary Material (SM) placed after the reference section.

Availability of data and materials

The sustainability reports used, processed and analysed during the current study can be publicly retrieved from the following websites: sasb.org [64] and responsibilityreports.com [65]. The complete list of the companies we considered can be found in Section SM21 of the supplementary material document. The ESG taxonomy adopted during the current study was extrapolated from Table IV (4) of the work [9] by Berg *et al.* It is also exhibited in Section SM19 of the supplementary material document. 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 raw findings are available from the corresponding author on reasonable request.

Funding

The work of JS has been partially funded by Ipazia S.p.A.

Author’s 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.

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).
- [3] European Union. Corporate sustainability reporting directive (csrd).
- [4] European Union. Corporate sustainability reporting.

- [5] Christina Wong and Erika Petroy. Rate the Raters 2020: Investor survey and interview results. Survey report, SustainAbility Institute by ERM, 2020.
- [6] Aaron K Chatterji, Rodolphe Durand, David I Levine, and Samuel Touboul. Do ratings of firms converge? implications for managers, investors and strategy researchers. *Strategic Management Journal*, 37(8):1597–1614, 2016.
- [7] Subhash Abhayawansa and Shailesh Tyagi. Sustainable investing: The black box of environmental, social, and governance (esg) ratings. *The Journal of Wealth Management*, 2021.
- [8] Monica Billio, Michele Costola, Iva Hristova, Carmelo Latino, and Loriana Pelizzon. Inside the esg ratings:(dis) agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5):1426–1445, 2021.
- [9] Florian Berg, Julian F Koelbel, and Roberto Rigobon. Aggregate confusion: The divergence of esg ratings. *Review of Finance*, 26(6):1315–1344, 2022.
- [10] Torsten Ehlers, Ulrike Elsenhuber, Kumar Jegarasasingam, and Eric Jondeau. Deconstructing esg scores: How to invest with your own criteria? *IMF Working Papers*, 2023(057):A001, 2023.
- [11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [12] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [13] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.
- [14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [16] Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309, 2023.
- [17] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [18] Laria Reynolds and Kyle McDonell. 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*, pages 1–7, 2021.
- [19] Yufei Wang, Wanjuan Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey, 2023.
- [20] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions, 2023.
- [21] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.
- [22] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.

- [23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- [24] Ridho Reinanda, Edgar Meij, Maarten de Rijke, et al. Knowledge graphs: An information retrieval perspective. *Foundations and Trends® in Information Retrieval*, 14(4):289–444, 2020.
- [25] Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, Alexander Wahler, Dieter Fensel, et al. Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*, pages 1–10, 2020.
- [26] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.
- [27] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, 9(1):77–129, 2018.
- [28] Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. Iterative zero-shot llm prompting for knowledge graph construction, 2023.
- [29] Lars-Peter Meyer, Claus Stadler, Johannes Frey, Norman Radtke, Kurt Junghanns, Roy Meissner, Gordian Dziwis, Kirill Bulert, and Michael Martin. Llm-assisted knowledge graph engineering: Experiments with chatgpt. *arXiv preprint arXiv:2307.06917*, 2023.
- [30] Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. Enhancing knowledge graph construction using large language models, 2023.
- [31] Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. Llm for knowledge graph construction and reasoning: Recent capabilities and future opportunities, 2023.
- [32] Armen S Asratian, Tristan MJ Denley, and Roland Häggkvist. *Bipartite graphs and their applications*, volume 131. Cambridge university press, 1998.
- [33] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, 2006.
- [34] Xiaohan Zou. A survey on application of knowledge graph. In *Journal of Physics: Conference Series*, volume 1487, page 012016. IOP Publishing, 2020.
- [35] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.
- [36] Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 792–802, 2021.
- [37] Bilin Shao, Xiaojun Li, and Genqing Bian. A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph. *Expert Systems with Applications*, 165:113764, 2021.
- [38] Dieter Fensel, U Simsek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler. *Knowledge graphs*. Springer, 2020.
- [39] Jihong Yan, Chengyu Wang, Wenliang Cheng, Ming Gao, and Aoying Zhou. A retrospective of knowledge graphs. *Frontiers of Computer Science*, 12:55–74, 2018.
- [40] Arzoo Katiyar and Claire Cardie. 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)*, pages 917–928, 2017.

- [41] Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*, 2020.
- [42] Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A survey on open information extraction. *arXiv preprint arXiv:1806.05599*, 2018.
- [43] Youngbin Ro, Yukyung Lee, and Pilsung Kang. Multi^2OIE: Multilingual open information extraction based on multi-head attention with BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1107–1117, Online, November 2020. Association for Computational Linguistics.
- [44] Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. Zero-shot information extraction as a unified text-to-triple translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1225–1238, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [46] Pai Liu, Wenyang Gao, Wenjie Dong, Songfang Huang, and Yue Zhang. Open information extraction from 2007 to 2022 – a survey, 2022.
- [47] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*, 2023.
- [48] Abhijeet Kumar, Abhishek Pandey, Rohit Gadia, and Mridul Mishra. Building knowledge graph using pre-trained language model for learning entity-aware relationships. In *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 310–315. IEEE, 2020.
- [49] Jiao Chen, Luyi Ma, Xiaohan Li, Nikhil Thakurdesai, Jianpeng Xu, Jason H. D. Cho, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Knowledge graph completion models are few-shot learners: An empirical study of relation labeling in e-commerce with llms, 2023.
- [50] Agnes Axelsson and Gabriel Skantze. Using large language models for zero-shot natural language generation from knowledge graphs, 2023.
- [51] Christine Chou, Robin Clark, and Steven O Kimbrough. What do firms say in reporting on impacts of climate change? an approach to monitoring esg actions and environmental policy. *Corporate Social Responsibility and Environmental Management*, 2023.
- [52] Viju Raghupathi, Jie Ren, and Wullianallur Raghupathi. Identifying corporate sustainability issues by analyzing shareholder resolutions: A machine-learning text analytics approach. *Sustainability*, 12(11):4753, 2020.
- [53] Régis MARODON, Jean-Baptiste Jacouton, and Adeline LAULANIE. The Proof is in the Pudding. Revealing the SDGs with Artificial Intelligence. Working Paper 85f81dba-c8e2-4255-878a-0, Agence française de développement, October 2022.
- [54] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.
- [55] SDG Prospector - artificial intelligence serving the sdgs. <https://www.sdgprospector.org>. [Accessed 22-09-2023].
- [56] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pre-trained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.
- [57] Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti-Senni, Dominik Stambach, Nicolas Webersinke, et al. Chatclimate: Grounding conversational ai in climate science. 2023.
- [58] ChatClimate grounded on the latest ipcc report. <https://www.chatclimate.ai>. [Accessed 22-09-2023].
- [59] Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stambach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, et al. Paradigm shift

- in sustainability disclosure analysis: Empowering stakeholders with chatreport, a language model-based tool. *arXiv preprint arXiv:2306.15518*, 2023.
- [60] TCFD. Recommendations of the task force on climate-related financial disclosures, 2017.
 - [61] Timothy M Doyle. Ratings that don’t rate: the subjective world of esg ratings agencies. *American Council for Capital Formation*, pages 65–71, 2018.
 - [62] OECD. *OECD Business and Finance Outlook 2020: Sustainable and Resilient Finance*. Number 6th (2020) in OECD Business and Finance Outlook. OECD, Paris, 2020.
 - [63] Elmira Aliakbari and Steven Globerman. The impracticality of standardizing esg reporting (esg: Myths and realities). 2023.
 - [64] SSAB. Sasb reporters.
 - [65] IR Solutions. Responsibilityreports.
 - [66] Hugging Face. Statistics on the number of monolingual models by language hosted on the hugging face platform.
 - [67] Refinitiv. Environmental, social and governance (esg) scores from refinitiv - may 2022.
 - [68] Özge Sahin, Karoline Bax, Claudia Czado, and Sandra Paterlini. Environmental, social, governance scores and the missing pillar—why does missing information matter? *Corporate Social Responsibility and Environmental Management*, 29(5):1782–1798, 2022.
 - [69] Eco-Business. Sustainability reporting: 4 things companies get wrong.
 - [70] Artifex. PyMuPDF — pypi.org. <https://pypi.org/project/PyMuPDF/>. [Accessed 22-09-2023].
 - [71] Nipun Sadvilkar and Mark Neumann. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online, 11 2020. Association for Computational Linguistics.
 - [72] Nipun Sadvilkar. PySBD — pypi.org. pypi.org/project/pysbd/. [Accessed 22-09-2023].
 - [73] W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading, 2010.
 - [74] Hannah Bast, Björn Buchhold, and Elmar Haussmann. Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval*, 10(2-3):119–271, 2016.
 - [75] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
 - [76] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search, 2022.
 - [77] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42, 2022.
 - [78] Stefan Buttcher, Charles LA Clarke, and Gordon V Cormack. *Information retrieval: Implementing and evaluating search engines*. Mit Press, 2016.
 - [79] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707, 2016.
 - [80] Wei Yang, Haotian Zhang, and Jimmy Lin. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*, 2019.
 - [81] NLP Group of The University of Hong Kong. Instructor-xl · Hugging Face — huggingface.co. <https://huggingface.co/hkunlp/instructor-xl>. [Accessed 25-09-2023].
 - [82] Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*, 2022.

- [83] Allen Institute for AI. Open information extraction - demo.
- [84] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022.
- [85] Eugene Yurtsev. Kor— pypi.org. <https://pypi.org/project/kor>. [Accessed 25-09-2023].
- [86] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [87] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.
- [88] Tom Jobbins. TheBloke/wizardLM-7B-HF · Hugging Face. <https://huggingface.co/TheBloke/wizardLM-7B-HF>. [Accessed 25-09-2023].
- [89] Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. *Natural language processing with transformers*. "O'Reilly Media, Inc.", 2022.
- [90] John Hewitt, Christopher D Manning, and Percy Liang. Truncation sampling as language model desmoothing. *arXiv preprint arXiv:2210.15191*, 2022.
- [91] Ernesto Estrada. *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, 10 2011.
- [92] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., USA, 2010.
- [93] Ulrik Brandes. *Network analysis: methodological foundations*, volume 3418. Springer Science & Business Media, 2005.
- [94] Junlong Zhang and Yu Luo. Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017)*, pages 300–303. Atlantis press, 2017.
- [95] Katherine Faust. Centrality in affiliation networks. *Social networks*, 19(2):157–191, 1997.
- [96] Marc Barthélemy. Betweenness centrality in large complex networks. *The European physical journal B*, 38(2):163–168, 2004.
- [97] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- [98] Luciano da F Costa. Further generalizations of the jaccard index. *arXiv preprint arXiv:2110.09619*, 2021.
- [99] Giulio Cimini, Tiziano Squartini, Fabio Saracco, Diego Garlaschelli, Andrea Gabrielli, and Guido Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1(1):58–71, 2019.
- [100] Gotelli, Nicholas J, and Werner Ulrich. Statistical challenges in null model analysis. *Oikos*, 121(2):171–180, 2012.
- [101] R Lyman Ott and Micheal T Longnecker. *An introduction to statistical methods and data analysis*. Cengage Learning, 2015.
- [102] Hervé Abdi. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA, pages 508–510, 2007.
- [103] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.
- [104] Correlation (pearson, spearman, and kendall).
- [105] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [106] Mark Lunt. Introduction to statistical modelling 2: categorical variables and interactions in linear regression. *Rheumatology*, 54(7):1141–1144, 05 2013.

- [107] Clara Dismuke and Richard Lindrooth. Ordinary least squares. *Methods and designs for outcomes research*, 93(1):93–104, 2006.
- [108] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [109] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction*, pages 109–139. Springer, 2022.
- [110] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.
- [111] Daniel Berrar. Cross-validation, 01 2018.
- [112] Dabao Zhang. A coefficient of determination for generalized linear models. *The American Statistician*, 71(4):310–316, 2017.
- [113] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.
- [114] Sungil Kim and Heeyoung Kim. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679, 2016.
- [115] Lloyd S Nelson. The anderson-darling test for normality. *Journal of Quality Technology*, 30(3):298, 1998.
- [116] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.
- [117] European Union. Eu emissions trading system (eu ets).
- [118] European Union. Fit for 55.
- [119] Sustainalytics. Esg risk ratings – methodology abstract, version 2.1.

Supplementary Materials for Glitter or Gold? Deriving Structured Insights from Sustainability Reports via Large Language Models

Marco Bronzini*, Carlo Nicolini, Bruno Lepri, Andrea Passerini and Jacopo Staiano

*Corresponding author. Email: marco.bronzini-1@unitn.it

1 Semantic clustering

Cluster label	Cluster elements
ESG Commitments	ESG Metrics ESG Incentives ESG Promotion ...
Product Sustainability	Sustainability Sustainable Products Supplier Sustainability ...
Supplier Assessment	Supplier Audits Supplier Relationship Supplier Risk Assessment ...

Table 1: Three examples of the semantic clusters of the ESG categories (cat) grouped through thresholded cosine similarity on embedded categories.

Cluster label	Cluster elements
Assessment of	Monitoring of Assessment on Assessment and monitoring of ...
Partnership with	Collaboration of Working together with Partnering with others to ...
Opportunity to	Creation of opportunity Opportunities to work on Opportunity to contribute to ...

Table 2: Three examples of the semantic clusters of predicates (pred) grouped through thresholded cosine similarity on embedded predicates.

Attribute	Original	Modified text
cat pred obt	Health and Safety Designed to continually improve Workplace safety culture	Employee Safety Improving Workplace safety culture
cat pred obt	Taxes Increase in Tax and royalty payment	Tax Increase of Tax and royalty payment
cat pred obt	Greenhouse Gas Emissions Follows The WRI and WBCSD GHG Protocol	Air Emissions Following The WRI and WBCSD GHG Protocol
cat pred obt	Resource Efficiency Increasing usage of Renewable and recycled raw materials	Energy Efficiency Increase of Renewable and recycled raw materials

Table 3: Three examples of replacing the attributes of the triples with the respective cluster labels.

2 Full model instruction

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.

```
``TypeScript
esg.actions: Array<{ //actions related to corporate's environmental, social or governance aspects
esg_category: "access.to.basic.services" | "access.to.healthcare" | "animal.welfare" | "anti.competitive.practices" |
"audit" | "biodiversity" | "board" | "board.diversity" | "business.ethics" | "chairperson.ceo.separation" | "child.labor" |
"climate.risk.management" | "clinical.trials" | "collective.bargaining" | "community.and.society" | "corporate.governance" |
"corruption" | "customer.relationship" | "diversity" | "esg.incentives" | "electromagnetic.fields" | "employee.development"
| "employee.turnover" | "energy" | "environmental.fines" | "environmental.management.system" | "environmental.policy"
| "environmental.reporting" | "financial.inclusion" | "forests" | "ghg.emissions" | "ghg.policies" | "gmoss" |
"global.compact.membership" | "green.buildings" | "green.products" | "hiv.programs" | "hazardous.waste" | "health.and.safety"
| "human.rights" | "indigenous.rights" | "labor.practices" | "lobbying" | "non.ghg.air.emissions" | "ozone.depleting.gases" |
"packaging" | "philanthropy" | "privacy.and.it" | "product.safety" | "public.health" | "remuneration" | "reporting.quality" |
"resource.efficiency" | "responsible.marketing" | "shareholders" | "site.closure" | "supply.chain" | "sustainable.finance" |
"systemic.risk" | "taxes" | "toxic.spills" | "unions" | "waste" | "water"
// an issue related to an ESG aspect
predicate: string //a nominalized verb that affects the ESG-related category
object: string //an entity related to the esg category that undergoes the predicate}>
``
```

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.

Input: In accordance with our ambitious goal, the water withdrawal of the data center decreased remarkably from 3.874 million litres to 2.367 million litres across the past three years.

Output: <json>{"esg.actions": [{"esg_category": "Water", "predicate": "Reduction of", "object": "The water withdrawal of the data center by 1.507 million litres"}]}</json>

Input: TotalEnergies introduced an innovative program at its European offices last year to address employees' concerns by creating a dedicated listening space.

Output: <json>{"esg.actions": [{"esg_category": "Employee Development", "predicate": "Introduction of", "object": "An innovative program"}]}</json>

Input: In San Antonio, Texas, our company reduced significantly the potable water usage of the data center by around 20% throughout 2020, providing economic and environmental benefits.

Output: <json>{"esg.actions": [{"esg_category": "Water", "predicate": "Reduction of", "object": "The data center's potable water usage by around 20%"}]}</json>

Input: In 2019, the ethics training program was completed by over 95% of our employees with outstanding results at our American training centre.

Output: <json>{"esg.actions": [{"esg_category": "Employee Development", "predicate": "Completion of", "object": "The ethics training program"}]}</json>

Input: Microsoft has invested €125 million in cutting-edge recycling technologies and smart waste management systems at its offices in Zwijndrecht, Belgium.

Output: <json>{"esg.actions": [{"esg_category": "Waste", "predicate": "Investment in", "object": "Cutting-edge recycling technologies and smart waste management systems"}]}</json>

Input: \$INPUT

Output:

Figure 1: The full model instruction used in our approach. INPUT represents an ESG-related sentence from which structured data needs to be retrieved. Sentences in the examples (in-context learning) do not represent any real facts.

3 Descriptive analysis of the generated triples

ESG category (cat)	Triples (%)	Companies (%)	Sectors (%)	Triples per company sector (%)
Corporate Governance	6.7	94	100	Industrials (7.8%), Financial Services (7.5%), Healthcare (7.4%)
Climate Risk Management	5.1	95	100	Real Estate (7.9%), Financial Services (7.0%), Industrials (5.5%)
Employee Development	4.8	90	100	Healthcare (5.9%), Communication Services (5.7%), Real Estate (5.6%)
Air Emissions	3.7	92	100	Energy (6.9%), Basic Materials (4%), Industrials (3.8%)
Water	3.1	85	100	Consumer Defensive (6.6%), Basic Materials (4.9%), Energy (4.4%)
Green Buildings	1.5	88	100	Real Estate (3.5%), Consumer Cyclical (1.8%), Technology (1.6%)
Packaging	0.9	49	100	Consumer Defensive (4.9%), Consumer Cyclical (1.5%), Technology (0.9%)
Business Ethics	0.3	61	100	Healthcare (0.7%), Utilities (0.5%), Industrials (0.4%),

Table 4: Sample of the ESG categories disclosed by companies in their sustainability reports. The table presents the category coverage using three key metrics: the total number of triples (triples), the count of companies that reported it (companies), and the number of company sectors covered (sectors). Additionally, the coverage is assessed by aggregating all the company triples by company sector and the percentage of sector triples concerning a category is reported (Triples per sector).

Action (cat:pred)	Triples (%)	Companies (%)	Sectors (%)	Triples per company sector (%)
AIR EMISSIONS: reduction of	0.6	70	100	Energy (1.2%), Basic Materials (0.8%), Industrials (0.7%)
PHILANTHROPY: donation by	0.4	60	100	Real Estate (0.8%), Communication Services (0.5%), Healthcare (0.5%)
ENERGY: reduction of	0.3	61	100	Real Estate (0.5%), Consumer Cyclical (0.4%), Communication Services (0.4%)
CLIMATE RISK MANAGEMENT: assessment of	0.3	56	100	Real Estate (0.7%), Industrials (0.3%), Financial Services (0.3%)
WATER: reduction of	0.3	50	100	Consumer Defensive (0.4%), Consumer Cyclical (0.4%), Energy (0.4%)
CORPORATE GOVERNANCE: establishment of	0.3	49	100	Healthcare (0.4%), Financial Services (0.4%), Technology (0.4%)
BIODIVERSITY: promotion of	0.2	40	91	Energy (0.5%), Utilities (0.4%), Financial Services (0.4%)
WASTE: reduction of	0.2	39	100	Consumer Defensive (0.5%), Industrials (0.3%), Consumer Cyclical (0.2%)
COMMUNITY AND SOCIETY: engagement in	0.1	37	82	Real Estate (0.5%), Healthcare (0.4%), Basic Materials (0.2%)
EMPLOYEE SAFETY: commitment and involvement with	0.1	25	100	Basic Materials (0.3%), Healthcare (0.1%), Industrials (0.1%)

Table 5: Sample of the actions disclosed by companies in their sustainability reports. The table presents the action coverage through the same four metrics of the above table.

4 Statistics of the bipartite graphs

Bipartite graph	Partition A	Partition B	Edges	Density (%)
category-predicate ($\mathbb{B}_{\text{catpred}}$)	542	4,864	19,574	0.7
company-category ($\mathbb{B}_{\text{cocat}}$)	124	542	7,455	10.8
company-action ($\mathbb{B}_{\text{coact}}$)	124	19,574	43,169	1.7

Table 6: Statistics of the bipartite graphs obtained from the original KG. The category-predicate and company-action graphs have a low density ($<1\%$), while the company-category graph exhibits a higher density (11%).

5 Degree distributions of the three bipartite graphs

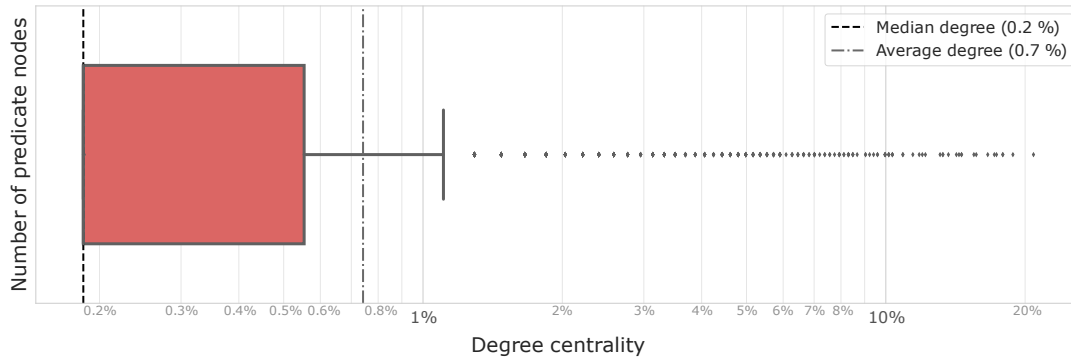


Figure 2: Numerical distribution of degree centrality concerning the bipartite graph $\mathbb{B}_{\text{catpred}}$.

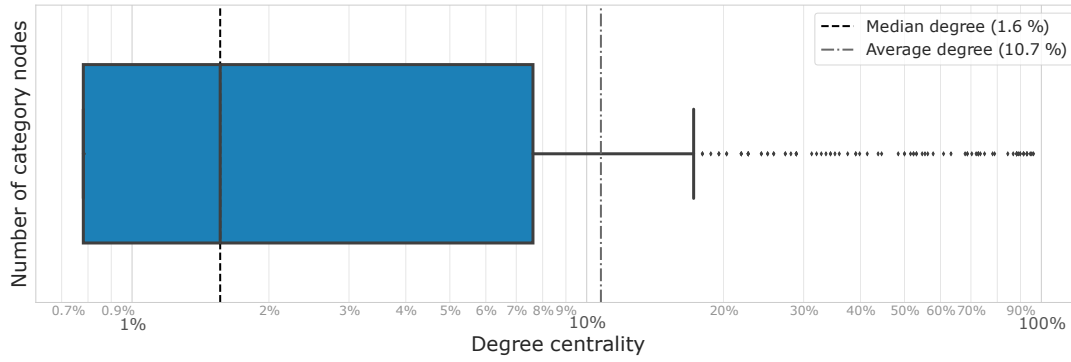


Figure 3: Numerical distribution of degree centrality concerning the bipartite graph $\mathbb{B}_{\text{cocat}}$.

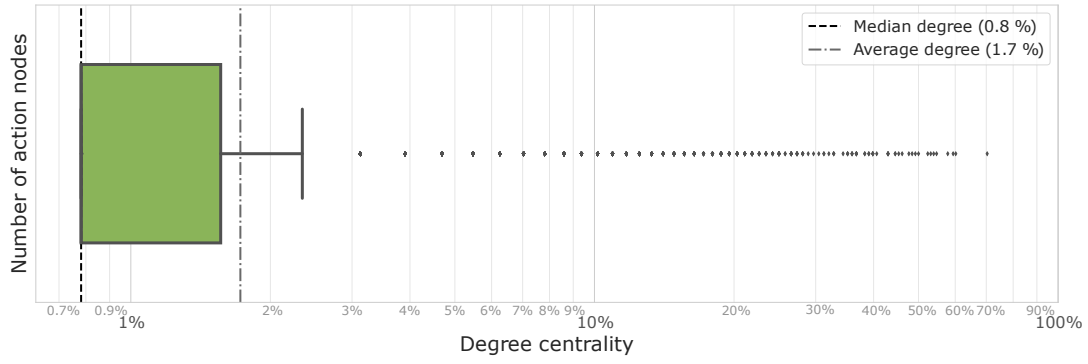


Figure 4: Numerical distribution of degree centrality concerning the bipartite graph B_{coact} .

6 Descriptive statistics of the three bipartite graphs

Predicate node (pred)	Degree (%)	Closeness (%)	Betweenness (%)
Commitment and involvement with	20.8	89.9	1.5
Advisory support for	18.8	89.0	1.2
Partnership with	17.9	88.2	1.2
Establishment of	17.3	88.3	1.1
Use of	17.2	88.9	1.1
Development and implementation of	16.6	89.0	1.0
Recognition of	13.1	87.6	0.8
Engagement in	12.2	84.2	0.8
Compliance with	10.9	85.3	0.6
Consideration of	10.3	83.9	0.6
Reduction of	10.1	77.3	0.5
Review of	10.0	84.6	0.5
Maintenance of	10.0	84.2	0.5
Launch of	10.0	82.3	0.5
The impacts of	9.6	84.2	0.6
Belief in	9.6	83.2	0.4
Accounting for	1.8	61.2	0.0
Becoming aware of	1.8	60.2	0.0
Re-institution of	0.4	51.0	0.0
Helping consumers achieve	0.2	49.9	0.0
Questioned by	0.2	34.8	0.0
Contesting	0.2	32.2	0.0

Table 7: Graph metrics of a sample of all predicate nodes of the bipartite graph $B_{catpred}$.

Category node (cat)	Degree (%)	Closeness (%)	Betweenness (%)
Environmental	96.1	99.0	2.3
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
Energy	93.0	97.9	2.1
Product Sustainability	93.0	97.7	2.2
Community and Society	91.4	98.1	2.1
Board Diversity	89.8	97.4	2.0
Philanthropy	88.3	96.8	2.0
Business Ethics	60.9	85.2	1.0
Anti-Competitive Practices	52.3	79.9	0.7
Packaging	50.0	77.7	0.6
Human Resources	39.8	74.2	0.4
Tax	39.1	74.6	0.4
Lobbying	32.8	71.8	0.3
Whistle-blowing	20.3	64.5	0.1
Recycling	8.6	59.1	0.0
LGBTQ+ Inclusion	4.7	55.4	0.0
Product Responsibility	2.3	53.2	0.0
Anti-Discrimination	2.3	52.1	0.0
Green-washing Behavior	1.6	51.2	0.0
Marketing Responsibly	0.8	47.8	0.0

Table 8: Graph metrics of a sample of all the 542 category nodes of the bipartite graph \mathbb{B}_{ccat} .

Action node (cat : pred)	Degree (%)	Closeness (%)	Betweenness (%)
AIR EMISSIONS: Reduction of	70.3	87.6	2.1
ENERGY: Reduction of	60.2	83.9	1.5
PHILANTHROPY: Advisory support for	59.4	81.9	1.4
PHILANTHROPY: Donation by	57.8	80.8	1.3
COMMUNITY AND SOCIETY: Advisory support for	54.7	78.2	1.1
CLIMATE RISK MANAGEMENT: Assessment of	53.9	77.2	1.0
BIODIVERSITY: Commitment and involvement with	53.1	77.1	1.1
CORPORATE GOVERNANCE: Commitment and involvement with	52.3	77.1	1.1
WATER: Reduction of	50.0	76.8	1.0
HUMAN RIGHTS: Assessment of	21.9	63.8	0.2
ENVIRONMENTAL: Participation in	21.9	62.4	0.2
EMPLOYEE DEVELOPMENT: Offering	21.9	61.5	0.2
ENVIRONMENTAL: Certification for	21.1	60.7	0.2
HUMAN RIGHTS: Recognition of	20.3	62.6	0.2
WATER: Introduction of	20.3	61.0	0.2
EMPLOYEE SAFETY: Advisory support for	20.3	60.5	0.2
WASTE: Evaluation of	4.7	52.8	0.0
SUPPLIER DIVERSITY: Establishment of	4.7	52.8	0.0
PRODUCT SUSTAINABILITY: Incorporation of	4.7	52.7	0.0
TRANSPORTATION: Opportunity to	0.8	50.4	0.0
INSURANCE: Related to	0.8	39.9	0.0
SUPPLIER ASSESSMENT: Internal risk based	0.8	38.6	0.0

Table 9: Graph metrics of a sample of all predicate nodes of the bipartite graph $\mathbb{B}_{\text{coact}}$.

7 Variability of ESG-related actions

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%)
Biodiversity	5.70	92	Commitment and involvement with (4.7%), Promotion of (4.2%), Increase of (2.1%)
Environmental	5.65	97	Development and implementation of (4.6%), Commitment and involvement with (3.1%), Establishment of (2.1%)
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%)
Employee Development	5.46	90	Introduction of (4.5%), Provision of (3.9%), Required training (3.3%)
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%)
Product Responsibility	1.79	2	Signatory to (16.7%), Commitment and involvement with (16.7%), Includes (16.7%)
Anti-Slavery Practices	1.61	4	Undertaking of (20%), Integration of (20%), Required training (20%)
Clean Energy	0.69	2	Helping clients deploy (50%), Engagement in (50%)
Conflict of Interest	0.00	1	Avoidance of (100%)

Table 10: A sample of ESG categories with the computed entropy values. The three most frequent category predicates are reported alongside the percentage of companies disclosing that category.

8 Company similarities according to disclosed actions

Company	Top three 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%), LG Display (6%), 3M (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%), Airbus (5%), BPER Banca (5%)
Mastercard	PHILANTHROPY: Advisory support for (x4) HUMAN RIGHTS: Commitment and involvement with (x4) FINANCIAL INCLUSION: Commitment and involvement with (x4)	Visa (6%), Home Depot (5%), American Electric Power Company (5%), BPER Banca (5%), Vodafone (4%)
Moderna	BIODIVERSITY: Commitment and involvement with (x4) PHILANTHROPY: Launch of (x3) PHILANTHROPY: Participation in (x3)	Vertex Pharmaceuticals (7%), AstraZeneca (5%), Broadcom (5%), Franklin Electric (4%), Alcon (4%)
Delta Air Lines	CLIMATE RISK MANAGEMENT: Evaluation of (x4) COMMUNITY AND SOCIETY: Recognition of (x3) EMPLOYEE SAFETY: Introduction of (x3)	American Electric Power Company (6%), Broadcom (6%), PPG Industries (5%), Canadian Pacific Railway (5%)
GlobalFoundries	WATER: Use of (x6) AIR EMISSIONS: Reduction of (x7) PHILANTHROPY: Donation by (x7)	Texas Instruments (7%), PPG Industries (7%), Imperial Oil (6%), Samsung (6%), Visa (6%)
Aluminum Corporation of China	CORPORATE GOVERNANCE: Establishment of (x3) CORPORATE GOVERNANCE: Improving (x3) CORPORATE GOVERNANCE: Organisation around (x3)	China Petroleum Chemical (6%), Geely Automobile (5%), Baidu (5%), United States Steel (4%), Tokyo Gas (4%)
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%), Korean Air Lines (6%), Alibaba Group (5%)
STMicroelectronics	AIR EMISSIONS: Reduction of (x4) SUPPLY CHAIN: Assessment of (x4) AUDIT: Conducted (x3)	Texas Instruments (6%), PPG Industries (5%), GlobalFoundries (5%), Monster (5%), TotalEnergies (5%)
Enel	AIR EMISSIONS: Reduction of (x10) COMMUNITY AND SOCIETY: Contribution to (x6) HUMAN RIGHTS: Commitment and involvement with (x6)	Saipem (7%), Snam (6%), Tokyo Gas (6%), TotalEnergies (6%), Banco Santander (6%), UniCredit SpA (5%)
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%), GlobalFoundries (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)	3M (4%), Coca Cola (4%), Croda (4%), GlobalFoundries (4%), United States Steel (4%), Samsung (4%)
VMware	CORPORATE GOVERNANCE: Overseeing (x4) PHILANTHROPY: Advisory support for (x4) BIODIVERSITY: Commitment and involvement with (x3)	Visa (5%), Oracle (4%), Mastercard (4%), The Home Depot (4%), Amazoncom (4%), Cisco (4%)

Table 11: A company sample with the most reported actions and the most similar companies for each. The company similarity is assessed by computing the Jaccard similarity on the action set of companies.

9 Features used for the correlation analysis of company similarities

Feature	Similarity Measure
Action Similarity	Jaccard Similarity
Sector	Cosine Similarity with min-max normalisation
Industry	
Country of Headquarters	
Minor Region of Headquarters	
ESG Score	Absolute difference with max-min normalisation
Environmental Pillar Score	
Social Pillar Score	
Governance Pillar Score	
ESG Financial Period	
Full-Time Employees	
Normalized EBITDA (USD billions)	
Total Liabilities (USD billions)	
Company Market Cap (USD billions)	

Table 12: Similarity measures of company features used in the bivariate correlation analysis concerning company similarities. All the measures are defined in the range between zero and one.

10 Statistical significance of the pairwise correlations

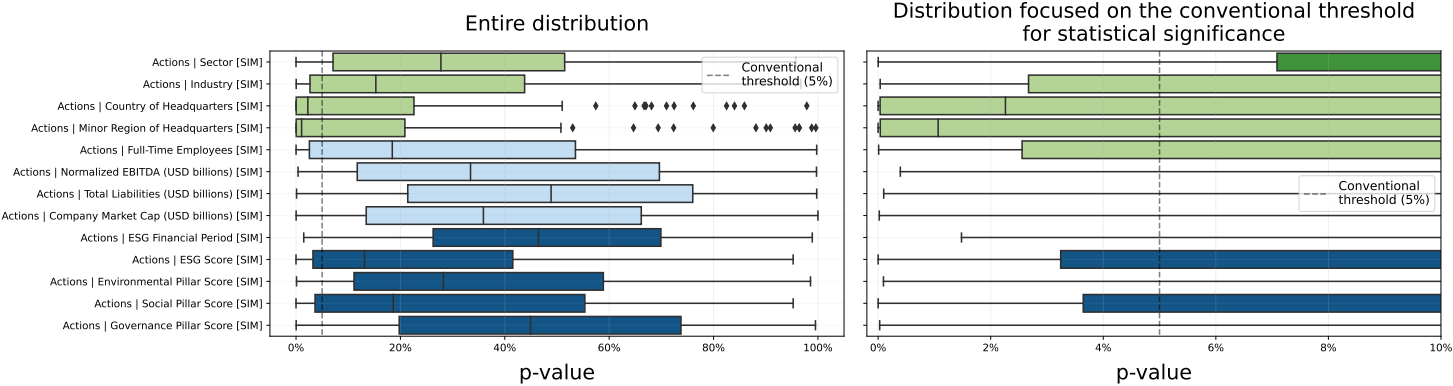


Figure 5: Distribution of the p-values for Kendall's correlation coefficient computed in the correlation analysis between company similarities in terms of actions disclosed and similarities in other company information. The right graph exhibits the distributions focused around the conventional threshold of statistical significance.

11 Monotonic correlations among company information

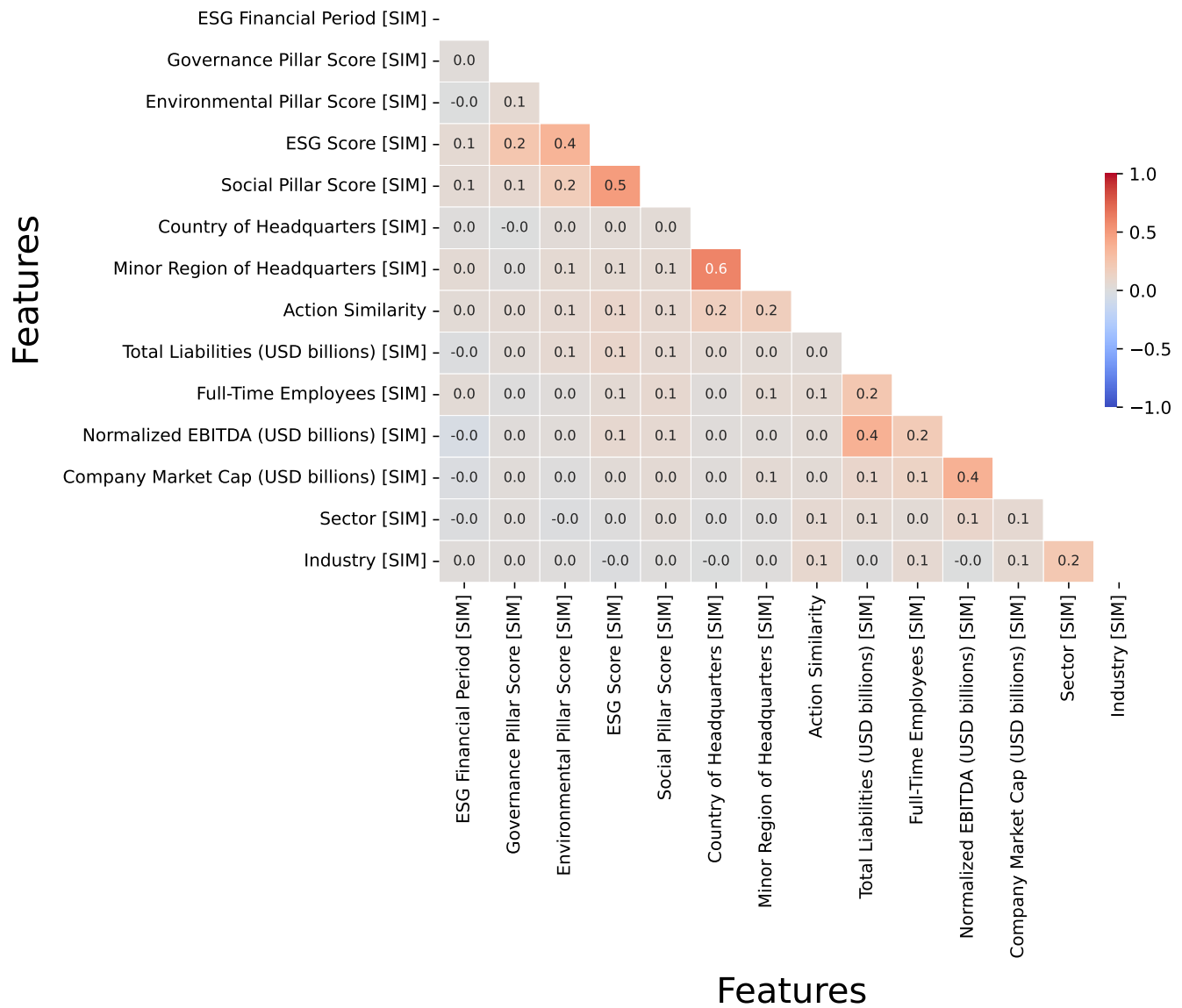


Figure 6: Pairwise monotonic correlations between company similarities.

12 Features for the ESG score inference for the interpretability analysis

Feature	Value	Source	Type
Category Entropy	3.6	Sustainability report	Continuous
Action Entropy	1.1		
EBITDA (USD Billions)	16.0	Refinitiv	
Liabilities (USD Billions)	191.8		
Market Cap (USD Billions)	158.6		
Employees	108900	Discrete	
Incorporation Year	1946		
Fiscal Year	2022		
SECTOR: Technology	True	Refinitiv	Binary categorical
SECTOR: Utilities	False		
...	...		
CONTINENT: Asia	True		
CONTINENT: Europe	False		
...	...		
REGION: Eastern Asia	True		
REGION: Northern Europe	False		
...	...		
COUNTRY: Japan	True		
COUNTRY: Canada	False		
...	...		
Social Categories	0.27	Sustainability report	Percentage
Environmental Categories	0.30		
Governance Categories	0.42		
CATEGORY: Corporate Governance	0.09		
CATEGORY: Environmental	0.07		
CATEGORY: Supply Chain	0.06		
CATEGORY: Waste	0.03		
CATEGORY: Board Diversity	0		
...	...		

Table 13: Example of the features used in the regression model. There are numerical features and categorical features (dummy variables). The data is exhibited beforehand standardization. The observation concerns the fiscal year 2022 of Sony.

13 Performance of the OLS model

Metric	Score
Coefficient of Determination (R^2)	0.707
Mean Absolute Error (MAE)	5.775
Root Mean Square Error (RMSE)	7.761
Weighted Mean Absolute Percentage Error (wMAPE)	0.079

Table 14: Performance of the first-order regression model. Elastic Net Regularization was used as a feature selection method with the optimal alpha parametric constant ($\alpha = 0.493$) discovered through a five-fold Cross-Validation approach.

14 Training performance of the OLS model using cross-validation

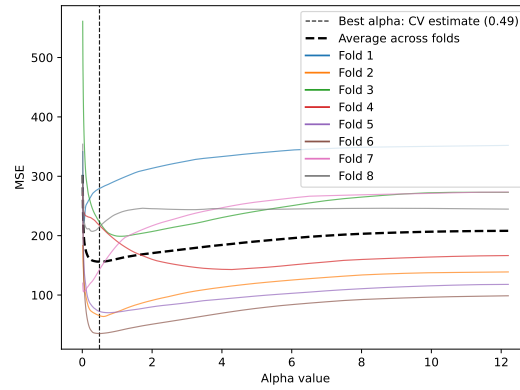


Figure 7: Performance of the OLS regression model during training. An eight-fold cross-validation approach, with the Elastic Net cost function based on Mean Squared Error (MSE), was adopted to estimate the optimal α parameter for the Elastic Net Regularisation.

15 Residual analysis of the OLS model

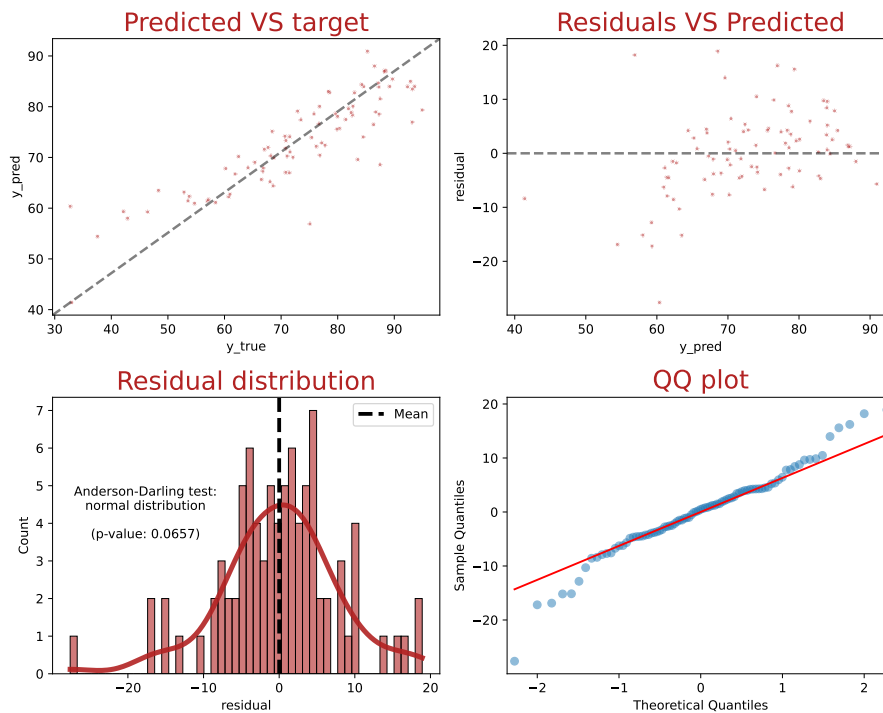


Figure 8: Graphical panels exhibiting different residual analyses. The plot at the top left exhibits the predicted scores versus the actual ones, while the top-right graph shows the residuals versus predicted scores in order to check homoscedasticity. At the bottom left, the histogram exhibits the distributions of the residuals alongside the results of the Anderson-Darling test on normal distribution. Lastly, a QQ plot of residuals versus normal distribution is displayed at the bottom right of the graphical panel.

16 ESG scores by region

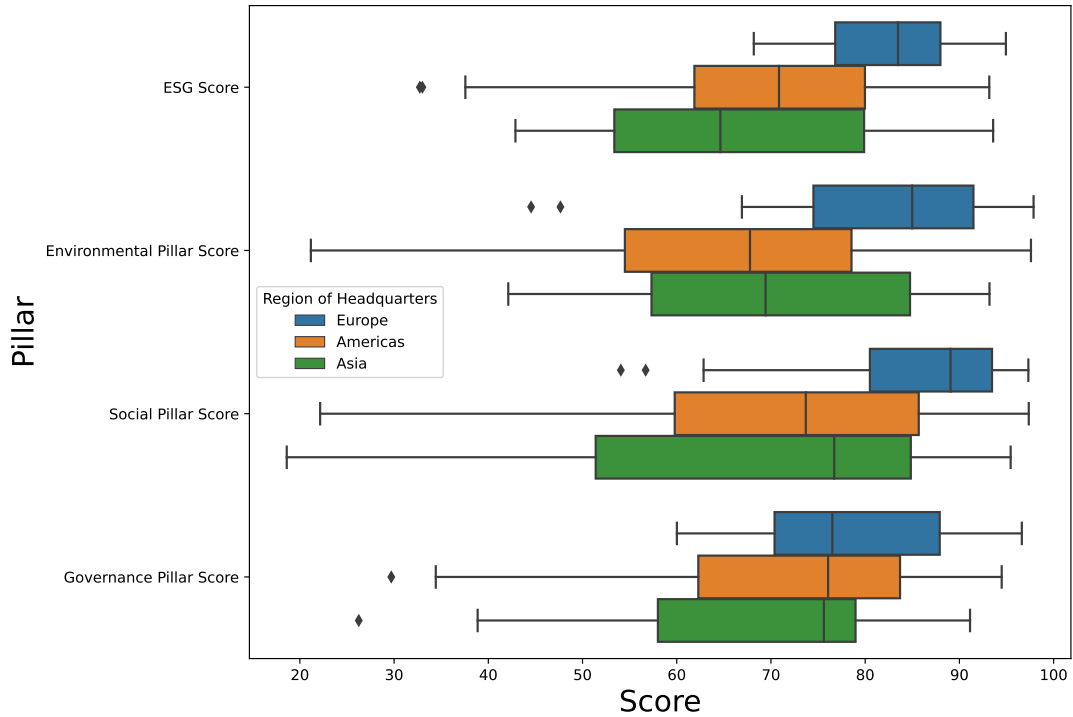


Figure 9: ESG scores of companies aggregated by continent. The graphical panel exhibits the E/S/G scores as well as the combined score. European companies have the highest average scores in each pillar.

17 Asian companies considered for the region-based ESG score interpretability

Company	Sector	Country	Incorporation Year	Fiscal Year	...	ESG score
Toshiba	Industrials	Japan	1904	2021	...	93.58
Sony	Technology		1946	2022	...	87.51
Toyota Motor	Consumer Cyclical		1937	2020	...	84.51
Tokyo Gas	Utilities		1885	2021	...	78.35
Geely Automobile	Consumer Cyclical	Hong Kong	1946	2022	...	75.42
Daikin Industries	Industrials	Japan	1934	2022	...	66.78
China Petroleum Chemical Corporation	Energy	China	2000	2021	...	62.49
Aluminum Corporation of China	Basic Materials		2001	2019	...	57.14
Baidu	Communication Services		2000	2020	...	53.55
China Evergrande	Real Estate		2006	2020	...	52.85
Alibaba	Consumer Cyclical	Saudi Arabia	1999	2022	...	48.33
Saudi Aramco	Energy		2018	2021	...	42.88

Table 15: The twelve Asian companies presented in our data. This subset includes businesses from Japan, China, Hong Kong and Saudi Arabia and was used for the region-based interpretability of ESG scores. The companies are ordered by ESG score.

18 Top features impacting the Asian companies' ESG scores

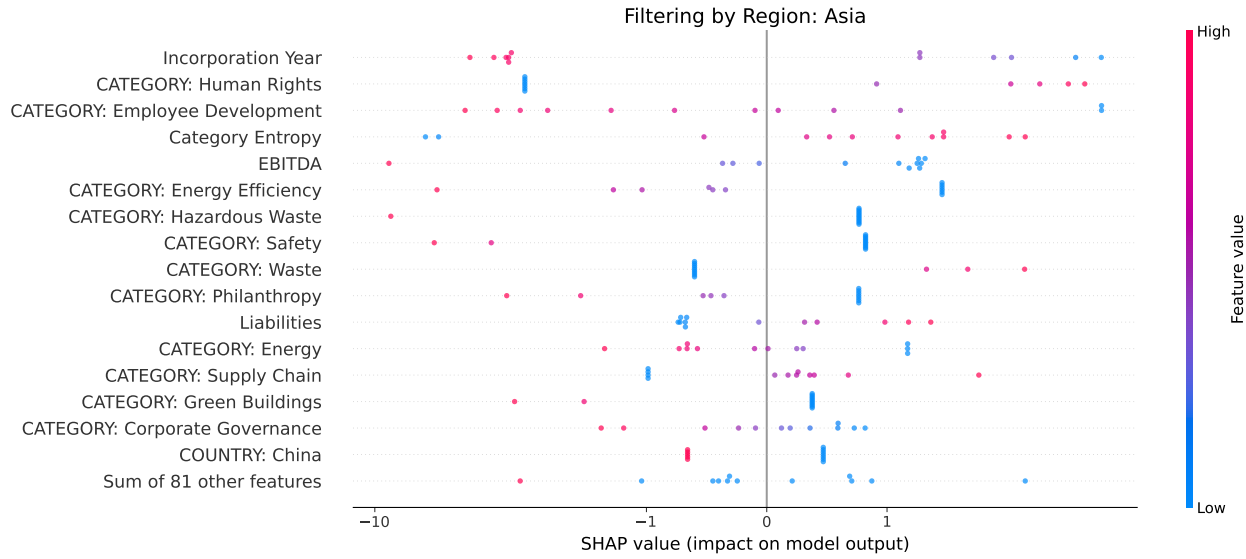


Figure 10: Summary of the sixteen primary factors with the greatest influence on determining the ESG score for Asian companies. The features are ordered according to their median shape value. The x-axis represents the degree of a positive and negative impact on model output. Each dot represents a company instance and colours represent the company values of the standardised feature.

19 Categories of the adopted ESG Taxonomy

ESG category	Pillar	Degree centrality (%)
Climate Risk Management	Environmental	94.5
Energy		93.0
Biodiversity		91.4
Waste		88.3
Green Buildings		86.7
Water		84.4
Product Safety		68.8
Forests		53.1
Packaging		50.0
Hazardous Waste		48.4
Animal Welfare		35.9
Green Products		35.2
Environmental Fines		10.8
Environmental Management System		10.8
Environmental Policy		10.8
Environmental Reporting		10.8
GHG Emissions		10.8
GHG Policies		10.8
Non-GHG Air Emissions		10.8
Ozone-Depleting Gases		10.8
Resource Efficiency		10.8
Sustainable Finance		10.8
Electromagnetic Fields		7.0
GMOs		2.3
Toxic Spills		2.3
Supply Chain	Social	95.3
Community and Society		91.4
Employee Development		89.1
Human Rights		88.3
Philanthropy		88.3
Customer Relationship		73.4
Financial Inclusion		70.3

Access to Basic Services		68.8
Access to Healthcare		44.5
Child Labor		33.6
Collective Bargaining		28.9
Employee Turnover		27.3
Public Health		19.5
Diversity		10.8
Health and Safety		10.8
Labor Practices		10.8
Responsible Marketing		10.8
Unions		10.8
Indigenous Rights		8.6
Clinical Trials		7.8
HIV Programs		0.8
Corporate Governance		93.0
Board Diversity		89.8
Audit		72.7
Remuneration		72.7
Business Ethics		60.9
Anti-competitive Practices		52.3
Lobbying		32.8
Systemic Risk		28.1
Global Compact Membership		21.8
Board	Governance	10.8
Chairperson-CEO Separation		10.8
Corruption		10.8
ESG Incentives		10.8
Privacy and IT		10.8
Reporting Quality		10.8
Shareholders		10.8
Taxes		10.8
Site Closure		10.2

Table 16: The table exhibits the ESG categories of the ESG taxonomy used in this work. Their degree is derived from our findings, although they represent only a tenth of the categories discovered through our methodology.

20 Distribution of the fiscal years of the sustainability reports considered

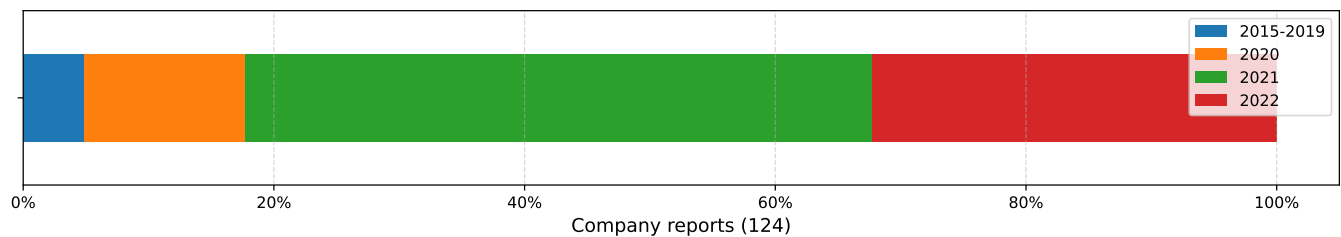


Figure 11: Distribution of the fiscal years of the sustainability reports considered for our study. We picked the latest available reports for each considered company. Almost all reports concern the 2020s (orange, green and red), whereas only 5% refer to fiscal years within the 2015-2019 period (blue).

21 Complete list of the company considered

Company	Sector	Industry
3M Corporation	Industrials	Conglomerates
3i Group plc	Financial Services	Asset Management
Activision Blizzard Inc	Communication Services	Electronic Gaming & Multimedia
Adecco Group AG	Industrials	Staffing & Employment Services
Adidas AG	Consumer Cyclical	Footwear & Accessories
Air Canada	Industrials	Airlines
Air Liquide SA	Basic Materials	Specialty Chemicals
Airbus SE	Industrials	Aerospace & Defense
Alcon Inc	Healthcare	Medical Instruments & Supplies
Alibaba Group Holding Limited	Consumer Cyclical	Internet Retail
Alphabet Inc	Communication Services	Internet Content & Information
Aluminum Corporation of China Limited	Basic Materials	Aluminum
Amazoncom Inc	Consumer Cyclical	Internet Retail
American Electric Power Company Inc	Utilities	Utilities: Regulated Electric
Amplifon	Healthcare	Medical Distribution
Apple Inc	Technology	Consumer Electronics
ArcelorMittal SA	Basic Materials	Steel
Assicurazioni Generali SpA	Financial Services	Insurance: Diversified
AstraZeneca PLC	Healthcare	Drug Manufacturers: General
BPER Banca SpA	Financial Services	Banks: Regional
Baidu Inc	Communication Services	Internet Content & Information
Banco Santander SA	Financial Services	Banks: Diversified
Bank of America Corp BofA	Financial Services	Banks: Diversified
Bayer AG	Healthcare	Drug Manufacturers: General
British American Tobacco PLC	Consumer Defensive	Tobacco
British Land Co PLC The	Real Estate	REIT: Diversified
Broadcom Inc	Technology	Semiconductors
Builders FirstSource Inc	Industrials	Building Products & Equipment
CF Industries Holdings Inc	Basic Materials	Agricultural Inputs
Campbell Soup Company	Consumer Defensive	Packaged Foods
Canadian Pacific Railway Limited	Industrials	Railroads
Canon Inc	Technology	Computer Hardware
CarMax Inc	Consumer Cyclical	Auto & Truck Dealerships
China Evergrande Group	Real Estate	Real Estate: Development
China Petroleum Chemical Corporation	Energy	Oil & Gas Integrated
Cisco Systems Inc	Technology	Communication Equipment
Coca Cola	Consumer Defensive	Beverages: Non-Alcoholic
Commonwealth Bank of Australia	Financial Services	Banks: Diversified
Croda International plc	Basic Materials	Specialty Chemicals
Daikin Industries Ltd	Industrials	Building Products & Equipment
Delta Air Lines Inc	Industrials	Airlines
Deutsche Bank AG	Financial Services	Banks: Regional
Deutsche Lufthansa AG	Industrials	Airlines
Deutsche Wohnen	Real Estate	Real Estate: Development
DuPont	Basic Materials	Specialty Chemicals
ENI SpA	Energy	Oil & Gas Integrated
Edison International	Utilities	Utilities: Regulated Electric
Enel SpA	Utilities	Utilities: Diversified
FedEx Corporation	Industrials	Integrated Freight & Logistics
First Republic Bank CA	Financial Services	Banks: Regional
Fox Corporation	Communication Services	Entertainment
Franklin Electric Co Inc	Industrials	Specialty Industrial Machinery
Geely Automobile Holdings Ltd	Consumer Cyclical	Auto Manufacturers
General Motors Co GM	Consumer Cyclical	Auto Manufacturers
GlobalFoundries	Technology	Semiconductors
Goldman Sachs Group Inc The	Financial Services	Capital Markets
Home Depot Inc The	Consumer Cyclical	Home Improvement Retail
Humana Inc	Healthcare	Healthcare Plans
Hyundai Motor Co	Consumer Cyclical	Auto Manufacturers
Imperial Oil Ltd	Energy	Oil & Gas Integrated
Intel Corp	Technology	Semiconductors
Intuitive Surgical Inc	Healthcare	Medical Instruments & Supplies
Iveco Group NV	Industrials	Farm & Heavy Construction Machinery
Johnson Johnson	Healthcare	Drug Manufacturers: General

Kia Corp	Consumer Cyclical	Auto Manufacturers
Korean Air Lines Co Ltd	Industrials	Airlines
Kraft Heinz Co The	Consumer Defensive	Packaged Foods
LG Display Co Ltd	Technology	Consumer Electronics
Leonardo SpA	Industrials	Aerospace & Defense
Lockheed Martin Corp	Industrials	Aerospace & Defense
Mastercard Inc	Financial Services	Credit Services
Meta Platforms Inc	Communication Services	Internet Content & Information
Microsoft Corporation	Technology	Software: Infrastructure
Moderna Inc	Healthcare	Biotechnology
Monster Beverage Corp	Consumer Defensive	Beverages: Non-Alcoholic
NVIDIA Corp	Technology	Semiconductors
National Grid PLC	Utilities	Utilities: Regulated Electric
Nestle SA	Consumer Defensive	Packaged Foods
Netflix Inc	Communication Services	Entertainment
Novo Nordisk A S	Healthcare	Biotechnology
Oracle Corporation	Technology	Software: Infrastructure
PPG Industries	Basic Materials	Specialty Chemicals
Paramount Resources Ltd	Energy	Oil & Gas E&P
Park Hotels Resorts Inc	Real Estate	REIT: Hotel & Motel
PepsiCo Inc	Consumer Defensive	Beverages: Non-Alcoholic
Petroleo Brasileiro SA Petrobras	Energy	Oil & Gas Integrated
Philip Morris International	Consumer Defensive	Tobacco
Poste Italiane	Industrials	Conglomerates
Prologis Inc	Real Estate	REIT: Industrial
Royal Bank of Canada	Financial Services	Banks: Diversified
Royal Dutch Shell PLC	Energy	Oil & Gas Integrated
STMicroelectronics	Technology	Semiconductors
Saipem SpA	Energy	Oil & Gas Equipment & Services
Samsung Electronics Co Ltd	Technology	Consumer Electronics
Saudi Aramco	Energy	Oil & Gas Integrated
Simon Property Group Inc	Real Estate	REIT: Retail
SkyWest Inc	Industrials	Airlines
Sligro Food Group NV	Consumer Defensive	Food Distribution
Snam SpA	Utilities	Utilities: Regulated Gas
Sony Corporation	Technology	Consumer Electronics
Sun Communities	Real Estate	REIT: Residential
Swisscom AG	Communication Services	Telecom Services
Telecom Italia SpA	Communication Services	Telecom Services
Tesco PLC	Consumer Defensive	Grocery Stores
Tesla Inc	Consumer Cyclical	Auto Manufacturers
Texas Instruments Inc	Technology	Semiconductors
Tokyo Gas Co Ltd	Utilities	Utilities: Regulated Gas
Toshiba Corp	Industrials	Conglomerates
TotalEnergies	Energy	Oil & Gas Integrated
Toyota Motor Corp	Consumer Cyclical	Auto Manufacturers
Uber Technologies Inc	Technology	Software: Application
UniCredit SpA	Financial Services	Banks: Regional
Uniper SE	Utilities	Utilities: Independent Power Producers
United States Steel Corp	Basic Materials	Steel
VMware Inc	Technology	Software: Infrastructure
Vertex Pharmaceuticals Inc	Healthcare	Biotechnology
Virgin Atlantic Ltd	Industrials	Transportation & Logistics
Visa Inc	Financial Services	Credit Services
Vodafone Group plc	Communication Services	Telecom Services
WESCO International Inc	Industrials	Industrial Distribution
Walmart Inc	Consumer Defensive	Discount Stores
Walt Disney Co	Communication Services	Entertainment
Washington Real Estate Investment Trust	Real Estate	Commercial Real Estate
Yamana Gold Inc	Basic Materials	Gold
Companies: 124		

22 All the sustainability reports originally gathered

Nationality	Companies	Companies (%)
United States	2059	47.0
Canada	391	8.9
Britain	191	4.4
Brazil	176	4.0
Taiwan	154	3.5
Japan	138	3.2
South Korea	133	3.0
...
United Arab Emirates	3	0.07
	4,222	100

Table 18: Nationality of the companies covered. Results are ordered by cardinality. The majority of the available reports come from North American companies (56%).

Language	Reports	Reports (%)
English (EN)	4,804	94.0
Spanish (ES)	77	1.5
Portuguese (PT)	76	1.5
Chinese (ZH)	61	1.2
Korean (KO)	31	0.6
Japanese (JA)	27	0.5
French (FR)	20	0.4
...
German (DE)	3	0.1
	6,456	100

Table 19: Percentage of the available reports by language. Results are ordered by cardinality. Almost all the non-financial reports are written in English.

Period	Fiscal year	Reports	Reports (%)
2020s	2021	3,483	54.0
	2020	1,588	24.6
	2022	621	9.6
2010s	2019	567	8.8
	2018	131	2.0
	2017	20	0.3
	2015	17	0.3

	2013	2	0.03
		6,456	100

Table 20: Percentage of the available sustainability reports by Fiscal year. The results are ordered by annual reports.