Department of Business and Law (Di.SEA.DE)

Department of Economics, Management and Statistics (DEMS)

PhD program **Business for Society**
Cycle **37°**

Innovating ESG and sustainable business practices through Generative AI

Surname: **Caggioni** Name: **Lorenzo**

Registration number 889694

Tutor: **Monica Rossolini**

Coordinator: **Elisabetta Marafioti**

Academic year 2023-2024

Every insight, every step,
has been illuminated
by a greater Presence,
to which I humbly
attribute everything.

SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Innovating ESG and sustainable business practices through Generative AI

Keywords: Large Language Models (LLMs), Generative AI (GenAI), Environmental Sustainability, Non-Financial Disclosures, ESG (Environmental, Social, and Governance), EU Taxonomy , Sustainability Reporting

Synopsis: This thesis explores the transformative potential of Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), in advancing corporate environmental sustainability. The increasing importance of ESG (Environmental, Social, and Governance) topics and the complexities of current reporting frameworks, characterized by heterogeneity and unstructured data, present significant challenges for stakeholders seeking consistent and comparable information. This research addresses how GenAI can overcome these hurdles, thereby enhancing transparency and accountability in sustainability reporting. The innovative empirical contribution of this thesis lies in its rigorous assessment of Generative AI's capability to evaluate corporate alignment with the EU Taxonomy. Specifically, it meticulously compares the accuracy and efficiency of two distinct LLM techniques—Retrieval Augmented Generation (RAG) and Long Context processing—in extracting pertinent information from unstructured non-financial disclosure documents. This direct comparison provides novel insights into the practical applicability and performance nuances of these cutting-edge AI methodologies within the critical domain of corporate sustainability reporting.

The methodology is structured in three distinct phases. First, a systematic literature review, leveraging GenAI itself, identifies the current landscape of GenAI applications in corporate environmental sustainability. This review classifies existing research into key thematic clusters, including applications for improving business operations, enhancing sustainability reporting and ESG analysis, and addressing the environmental impact of GenAI itself. Second, the thesis empirically assesses GenAI's capability in evaluating corporate alignment with the EU Taxonomy. This phase specifically compares the accuracy and efficiency of Retrieval Augmented Generation (RAG) and Long Context processing techniques in extracting relevant information from non-financial disclosure documents, demonstrating the superior performance of Long Context models in this task. Finally, a case study illustrates the practical application of

GenAI in supporting ESG researchers. Focusing on Italian transport infrastructure companies, this case study showcases how LLMs can automate data extraction, identify trends and patterns, and generate actionable insights from complex sustainability reports, facilitating cross-company comparisons and identifying opportunities for best practice adoption.

The key findings confirm that GenAI is a powerful tool with the potential to significantly contribute to the field of sustainable business practices. Our results provide valuable insights into the nascent but rapidly accelerating literature on GenAI and environmental sustainability. From a practical standpoint, the demonstrated efficacy of LLM-based tools in assessing EU Taxonomy alignment suggests a promising avenue for more scalable and consistent regulatory monitoring, potentially reducing the resource burden for companies, especially SMEs. For academia, this research lays a foundational understanding of GenAI's impact on corporate sustainability and outlines fertile ground for future research, including exploring other GenAI models and refining methodological approaches. Ultimately, this work advocates for the integration of GenAI into ESG auditing and reporting frameworks to foster a more transparent, accountable, and sustainable business environment.

Despite the promising potential highlighted, this work also acknowledges several limitations. A critical aspect for future research involves developing robust methodologies for validating the reliability and accuracy of AI models used in this domain, ideally against a "golden dataset" to ensure replicability of results. Furthermore, while Generative AI offers significant advantages, the continuous evolution of this technology necessitates a strategic approach: creating bespoke, dedicated models for ESG analysis may not be cost-effective given the rapid pace of innovation and the constant emergence of new, powerful out-of-the-box solutions. Instead, grounding existing highly-capable LLMs with specific ESG information appears to be a more viable and future-proof strategy, allowing continuous leveraging of their advanced features. The ability to ground information with the source of the answer is also crucial for maintaining transparency and trustworthiness in AI-generated insights.

Main field: The main field of this research lies at the intersection of Generative AI, Business Management, and Environmental Science.

Research Question:

- What are the primary applications of Generative AI in enhancing corporate environmental sustainability, and what are the potential benefits and challenges?
- How can Generative AI automate the extraction of key data points from non-financial documents to improve the efficiency and accuracy of sustainability reporting?
- How effective are LLMs in assessing the alignment of corporate disclosures with the EU Taxonomy, and how do different LLM techniques (RAG vs. Long Context) compare in this task?

1. Introduction

1.1. The EU Taxonomy and its importance

In recent years the European Union (EU) has been at the forefront for the production of legislative frameworks that should promote firms' environmental sustainability. The most recent regulatory development in this sense has been the EU Taxonomy (EU Regulation 2020/852). The EU Taxonomy is a classification system developed to provide a clear framework for identifying environmentally sustainable economic activities. It establishes science-based criteria across six environmental objectives to guide investors, companies, and policymakers in aligning financial flows with the EU's sustainability goals. As part of the broader EU Green Deal and Sustainable Finance agenda, the Taxonomy aims to enhance transparency, reduce greenwashing, and support the transition toward a low-carbon, resilient economy (*Hummel & Jobst, 2024*).

At its core, the EU taxonomy defines six fundamental environmental objectives:

Climate Change Mitigation: Focused on reducing greenhouse gas emissions and transitioning to a low-carbon economy, this objective encourages businesses to adopt renewable energy sources, invest in energy efficiency measures, and reduce their carbon footprint.

Climate Change Adaptation: Recognizing the urgency of climate change, this objective aims to enhance the resilience of businesses and communities to climate-related risks. It promotes adaptation strategies like implementing flood defenses, developing drought-resistant crops, and improving disaster preparedness.

Circular Economy: Promoting the efficient use of resources and minimizing waste, this objective encourages businesses to adopt circular economy principles. It includes designing products for longevity, promoting reuse and repair, and establishing effective waste management systems.

Pollution Prevention and Control: Aiming to reduce pollution and safeguard human health and the environment, this objective encourages businesses to adopt cleaner production technologies, implement emission control measures, and responsibly manage hazardous substances.

Protection and Restoration of Biodiversity and Ecosystems: Recognizing the significance of biodiversity and healthy ecosystems, this objective aims to conserve and restore natural habitats, protect endangered species, and promote sustainable land use practices.

Sustainable Use and Protection of Water and Marine Resources: Emphasizing the sustainable management of water resources and the protection of marine ecosystems, this objective encourages businesses to adopt water-efficient practices, reduce water pollution, and promote sustainable fishing practices.

The EU taxonomy's significance lies in its ability to drive positive change and promote sustainable economic activities (*Schütze & Stede, 2024*). It provides a robust framework for businesses to assess and report on their sustainability performance. This information empowers investors to make informed decisions, allocating capital towards companies that align with their sustainability values. Moreover, the EU taxonomy serves as a valuable tool for policymakers to develop targeted policies and regulations that support sustainable economic growth. By aligning with the taxonomy's objectives, policymakers can create an enabling environment for businesses to innovate, invest, and thrive sustainably.

The EU taxonomy is still evolving, but its impact is already being felt throughout the business landscape (*Hummel & Bauernhofer, 2024; O'Reilly et al., 2024*). Forward-thinking companies have embraced the taxonomy, using it to develop comprehensive sustainability reports and demonstrate their commitment to responsible business practices. Investors are also increasingly incorporating the EU taxonomy into their investment strategies. By aligning their portfolios with the taxonomy's objectives, investors can contribute to a more sustainable and resilient economy.

The EU taxonomy exemplifies the European Union's steadfast commitment to fostering sustainability. It furnishes a robust instrument for enterprises, investors, and policymakers to instigate salutary transformations and cultivate a future wherein economic advancement and environmental preservation coalesce. As the EU taxonomy undergoes continuous refinement, it harbors the capacity to modulate global sustainability endeavors and exert influence upon businesses and investors globally. Its ramifications are likely to transcend the confines of the European Union, thereby galvanizing analogous undertakings and regulatory constructs in diverse geographical domains and legal frameworks. The EU taxonomy constitutes a momentous stride in advancing sustainability and conscientious business modalities. By proffering a common lexicon and a standardized operational framework, it empowers enterprises, investors, and policymakers to collaborate synergistically toward a more sustainable future.

1.2. The challenge of processing non-financial documents

The increasing demand for transparency and accountability in corporate sustainability reporting has led to a surge in the volume of non-financial disclosures (NFDs) produced by companies. This surge is driven by firms' efforts to comply with regulatory requirements (*Hummel & Jobst, 2024*). While these documents offer valuable insights into a company's environmental, social, and governance (ESG) performance, their production and analysis can be a daunting task. This difficulty arises not only because many firms lack the necessary in-house expertise, leading to suboptimal information quality (*O'Reilly et al., 2024*), but also because sustainability disclosure frameworks often overlap in their information requirements (*Moreno & Caminero, 2024*). Consequently, the inherent features of the European context introduce significant heterogeneity in these documents.

NFDs manifest in various formats, such as integrated reports, sustainability reports, and CSR reports, making it challenging to extract consistent and comparable information across companies. Even within the same reporting format, there is considerable variation in the type and detail of information provided (*Breijer et al., 2025*). For instance, one company's sustainability report might prioritize environmental issues, while another focuses on social or governance factors. This inconsistency hinders comparisons of companies' ESG performance and the identification of best practices, even within the same sectors. Such a lack of standardization complicates informed decision-making for investors, analysts, and other stakeholders based on ESG information (*Ong et al., 2025*). As a result, analyzing these reports can become profoundly subjective, leading to excessive variability in firm assessments (*Büyüközkan & Karabulut, 2018; Hope, 2003*). For example, one analyst might interpret a company's sustainability commitment as strong ESG performance, while another might view it as mere greenwashing. Therefore, extracting meaningful information from NFDs is often a time-consuming and labor-intensive process.

However, the recent developments in the artificial intelligence domain might offer a new landscape for analysts and stakeholders to assess the performances of firms based on the information provided through the NFDs (*Föhr et al., 2023*). For instance, traditional ML models are capable of processing vast amounts of data to extract meaningful insights from structured, tabular datasets, thereby mitigating the subjectivity inherent in human judgment (*Ong et al., 2025*). A key limitation of these models lies, though, in their reliance on structured inputs, whereas NFDs are predominantly unstructured in nature: these documents typically combine numerical data with textual narratives and visual elements such as charts and images. For example, a sustainability report may include descriptive accounts of a company's environmental initiatives alongside infographics to support its claims. This unstructured format poses significant challenges for the application of conventional data analysis techniques (*Maibaum et al., 2024*). Thus, given the peculiar structure of NFDs, another kind of AI application is becoming the preferred option to analyze this type of documents, i.e. LLMs (*Cao & Zhai, 2023; Maibaum et al., 2024*), thanks to their ability to deal with unstructured data.

LLMs are a class of AI systems trained on massive corpora of textual data, enabling them to understand, generate, and summarize human language with a high degree of contextual awareness (*Sedkaoui & Benaichouba, 2024*). Unlike traditional machine learning models, LLMs are specifically designed to process unstructured data, such as the narrative components commonly found in NFDs. Their ability to comprehend context, extract relevant information, and interpret nuanced language makes them particularly well-suited for analyzing the heterogeneous and complex content of non-financial disclosures. Specifically, with regards to corporate sustainability, LLMs can be particularly proficient in three main tasks (*Föhr et al., 2023; Zhuang & Wu, 2023*). First, LLMs can automate the data extraction process. These models, in fact, can be exploited, even without a specific training (*Bommasani et al., 2021*), to extract key data points from non-financial documents with high accuracy and efficiency. This includes extracting information such as:

- ESG metrics: LLMs can extract ESG metrics such as carbon emissions, water usage, and waste production from sustainability reports, press releases, and other non-financial documents.

- Risk factors: LLMs can identify risk factors related to ESG issues, such as climate change or supply chain disruptions.
- Regulatory compliance data: LLMs can extract regulatory compliance data, such as permits, licenses, and certifications, from legal documents and government filings.

By automating this process, LLMs can significantly reduce the time and cost associated with manual data extraction, which is often a tedious and error-prone task. The extracted data can then be easily integrated into existing data systems and utilized for further analysis.

Second, LLMs can identify trends and patterns that are not readily observable by a human analyst. For instance, they are capable of identifying companies that demonstrate consistent improvements in their ESG performance over time, as well as highlighting specific sectors that are encountering persistent ESG-related challenges. Furthermore, LLMs can detect emerging themes, such as the increasing emphasis on biodiversity, that may signal shifts in stakeholder expectations or regulatory priorities. These insights can support investors and analysts in making more informed investment decisions and in identifying strategic opportunities for engaging with companies on ESG matters. Third, LLMs can generate valuable insights from NFDs by identifying relationships between disparate pieces of information and drawing inferences from the extracted data. This capability allows the detection of potential ESG risks and opportunities, the evaluation of how ESG factors may influence financial performance, and the identification of best practices in ESG reporting. Such insights enhance stakeholders' understanding of a company's ESG performance and support decision-making processes that are aligned with specific sustainability priorities and investment strategies.

By addressing the challenges of processing NFDs, LLMs can unlock the value of the information contained in those documents and enable stakeholders to make more informed decisions (Gu et al., 2024). This can contribute to a more sustainable and transparent business environment, where companies are held accountable for their ESG performance.

1.3. Assuring a sustainable Europe via LLMs

Complementing the EU Taxonomy is the Corporate Sustainability Reporting Directive (CSRD), a significant piece of EU legislation that aims to enhance the transparency and accountability of corporate sustainability reporting. Building upon the previous Non-Financial Reporting Directive (NFRD), the CSRD introduces more detailed reporting requirements across a wide range of ESG issues (*Hummel & Jobst, 2024*). The directive's implementation is being phased in, with the first set of companies, those already subject to the NFRD, required to apply the new rules in the 2024 financial year for reports published in 2025. This initial phase will be followed by the inclusion of all large companies and listed SMEs in subsequent years. The CSRD mandates reporting according to the European Sustainability Reporting Standards (ESRS), which are being developed to provide specific guidance on the information to be disclosed, ensuring comparability, reliability, and verifiability.

However, the recent Decreto OMNIBUS has introduced a notable development, effectively reducing the scope of companies required to comply with the CSRD, at least in the short term. This decision means that fewer European-based companies will be immediately obligated to adhere to the new, more stringent reporting standards. Building on the delayed expansion of the CSRD's scope, it is precisely in this interim period that the need for robust, AI‑driven analysis becomes most pressing. In the short term, we are likely to witness a patchwork of reporting quality and content. In fact, on the one hand, multinational corporations and early adopters will continue to refine their reporting processes, drawing on in‑house sustainability teams or external consultants to meet emerging best practices. On the other hand, a sizable group of SMEs, often lacking dedicated ESG expertise and standardized data‑collection systems (*O'Reilly et al., 2024*), will produce disclosures that vary widely in structure, terminology, and data granularity. This divergence will not only make longitudinal analyses of corporate environmental performance more complicated but may also create a sort of "reporting fatigue" within smaller firms, increasing the risk that sustainability-related information remains incomplete or non‑comparable.

Moreover, external stakeholders, who depend on reliable sustainability data will face significant challenges. For example, under the CSRD's ultimate scope, banks will be required to report the climate‑alignment of their credit portfolios (*Brühl, 2023*). However, in the absence of uniformly disclosed metrics on emissions, resource use, and climate mitigation strategies from borrowing firms, financial institutions will struggle to offer a comprehensive disclosure on how they are financing the green transition. The current landscape of voluntary environmental reporting thus poses a significant risk to the efficacy of the EU's sustainable-finance agenda. While regulations like the Corporate Sustainability Reporting Directive (CSRD) and the EU Taxonomy do provide a detailed structure, their complexity makes them incredibly difficult to implement. The sheer volume of specific data points required, coupled with the technical nature of criterias, creates significant practical hurdles for companies. This often leads to varied interpretations and inconsistent reporting. This disconnect between the ambitious regulatory framework and the practical challenges of its implementation is a key reason why a truly uniform disclosure format remains elusive. Companies struggle to collect, verify, and report the required information. As a result, while the regulations exist in principle, the actual execution leads to a lack of standardization, which can ultimately hinder the goals of the sustainable-finance agenda.

This situation underscores the continued relevance of tools and methodologies needed to process non-financial documents published by firms and to analyze historical non-financial disclosure documents. Analyzing past reporting practices can provide valuable insights into trends, identify areas for improvement in future reporting, and aid in the development of methodologies and tools, including those powered by AI, to better assess and ensure compliance with evolving reporting standards.

Building on these premises, this thesis proposes a framework for using LLMs to analyze corporate sustainability disclosures.

Specifically, the first chapter provides a critical review of the emerging literature on the use of LLMs to enhance corporate sustainability practices. Two main findings emerge from this analysis. First, although the academic literature on the topic remains in its early stages, there is a clear and accelerating trend toward the adoption of LLM-based solutions. These applications range from LLM-supported generation and analysis of corporate sustainability disclosures to more advanced use cases, such as tailoring LLM-driven tools for operational improvements within companies. Second, this chapter demonstrates, through a sort of meta-analytical approach, how ready-made LLM-based tools can be used to conduct literature reviews, highlighting their potential as valuable instruments for researchers. This is particularly significant because it eliminates the technical barrier to entry, enabling scholars without deep expertise in computer science or machine learning to leverage these tools effectively in their academic work.

The second chapter delves deeper into the potential of LLMs to support understanding and extraction of concepts related to the EU Taxonomy for sustainable activities. Using Gemini, Google's proprietary LLM, the chapter compares different prompt-based approaches to retrieve taxonomy-relevant information from the corporate disclosures of two companies in the transportation sector. This analysis not only highlights the models' ability to identify and extract sustainability-relevant content but also provides a preliminary benchmark in terms of time and cost efficiency, metrics that will be essential for assessing the scalability and feasibility of future applications in this space.

Building on this foundation, the third chapter presents a detailed case study illustrating how researchers, or corporate professionals responsible for sustainability reporting, can leverage LLM-based tools to systematically document existing ESG-related activities or to extract taxonomy-aligned information from existing disclosures. This practical application aims to demonstrate the real-world utility of generative AI models in facilitating structured, consistent, and meaningful sustainability reporting.

In sum, this thesis serves as a critical analysis of how recent advancements in genAI are generating a disruption in the field of corporate sustainability disclosure. It examines both sides of the disclosure ecosystem: the internal stakeholders responsible for producing such information, and the external stakeholders who depend on high-quality and comparable data. By acknowledging this ongoing transformation, the goal of this work is to equip practitioners and policymakers with scalable, data-driven tools that can enhance the transparency, comparability, and ultimately the credibility of sustainability information across the corporate landscape.

2. Generative AI & corporate environmental sustainability: a literature review

Climate change has gained increasing importance over the past decades, driven by significant environmental changes that are reshaping the way firms operate and people live their everyday lives (*Rau & Yu, 2024*). In this context, environmental concerns surged to a critical role in each organisation's operations, as companies are increasingly recognising that addressing environmental sustainability is not merely a rhetorical exercise but a critical consideration in strategic planning (*N. Wang et al., 2023*). In fact, public authorities around the world are voicing concerns about the perils created by the reckless negligence of environmental issues (IPCC, 2018). As a result, environmental issues have transitioned from peripheral issues to fundamental factors that firms must integrate into their core strategies to remain competitive and socially responsible.

The increasing importance given to environmental sustainability has pressured firms across various industries to seek innovative solutions for reducing their ecological footprint, and as companies grapple with the complexities of integrating sustainable practices into their operations, the advent of Generative Artificial Intelligence (AI), also referred to as GenAI, presents a transformative opportunity (*Mariani & Dwivedi, 2024*). Generative AI, a subset of artificial intelligence that includes models capable of producing novel content, such as text, images, and videos, has rapidly gained traction for its potential to revolutionize various business functions (*Sedkaoui & Benaichouba, 2024*). As such, these technologies are beginning to influence the ways in which firms address sustainability challenges.

However, despite the growing interest in Generative AI, there remains a dearth of research exploring its implications for environmental sustainability, especially concerning how these technologies can be exploited to improve sustainable business practices. Therefore, this study tries to systematise the existing knowledge about the current use of Generative AI for enhancing corporate environmental sustainability. Specifically, we employ the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) approach (Page et al., 2021), a methodology used to conduct literature reviews based on a transparent and systematic review process, to answer the following research questions:

RQ1. What are the major topics discussed in the literature?

RQ2. What are the potential future research directions?

Through the PRISMA approach, we classify the current research into three thematic clusters, namely "*Applications of Generative AI for improving Sustainability of Business Operations", "Generative AI for Sustainability Reporting and ESG Analysis",* and *"Addressing Environmental Impact of Generative AI"*, offering insights into the emerging applications and ethical considerations of these technologies.

The relevance of this analysis is motivated by the ongoing pressure on companies to reduce their environmental impact on our planet. With their ability to process vast amounts of unstructured data, Generative AI can contribute to climate action by enabling firms to better analyse and mitigate their environmental footprint, thus playing a crucial role in the global effort to combat climate change (*Alzoubi & Mishra, 2024*). Additionally, as environmental regulations become more stringent globally, companies need advanced tools to comply with these standards (*De Villiers et al., 2024*). Generative AI can help firms meet these regulatory requirements more effectively by improving their environmental reporting and ensuring transparency in their sustainability practices (*Zhuang & Wu, 2023*). As a consequence, firms that will be able to integrate these technologies to achieve sustainability goals will be better positioned to adapt to future environmental challenges and economic shifts.

Given the growing importance of both environmental sustainability and technological advancements, we seek to provide a foundational understanding of how Generative AI can be leveraged to drive sustainable practices within firms while also identifying areas for future research and practical application.

The remainder of this chapter is as follows: Section 2 presents the methodology to collect and analyse our data. Section 3 reports the main results of the analysis. Section 4 discusses the main results and presents some potential future research avenues. Finally, Section 5 concludes the study.

2.1. Design of the study

Research Protocol

To provide an overview of the extant literature on the impact of Large Language Models on business sustainability, we adopt a systematic review approach through the PRISMA method. This method lets the researcher summarise the current academic knowledge by relying on a meticulous, transparent and evidence-based stepwise iterative process (*Page et al., 2021*). Consequently, The PRISMA method will help us acknowledge the main themes discussed in the literature so far and understand what research questions need to be answered in the future. Given the perks of this methodology, many studies in the management literature are increasingly adopting the PRISMA method (e.g., *Galletta et al., 2024*; *Scartozzi et al., 2024*).

Our PRISMA diagram is outlined in a three-step process, as depicted in Figure 1. In the first step, the identification phase, the inclusion and exclusion criteria must be defined. To gather the necessary sources, the two authors independently collected a set of keywords for the Boolean search. After the merge of the two sets, the final adopted string was as follows:

```python
( TITLE-ABS-KEY ( llm* OR "large language model" OR *gpt* OR chatgpt OR openai OR "generative ai" OR "language model" OR "genai" ) AND TITLE-ABS-KEY ( environ* OR carbon OR sustainab* OR climat* OR ecologic* OR footprint* OR emission* OR pollution OR greenhouse OR ghg OR biodiversity OR conservation OR renewab* OR energ* OR waste OR recycl* OR taxonomy OR green OR esg ) )
```

Moreover, we also set the following additional inclusion criteria: the articles must be complete and available online, must be written in English, and must be published in the subject areas of Business ("BUSI"), Economics ("ECON"), or Environmental Sciences ("ENVI"). The exclusion criteria were as follows: we exclude books, book chapters, and other non-academic reports[1], as well as articles in languages other than English.

[1] Given that the literature on the use of LLM for business applications pertains to an interdisciplinary approach which brings together computer science and management studies, we decided to keep also the conference papers since these are much valued avenues for the computer science field.

The abovementioned identification was performed on the Scopus database on 23/07/2024. Scopus is a bibliographic database and abstracting service owned by Elsevier. It contains abstracts and citations for academic journal articles, books, book chapters, and conference papers. Scopus is one of the world's largest abstract and citation databases, and it is used by researchers, librarians, and students to find information on a wide range of topics. The choice was motivated by the fact that Scopus provides excellent coverage of peer-reviewed articles from different disciplines and publishers (*Vieira & Gomes, 2009*) and is, in fact, widely used for systematic literature reviews (*Martín-Martín et al., 2018*). Moreover, using Scopus instead of directly selecting only relevant journals allows the researcher to avoid the potential bias of omitting essential sources in the final choice of documents (*Scartozzi et al., 2024*).

This initial process yielded a total of 1,070 documents across journal articles, conference papers, books and book chapters. From this dataset, we exclude all books, book chapters and sources published before 2018, the year of the introduction of the GPT-1 architecture, which opened the path for large-scale production of generative artificial intelligence (*Han et al., 2021*).

The resulting dataset of 694 sources served as the basis for the second phase. In this screening process, all the sources' titles, keywords and abstracts were perused and assessed to exclude documents deemed out of scope.

During this screening process, all the sources' titles, keywords and abstracts were perused and assessed to exclude documents deemed out of scope. The two primary criteria for exclusion focused on removing papers that were obviously not related to generative AI but more related to the use of a broader AI methodology or tool.

The resulting 37 articles were then further assessed for the eligibility criteria set in the first step: we retrieved the full texts for all documents and read them in their entirety to ensure that each document fell within the scope of our research. During this phase, one article was excluded because it was not possible to retrieve its full text via official databases. This final phase led to the inclusion of a total of 36 documents.

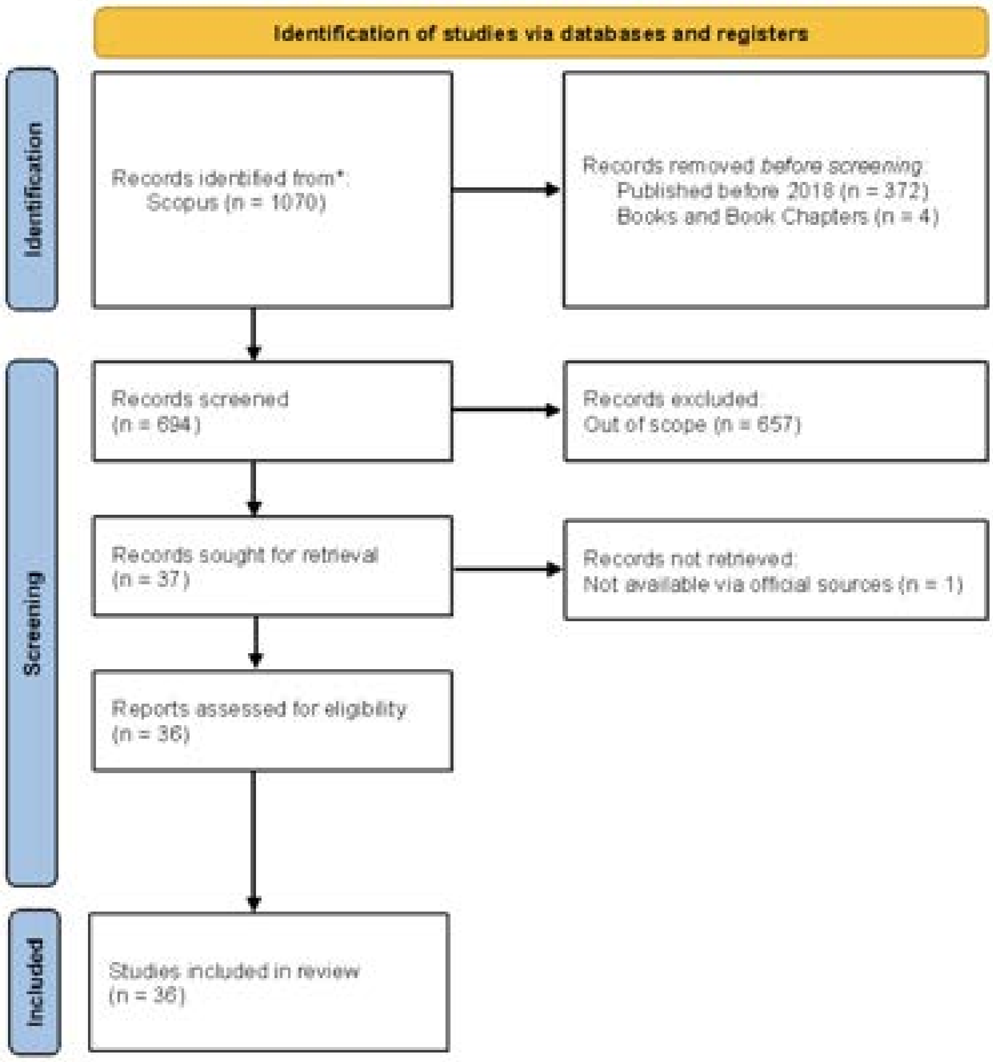

Figure 1: PRISMA flow diagram for systematic literature reviews.

Content Analysis

The final dataset of 36 papers is used as the basis for the systematic literature review in order to map the evolution of the literature comprehensively. Specifically, given the size of the dataset, we opt to conduct a content analysis. Content analysis is a qualitative approach that tries to discover common topics among different texts collected in a corpus (*Krippendorff, 2019*). Typically, content analysis is performed manually by the scholar, limiting the number of documents that can be perused (*Donthu et al., 2021*; *Snyder, 2019*): the novelty of this study is that this step of the analysis is assisted by an out-of-the-box Large Language Model tool called

NotebookLM, which is based on the novel Google-developed Gemini 1.5 Pro multimodal model. NotebookLM is "*an experimental application that helps users 'think better' by organising information and facilitating the generation of insights from various sources. When a user loads sources such as Google Docs, presentations, PDFs, text files or web URLs, NotebookLM creates a guide that summarises the documents and suggests key topics or questions to ask. NotebookLM allows users to ask specific questions about particular sources and receive answers with inline quotations linked to the original text. Users can upload up to 50 sources containing less than 500,000 words in a single notebook. NotebookLM also supports the inclusion of images in presentations and documents to allow users to ask questions about visual information*[2]*"*. As the definition implies, NotebookLM relies on a mix of Retrieval Augmented Generation (RAG) and Long Context technique to overcome one of the most critical issues of LLMs, i.e. they tend to show deteriorated performance when knowledge-intensive tasks must be performed (*Lewis et al., 2020*), since they usually heavily rely on the provided prompt to generate meaningful answers. In other words, when the question being asked is too specific, and the LLM is not trained on data pertaining to the specific knowledge required to answer the question, the LLM might hallucinate and produce nonsensical or completely made-up responses (*Burger et al., 2023*). The RAG and Long Context technique, used by the NotebookLM tool, solves this issue by combining text generation with the retrieval of relevant information from an external data corpus. With RAG, an information retrieval component is introduced into the loop using user input to extract information from a new data source. The user's request and the relevant information are both provided to the LLM. The LLM uses the new knowledge and its training data to create better answers (*Li et al., 2022*). This allows the model to provide more precise and relevant answers based on up-to-date and accurate information. An additional advantage of this tool is that it does not require any pre-existing knowledge about coding and can, thus, be used by any researcher who wants to enhance their knowledge mapping possibilities with AI tools.

In this step of our analysis, we exploit NotebookLM to facilitate the content analysis of the retrieved documents. More specifically, in the first part, we use this tool to provide an initial evaluation of the major topics discussed in the literature so far. In the second part, this tool is

[2] This definition was provided directly by the NotebookLM through a prompt provided to its chat interface.

used to discuss further developments in the literature and generate meaningful prompts for future research directions. Since NotebookLM is only leveraged as an automated research assistant, in each step of the analysis the responses of the LLM are presented and then critically discussed by the authors after a careful reading of the studies included in the review.

2.2. Results

Description of the Dataset

The dataset contains 32 journal articles and 4 conference papers published in the period 2022-2024. More specifically, we have 1 article in 2022, 11 in 2023, and 24 until July of 2024, highlighting both how the literature on the topic is still in its infancy and how it is gradually accelerating. On average, each document has been cited 10.78 times. To get an initial grasp of the topics discussed in our sample, in Figure 2, we present a WordCloud graph of the top 200 words used in either the title, the abstract or the keywords of the 36 analyzed documents. We can observe how the most common words can be subdivided into two major clusters. First, a cluster pertaining to the artificial intelligence field, with words such as “AI”, “ChatGPT”, “model” and “generative”. Second, a cluster related to corporate environmental sustainability, with words such as “climate”, “sustainability” and “industry”. This initial evidence highlights how our sampling methodology has effectively sourced all the relevant documents for our analysis. Moreover, we note how, in this particular context, the literature has focused primarily on the use of Large Language Models as GenAI models, even though other examples exist, such as image-, audio- or video-generative models.

Figure 2: Word Cloud of the 200 most common words in the Title-Abstract-Keywords set of the documents in our sample.

Regarding the journals, 30 different outlets have published the 36 documents, with an average of 1.2 publications and 12.93 citations per outlet. Table 1 presents the list of the five most prolific journals in terms of published articles. We can observe how all outlets but one, i.e. Journal of Chinese Economic and Business Studies, have a quite high impact factor, with a minimum level of 7.3 and a maximum of 11.2. We also note how there is an elevated heterogeneity of the journals' fields: we have a Finance journal (*Finance Research Letters*), a multidisciplinary economics journal (*Journal of Chinese Economic and Business Studies*), a manufacturing technology journal (*Journal of Manufacturing Technology Management*) and two transdisciplinary journals focused on sustainability issues (*Journal of Cleaner Production* and *Resources, Conservation and Recycling*). This highlights the interdisciplinarity of the field, and articles do not focus on specific subjects of business and economics since the impact of firms on the environment needs to be investigated from multiple perspectives.

Table 1: Most prolific journals.

Journal	Articles	N. Citations	Avg. Citations	IF (2023)	AJG (2021)
Journal of Cleaner Production	3	32	10.67	9.7	2
Finance Research Letters	2	72	36	7.4	2
Journal of Chinese Economic and Business Studies	2	34	17	2.4	1
Journal of Manufacturing Technology Management	2	2	1	7.3	1
Resources, Conservation and Recycling	2	2	1	11.2	/

Note: this table lists the five most prolific journals in terms of published articles. IF is an abbreviation for "Impact Factor", while AJG (Academic Journal Guide) is the ranking provided by the Chartered Association of Business Schools.

In terms of authors, our dataset comprises 113 authors, with 2 single-author documents and 34 multi-author documents. The average collaboration index is 3.17, meaning that the average document is authored by at least 3 people. Moreover, regarding the authors' country affiliation, in Figure 3, we can observe that China has the highest number of authors, with 27 researchers, followed by the USA, with 13, and the United Kingdom, with 12.

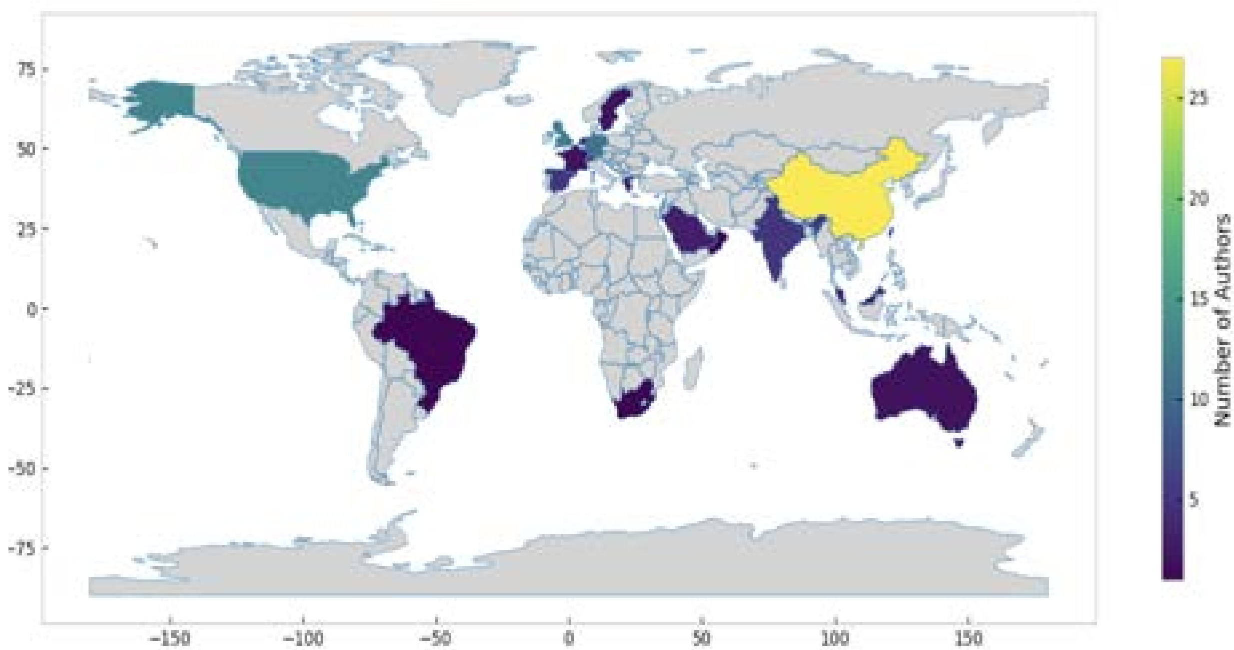

Figure 3: Authors' country affiliation distribution.

Identification of the Thematic Clusters

To answer RQ1, in this section we will review the articles that were identified using the methodology described in Section 2.1. Before this review process, however, we proceed to cluster the articles into groups of homogeneous documents based on the discussed topics. To do that, we ask NotebookLM to provide an initial description of the number of topics and of the content of each topic. In this phase of our analysis, we utilize NotebookLM for the content analysis of the compiled documents. NotebookLM excels in identifying thematic elements across a corpus of uploaded documents, serving as a robust instrument specifically engineered for rigorous research and analytical undertakings. Its design is based on the use of the provided sources, which helps reduce errors and ensures that the insights are supported by the original material. This structure for document analysis, alongside functions like synthesis, accurate referencing, and the ability to work with different formats, makes it a reliable tool for studying specific textual sources and organizing research in an efficient way.

After a brief trial-and-error phase, we ended up with a prompt which provided consistent and satisfactory results. The final prompt we used is as follows:

Python
```python
"You are a researcher in the economics and business field. You are doing a systematic literature review about the implications of the introduction of Generative AI for the environmental sustainability of firms. You want to cluster all the uploaded documents into different groups depending on the major topic discussed in each document so that you can subsequently make more in-depth analysis. How many clusters can you identify? Make the clusters as comprehensive and specific as possible. Justify your answer. I suggest making as many clusters as possible and then merge the clusters that are too similar to be stand-alone clusters."
```

This prompt generated the following output:

```Python
"Following your suggested approach, a comprehensive analysis of the provided sources reveals seven potential clusters, some of which can be merged due to their close relationship:
- Generative AI-Enhanced Sustainability Reporting […];
- Combating Greenwashing with Generative AI […];
- Generative AI for Sustainable Manufacturing and Operations […];
- Generative AI for Sustainable Product Design and Lifecycle Assessment […];
- Green AI Initiatives and Tools […];
- Generative AI for ESG Risk Assessment and Sustainable Finance […];
- Generative AI and Responsible AI Development for Sustainability […].
In summary, the sources point to five overarching clusters:
- Generative AI for Enhanced Sustainability Reporting.
- Generative AI for Sustainable Manufacturing and Operations.
- Generative AI for Sustainable Product Design and Lifecycle Assessment.
- Green AI Initiatives and Tools.
- Generative AI for ESG Risk Assessment and Sustainable Finance."
```

Based on this initial clustering, we analyzed all documents and further merged the clusters provided by the LLM into three main topics emerging from the literature: Applications of Generative AI for improving Sustainability of Business Operations, Generative AI for Sustainability Reporting and ESG Analysis, and Addressing the Environmental Impact of Generative AI.

Cluster 1: Applications of Generative AI for Improving Sustainability of Business Operations

The first cluster discusses how Generative AI can be used to improve the environmental sustainability of business operations and practices. While the use of GenAI poses both challenges and opportunities, if the application of this new technology is designed with a holistic approach in mind, the gains can more than counterweigh the potential hindrances, letting firms combine economic profitability with environmental sustainability (*Ghobakhloo et al., 2024*). Indeed, thanks to its flexibility and almost unlimited customizability, GenAI can play an essential role in product and process optimisation (*Kar et al., 2023*). One example of this kind of optimisation is the use of LLM to improve the life cycle assessment (LCA), i.e. the evaluation of the potential environmental impacts of products/services during their life cycle (*Cornago et al., 2023*). Indeed, the use of LLM can tremendously expedite this task in terms of data retrieval and text processing, improving the quality of the LCA process while simultaneously making it more time-efficient and letting the human expert focus on more intellectually-demanding tasks (*Preuss et al., 2024*). *Feng et al. (2024)*, for example, implement an advanced management system combining GenAI and a two-stage auctions model to optimise the recycling process of electric vehicle batteries, showing how their method not only improves the overall environmental impact of the entire recycling process but also guarantees a satisfactory level of profitability for all the economic stakeholders involved.

Another example of how GenAI can also help improve firms' operations is the management of supply chains (*Kar et al., 2023*). Similarly to what is happening in the case of LCA, the ability of LLMs to manage vast amounts of textual data and find trends that may not be obvious to the human eye can help firms truly incorporate sustainability into supply chain management and planning (*Fosso Wamba et al., 2024*; *Haddud, 2024*). This can improve the overall efficiency of a firm's logistics, reducing waste and creating an actual green supply chain (*Susithra & Vasantha, 2024*). Moreover, LLMs can also be used as an instrument for generating accurate predictions on resources utilisation providing organisations with insights to make informed decisions and contribute to sustainable development (*H. Yang & Zhou, 2023*).

Nevertheless, also the production process itself can be strongly impacted by the introduction of LLMs. For example, *Akhtar et al. (2024)* examine the integration of Generative AI into Smart Product Platforming to advance personalised and sustainable manufacturing within the circular economy framework. They underscore how these technologies enhance product design, optimise resource utilisation, and support environmentally sustainable production. The study, focusing on the automotive industry, demonstrates that GenAI-driven Smart Product Platforming is essential for fostering the sustainability of the manufacturing process. On another front, *T. Yang et al.*

(2024) discuss how LLMs have significant implications for intelligent agriculture production by enhancing the precision and efficiency of question-answering systems. They enable accurate decision-making through better contextualisation of complex agricultural data, from weather patterns to soil conditions, supporting sustainable practices. LLMs can integrate and analyse diverse data sources, offering real-time insights for optimising crop management with minimal environmental impact. However, to fully realise their potential in advancing smart farming technologies and promoting environmental sustainability, challenges such as domain-specific training and data standardization must be addressed.

Finally, the literature highlights how, even though the opportunities are promising, the threats coming from the introduction of this innovation need to be considered. The fact that GenAI is a completely new and disruptive technology may generate downsides that many firms may still not be able to hedge (*Ghobakhloo et al., 2024*), such as data privacy issues or model complexities that require expert knowledge (*Fosso Wamba et al., 2024*). However, in this context, given that AI adoption can in fact improve firms' environmental performances (*C.-T. Chen et al., 2024*), the Governments should play a key role by supporting firms through private-public partnerships that can enhance the sustainability, both economic and environmental, of such projects (*S. Wang et al., 2024*). On this front, *Alkire et al. (2024)* develop a strategic framework for integrating AI in a responsible way into industries: their framework emphasises the need for a collaborative development of these projects that align with both societal well-being and sustainable development.

Cluster 2: Generative AI for Sustainability Reporting and ESG Analysis

The second cluster encompasses two main topics related to the use of Generative AI: the use of LLMs to improve the quality of sustainability reporting, on one side, and the design of AI-powered techniques to develop innovative ESG assessment tools, on the other side. Given that sustainability reporting and ESG analyses are essentially based on annual reports of firms, LLMs, with their ability to deal with large amounts of unstructured texts, are the preferred configuration of GenAI used in these kinds of applications (*Cao & Zhai, 2023; Maibaum et al., 2024*).

One of the main issues that firms face when dealing with environmental reporting is that many accounting methods exist, thus leading to a situation in which private organisations may need to report the same data under several, and sometimes overlapping, frameworks (Ramautar et al., 2024). This creates a paradoxical situation in which companies might report the same data multiple times, wasting valuable resources that could otherwise focus on more purely intellectual efforts (*Zhuang & Wu, 2023*). However, as *Ramautar et al. (2024)* show, LLMs can effectively identify intersections between different environmental accounting methods to create a unified framework that firms can efficiently employ to report their environmental performances. As already discussed, the integration of Generative AI in sustainability reporting offers both opportunities and challenges. Indeed, as *De Villiers et al. (2024)* highlight, GenAI can enhance efficiency and standardisation, streamlining report production and reducing costs. However, human judgment remains essential to ensure the accuracy and completeness of AI-generated content, as management is accountable for these reports. A significant risk is the potential for AI to perpetuate greenwashing by replicating biased or misleading information, thus undermining report credibility. Additionally, AI's influence may shift power dynamics within organisations, favouring those with technological expertise, and potentially embedding certain sustainability ideologies that do not reflect the diverse values of all stakeholders. Therefore, careful oversight and a balanced approach are critical when leveraging AI in sustainability reporting. As a final remark, the literature also implies that LLMs like ChatGPT will be leveraged in the future not only in sustainability reporting but also in developing internal sustainability strategies with an overall improvement of the corporate social responsibility of the firm (*Zhuang & Wu, 2023*).

The second major theme discussed in this cluster relates to the use of LLMs to evaluate a company's ESG performance and potential risks. These models can, in fact, effectively be used to identify topic-related discussions, especially with regard to ESG-related ones, both in official corporate reports (*Huang et al., 2023*) and in social media (*Guo et al., 2024*).

Bingler et al. (2022), for example, evaluate corporate climate risk disclosures by firms that support the Task Force for Climate-related Financial Disclosures (TCFD). The authors utilise *ClimateBERT*, a deep learning model fine-tuned from *BERT*, to examine the content of climate-related disclosures from 818 firms for the fiscal periods 2014-2019. The study's key finding is that the support for TCFD by these firms often amounts to "cheap talk", meaning that firms tend to selectively disclose non-material climate information rather than comprehensive, actionable data. The analysis highlights significant cherry-picking behaviour, where firms prioritise less relevant information, particularly in the categories of strategy and metrics, which are crucial for effective climate risk management by investors and other stakeholders. This selective disclosure undermines the intended transparency and utility of TCFD reports in assessing and managing climate-related risks.

On the same front, *Moreno & Caminero (2024)* conduct an analysis of European financial institutions' progress in climate-related disclosures using text-mining techniques. The authors create a Greenhouse Gas Disclosure Index (GHGDI) by analysing 23 granular disclosures in the ESG reports of significant banks under European Central Bank supervision from 2019 to 2021. This index is then compared to the scores given by the Carbon Disclosure Project (CDP). The study finds a moderate correlation between banks that do not report to CDP and a low GHGDI. However, banks with high CDP scores do not necessarily have high GHGDI, suggesting inconsistencies between different reporting standards. The analysis highlights the potential of combining traditional Natural Language Processing methods with (LLMs) to improve the accuracy and precision of context verification in text mining. The study concludes that while GHGDI shows promise in assessing progress towards better climate-related disclosures, further refinement is needed to fully capture the complexities of corporate climate reporting.

From these two examples, we can understand how, in this field, LLMs can be exploited to standardise pieces of information coming from different sources and from different reporting methods, rendering the ESG scoring process more streamlined, even for smaller firms that usually do not receive such attention (*Nandiraju & Kanthi, 2024*). However, one caveat that comes from the literature is that publicly available LLMs are typically general-purpose, meaning that they are trained on a large corpora of documents that cover many different topics. Therefore, they may struggle when the language used is context-specific, such is the case for financial and ESG reports. Thus, to get the best performances, LLMs need to be fine-tuned[3] to specifically

[3] *Fine-tuning* refers to the process of taking a pre-trained model, which has already been trained on a large and general dataset, and adapting it to a specific task or dataset by continuing the training process with a smaller, task-specific dataset.

recognise corporate climate-related vocabulary (*Trajanov et al., 2023*). And even though this process can become very expensive in the case of generative models, fine-tuned LLMs tend to achieve the greatest performances in this specific task compared to other traditional NLP techniques, and are slowly becoming the preferred tools to extract insights from financial texts (*Huang et al., 2023*).

Cluster 3: Addressing Environmental Impact of Generative AI

The third cluster presents a comprehensive overview of the potential environmental impacts associated with the development and deployment of Generative AI in industry and civil society. In fact, Generative AI can be seen as a double-edged sword in terms of environmental impacts. On the one hand, artificial intelligence can be used by firms to implement the most effective environmentally friendly strategies and by academia to support climate change research (*Agathokleous et al., 2023*; *B. Chen et al., 2023*). For example, *Toetzke et al. (2023)* suggest that one of the potential deployments of LLMs is the monitoring of climate technology innovation: this can help policymakers in designing effective measures to support firms that are committed to developing environmentally friendly innovations. On the other hand, it is known that the AI industry is one of the most energy-intensive in the information technology domain (*Stahl & Eke, 2024*; *Xu et al., 2024*). To give a hint of the impact of GenAI on the environment, the estimated energy consumption of the training phase of GPT-3, i.e. the first GPT iteration available in the preliminary version of ChatGPT, is equivalent to that of 123 gasoline-powered passenger cars for an entire year (*Alzoubi & Mishra, 2024*).

Therefore, given the double-faced nature of AI, a multi-stakeholder perspective is needed, one that allows the confluence of the interests of both industry and academia in the development of a responsible AI (*Polyportis & Pahos, 2024*; *Stahl & Eke, 2024*). Regarding academia, universities should start developing specific curricula that encompass both the new technological advancements and the need for expertise on sustainable development (*Abulibdeh et al., 2024*): Generative AI tools can be leveraged as catalysts for reshaping the learning experience (*Rillig & Kasirzadeh, 2024*; *Sommer & Von Querfurth, 2024*). The integration of AI into the realm of education for sustainable development should create a new workforce with both the expertise and the sensibility to implement these new technologies in their everyday life. Regarding industry, firms should start assessing available sustainable AI tools that can be implemented within organisations in order to reduce the negative environmental impact that information technology brings about (*Alzoubi & Mishra, 2024*). In this, both academia and the GenAI itself can help in identifying the best investment opportunities to develop projects that generate the highest long-term environmental returns (*Abulibdeh et al., 2024*; *Leippold, 2023*).

2.3. Discussion and Future Research Avenues

Based on the clusters identified in the previous section, we answer RQ2 by developing future research directions for the researchers. This step of the analysis was aided by the use of NotebookLM. More in detail, we initially asked NotebookLM to suggest some potential research directions that could depart from the clusters analysed in Section 3.2. The prompt used in this phase is as follows:

Python
```python
"You are a researcher in the economics and business field. You are doing a systematic literature review about the implications of the introduction of Generative AI for the environmental sustainability of firms. You want to identify the future research directions of the literature based on the three following clusters: Cluster 1: Applications of Generative AI for improving Sustainability of Business Operations, Cluster 2: Generative AI for Sustainability Reporting and ESG Analysis, Cluster 3: Addressing Environmental Impact of Generative AI. Justify your answers."
```

The generated output is the following:

None
```
Here are some potential future research directions for the literature on Generative AI and environmental sustainability of firms, based on the three clusters:
**Future Research Directions on Generative AI and Environmental Sustainability**
```

```
Cluster 1: Applications of Generative AI for Improving Sustainability of Business Operations
- Investigate the use of Generative AI for designing eco-friendly products and processes. […]
- Explore how Generative AI can be used to optimize supply chains for sustainability. […]
- Examine the potential of Generative AI to enhance resource efficiency in various business operations. […]
Cluster 2: Generative AI for Sustainability Reporting and ESG Analysis
- Examine the potential of Generative AI to automate the process of sustainability reporting, […]
- Investigate how Generative AI can enhance the quality and comparability of sustainability reports across different firms. […]
- Address the ethical and practical challenges associated with using Generative AI in sustainability reporting. […]
- Explore the use of Generative AI for ESG analysis and investment. […]
Cluster 3: Addressing Environmental Impact of Generative AI
- Develop a more comprehensive understanding of the carbon footprint of Generative AI. […]
- Explore strategies to mitigate the environmental impact of Generative AI. […]
- Investigate the use of Generative AI itself to address its own environmental impact. […]
```

This output served as a trigger for a critical discussion of the further developments of the literature, which we present following the clusters developed previously. *Table 2* presents some potential challenges that could be addressed in future research.

With regards to the first cluster, i.e. "Applications of Generative AI for improving Sustainability of Business Operations", the main challenges encompass the use of Generative AI for optimising the sustainability of production and processes. While literature highlights that LLMs show great potential for implementation in firms' operations (*Kar et al., 2023*), there is still a dearth of studies presenting the actual use of GenAI to improve the sustainability of business processes. *Preuss et al. (2024)*, for example, suggest how LLMs could be greatly exploited by LCA practitioners to improve the efficiency and the timing of their assessment; moreover, since LLMs are AIs specialised in interpreting and generating texts, they could be used in the future also to help the dissemination of LCA results to the general public. This suggestion, although intriguing, needs a careful consideration of the potential threats related to the hallucination risk that all GenAI models are prone to (*Cornago et al., 2023*): this could potentially pave the way for future researches that specifically address the comparison of the question-answering process between LLMs and human experts.

Moreover, GenAI can potentially also improve the sustainability of supply chains. Some authors, in fact, highlight how LLMs, and ChatGPT in particular, could foster green supply chain practices by allowing more precise tracking of the whole process (*Haddud, 2024*; *Kar et al., 2023*). These suggestions, however, still need to be tested in practice: given the ability of LLMs to grasp diverse concepts and pieces of information coming from different sources, researchers could investigate, for example, whether these models can effectively be leveraged to reduce the environmental impact of supply chain operations.

Finally, on the same front, also the production process could greatly benefit from the implementation of LLMs in the design process since these models could be used to identify potential opportunities for waste reduction, reuse, and recycling, leading to a more circular economy.

The second cluster deals with two main issues in the economics literature: sustainability reporting and the analysis of firms' environmental impact.

When dealing with sustainability reporting, firms are confronted with many accounting methods, which often overlap in terms of reported items (*Ramautar et al., 2024*). Therefore, LLMs could help address the issue of heterogeneity of environmental accounting frameworks by supporting regulators in developing a standardised reporting taxonomy for sustainability data, ensuring consistency and comparability across different firms and industries. Moreover, considering the generative nature of these models, GenAI could be further exploited to automatise some steps of the reporting process, thus leaving human experts to the actual evaluation of the environmental impact of the firms. Obviously, this suggestion may raise concerns regarding the ethical considerations and the potential biases associated with GenAI-generated sustainability reports. Therefore, future research advocates for a thorough investigation of the possible safeguards and validation mechanisms that could be put in place to ensure the accuracy and reliability of such AI-generated reports.

On the other side, the heterogeneity of accounting and reporting standards could further pave the way to the implementation of Generative AI in the field of ESG analysis. LLMs are, in fact, able to identify ESG-related discourse both in long unstructured texts, such as company reports (*Huang et al., 2023*), and in shorter social media posts (*Guo et al., 2024*). Therefore, these models could be used to extract meaningful information from a myriad of different sources in order to develop a comprehensive and real-time assessment of a company's environmental performance, which should not be affected by the specific accounting standard adopted by each firm. Thus, future studies should investigate whether these AI-generated measures produce results that are consistent with, or even superior, to indicators already existing in the literature. Additionally, LLMs could also be exploited to produce measures for theoretical constructs that are difficult to grasp for traditional text mining techniques. For example, GenAI might be leveraged to analyse corporate commitment towards environmental policy by inputting texts extracted from corporate financial statements or from earnings calls: given their understanding of different contexts, LLMs can create indicators that could surpass the simplistic dictionary-based measures typically introduced in the literature (*Loughran & McDonald, 2020*).

Lastly, the key challenges highlighted in the third cluster relate to the environmental impact of Generative AI. Indeed, it is well known that the whole training and employment process of these models is quite energy-intensive (*Xu et al., 2024*). However, it would be interesting to understand which steps in this process have the most significant impact on the carbon footprint (*Alzoubi & Mishra, 2024*): therefore, researchers should investigate the main determinants of energy consumption and emissions throughout the lifecycle of Generative AI, as well as understand how these drivers can be effectively measured and benchmarked to assess the environmental footprint of different GenAI applications, both in terms of proprietary and open source models.

Moreover, our results show how there is a renewed awareness towards the introduction of "green AI" tools as a response to the environmental impacts of artificial intelligence systems. However, as highlighted by *Alzoubi & Mishra (2024)*, these instruments are still in the infancy stage and firms look at them cautiously because, while showing great potential, they are also affected by a series of critical weaknesses, such as privacy concerns and the technical complexity that retains firms without specific know-hows from using them. It follows that further studies could thoroughly investigate why and how firms should implement these green AI initiatives and the requirements for their effective implementation within corporate strategies and processes. Additionally, research is needed to validate the discussions around green AI and explore new initiatives in this rapidly evolving field.

Finally, a fascinating suggestion that stems from the analysis of the third cluster revolves around the use of GenAI as a meta-tool to improve itself. Since these models are, in fact, capable of understanding context and producing insightful solutions by leveraging different sources, it could be interesting to investigate the use of GenAI to address its own environmental impact. This solution, if feasible, could not only have a non-negligible impact on the reduction of the environmental footprint of GenAI models but could also serve as a starting point to develop initiatives that are tailor-made to the specific needs of firms interested in implementing AI in their operations.

Overall, this study provides three valuable preliminary insights into how Generative AI has been used so far for improving corporate environmental sustainability. First, it shows that the three main applications of GenAI in this context are life cycle assessments, logistics, and corporate reporting. Those are the business areas most affected by environmental concerns and, consequently, are the first application fields that raised interest in researchers.

Second, the review shows that the literature is still in its infancy since the capabilities of these models are largely unexplored in many fields, not only in terms of corporate functions but also in terms of algorithms employed. Indeed, as tools such as ChatGPT become more and more integrated into the work-life of employees, AI-driven technologies will probably have a durable effect on how each specific function will be carried away (*De Villiers et al., 2024*), from marketing to finance, from customer care to human resources management. However, given the recency of the introduction of such instruments, researchers are tasked with investigating how this process is taking place. Moreover, we also highlight how, in our review, the concept of Generative AI is strictly linked to the concept of Large Language Models. This should not surprise, considering that ChatGPT was the first GenAI model available to the general public and is, thus, the only model that most people could have had experience with. However, many other models with different capabilities exist, such as DALL-E, for image generation, or Sora, for text-to-video generation. All these models could have a different impact on how firms will approach environmental sustainability: to the best of our knowledge, however, no theoretical or empirical research has been conducted so far on the application of these models in the corporate context.

Third, the identification of the three clusters has laid the foundation for developing a framework that enhances our understanding of the various ways generative AI can positively (or negatively) impact corporate sustainability. This framework has guided the identification of future research avenues in this domain, particularly highlighting the need for studies on generative AI to explore the mechanisms connecting different impact areas within complex and interconnected sustainability challenges. Addressing the identified research directions will deepen our theoretical insights into the ethical and environmental implications of applying generative AI in corporate sustainability.

Table 2: Potential future research questions.

Cluster	Potential Research Questions
Cluster 1	How can GenAI enhance supply chain transparency and traceability by tracking material flows, identifying sources of environmental impact, and enabling better monitoring of ethical sourcing practices?
	Can GenAI be used to generate designs that minimise waste during production and facilitate more straightforward repairs and component replacements?
	How can GenAI enhance supply chain transparency and traceability by tracking material flows, identifying sources of environmental impact, and enabling better monitoring of ethical sourcing practices?
	To what extent can GenAI contribute to reducing water usage in manufacturing processes by identifying optimisation opportunities, simulating different scenarios, and proposing alternative process designs or water recycling methods?
	Can GenAI-powered systems effectively identify opportunities for waste reduction, reuse, and recycling, leading to a more circular economy?
Cluster 2	How can GenAI address the challenges of data heterogeneity and subjective interpretations of sustainability indicators?
	How can GenAI be utilised to develop standardised reporting frameworks and taxonomies for sustainability data, ensuring consistency and comparability across different firms and industries?
	What are the ethical considerations and potential biases associated with using GenAI to generate sustainability reports?
	What safeguards and validation mechanisms are necessary to ensure the accuracy and reliability of GenAI-generated reports?
	Can GenAI-based tools assist regulators and investors in assessing the credibility and authenticity of corporate sustainability disclosures?
	How can GenAI enhance the analysis of unstructured data sources to provide a more comprehensive and real-time assessment of a company's environmental performance?
Cluster 3	What are the main determinants of energy consumption and emissions throughout the lifecycle of Generative AI?
	How can these drivers be effectively measured and benchmarked to assess the environmental footprint of different GenAI applications?
	Can GenAI itself contribute to developing more sustainable hardware and infrastructure for its own operations?
	What are the barriers and incentives for adopting green GenAI practices within the tech industry?
	How can GenAI contribute to accelerating the transition to a low-carbon economy?

2.4. Key takeaway

Through a PRISMA systematic literature review, this study investigates how the introduction of Generative AI has affected the way firms are dealing with environmental sustainability. Results show that the literature is still in its infancy phase, the reason being that this new kind of artificial intelligence has only been brought to the general public in recent years. Therefore, the application fields have been quite restricted so far. Specifically, the LLM-enhanced content analysis allowed us to classify this area of research into three thematic clusters: Cluster 1 delves into the applications of Generative AI for improving the sustainability of firm operations, with a specific focus on logistics and life cycle assessments; Cluster 2 deals with the use of Generative AI for improving corporate reporting and environmental analysis; finally, Cluster 3 focuses on the environmental and ethical impact of the use of GenAI in business and society.

These findings highlight the significant potential for advancing research on the role of Generative AI in enhancing corporate environmental sustainability, not only regarding the business functions impacted but also concerning the models employed. Regarding the former, there is a clear opportunity to explore how Generative AI can be integrated into a broader range of business activities, extending beyond the currently limited scope of logistics, LCA and reporting. This expansion would allow for a more comprehensive understanding of how Generative AI can drive sustainability across various sectors. Concerning the models, future research could benefit from exploring not only the refinement of LLMs but also the potential of other Generative AI models, such as text-to-image models like DALL-E and text-to-video models like SORA. These models offer unique capabilities that could be leveraged to address specific sustainability challenges, such as visualising environmental impacts, enhancing communication strategies around sustainability, or creating more engaging and informative content for stakeholders.

Our findings also have practical implications both for academia and industry. First, the identification of thematic clusters provides a structured framework that researchers can use to explore how Generative AI might enhance the environmental sustainability efforts of firms. Second, we provide a methodological framework that leverages the potential of LLMs as an automated research assistant. NotebookLM is, in fact, an off-the-shelf AI-powered tool that can be used without any prior knowledge of programming that can help researchers identify common patterns across large corpora of texts. Since LLMs are going to affect the research process (*Burger et al., 2023*; *Klarin, 2024*), we show

how these models can be effectively integrated into the researcher's analysis process without affecting the integrity of the results. Third, by highlighting the potential of Generative AI models, the study encourages businesses to experiment with diverse AI technologies, which could lead to innovative solutions for designing, and communicating, sustainability initiatives. Moreover, the study underscores the importance of considering the ethical and environmental impacts of deploying Generative AI in business contexts. This awareness could guide firms in developing responsible AI strategies that not only improve sustainability outcomes but also align with broader societal and ethical expectations.

As with any study, this work is not without limitations. In particular, as highlighted above, Generative AI has only recently been introduced to the general public: as such, much of the academic research may still be unpublished, hence the limited size of our sample. Therefore, a more in-depth study would likely benefit from considering the grey literature, which could potentially complement our understanding of how these models are being implemented in firms' strategies. Moreover, the investigation process itself could be improved in many ways. For example, in our case, the interrogation of NotebookLM required a long process of trial and error before giving consistent and satisfactory results. Further research could, thus, exploit different tools to understand whether similar, or better, results could be obtained.

In conclusion, while this systematic literature provides a foundational understanding, there remains ample space for further investigation and development in this rapidly evolving field.

3. Generative AI to assess EU Taxonomy alignment in Non Financial Disclosures documents

The rapid advancement of GenAI, and particularly LLMs, is revolutionizing the way we process and interpret written text. This has significant implications for various fields, including the analysis of corporate disclosures and the assessment of alignment with regulatory frameworks like the EU Taxonomy (*Föhr et al., 2023*). NFD documents, which companies issue to communicate their efforts and performances in the ESG domain, are often characterized by their length, complexity, and lack of standardization (*Garcia-Torea et al., 2024*). Moreover, these documents can be presented in various formats and under different nomenclatures, such as integrated reports, sustainability reports or CSR reports.

This heterogeneity poses a substantial challenge for stakeholders seeking to extract consistent and comparable information across different companies. Furthermore, a significant portion of the information within these documents is unstructured, consisting of text, images, and tables, which makes it difficult to analyze using traditional data analysis techniques. The EU Taxonomy, a complex classification system designed to define environmentally sustainable economic activities, adds another layer of complexity (*Hummel & Jobst, 2024; O'Reilly et al., 2024*). The Taxonomy's complex technical screening criteria (TSC) specify the conditions that economic activities must meet to be considered sustainable. Matching the information in non-financial disclosure documents with the specific requirements of the EU Taxonomy is a complex task, requiring a deep understanding of both the regulatory framework and the nuances of corporate reporting, as well as a strong technical knowledge of the business under consideration in each case (*Hummel & Bauernhofer, 2024*).

To address these challenges, various techniques can be employed, leveraging the capabilities of AI to automate the analysis and assessment process. This chapter will focus on two techniques that have recently gained momentum in the field of natural language processing (NLP) and have shown promise in handling complex textual data: Retrieval Augmented Generation (RAG) and Long Context processing. RAG enhances the capabilities of LLMs by integrating external knowledge sources, allowing them to generate more accurate and contextually relevant responses (*Lewis et al., 2020*). Long Context models, on the other hand, are designed to process vast amounts of textual data in a single pass, enabling them to capture broader contextual information within lengthy documents *(Liu et al.,* *2023*). These two approaches are employed to evaluate whether an LLM is capable, first, of understanding[4] the core concepts and criteria outlined in the EU Taxonomy, and second, of identifying the presence and implementation of these concepts within NFD documents produced by companies. This dual focus allows for a comprehensive assessment of the LLMs' ability to bridge the gap between regulatory language and corporate reporting practices. The findings of this research are particularly relevant, as they demonstrate that LLMs can be effectively leveraged for this task, offering a valuable tool for a wide range of stakeholders who seek to assess corporate sustainability efforts. Importantly, this can be achieved without requiring deep technical expertise in the specific economic activities defined by the EU Taxonomy, thereby enhancing accessibility and interpretability in the broader context of sustainable finance and corporate accountability. Based on these premises, in this chapter, we will explore the application of RAG and Long Context techniques to evaluate the feasibility of using LLMs for EU Taxonomy alignment assessment, illustrating their strengths, limitations, and potential use cases through the analysis of NFD documents.

3.1. Methodology

Understanding the EU Taxonomy with GenAI

To assess how Gemini can understand the EU Taxonomy, both RAG and Long Context approaches will be employed. To evaluate the comprehension of the EU Taxonomy by LLMs, it is essential to establish a clear and comprehensive set of reference documents that serve as the basis for both the model's understanding and the subsequent human evaluation. Thus, the knowledge base is constituted by a comprehensive collection of official EU Taxonomy documents sourced from the European Parliament website[5] to ensure the accuracy and relevance of the assessment. The documents included in the analysis are:

- *REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*: This foundational regulation establishes the overarching framework and objectives of the EU Taxonomy.
- *COMMISSION DELEGATED REGULATIONS*: A series of delegated regulations, including those dated 4 June 2021 (2021/2139), 6 July 2021 (2021/2178), 9 March 2022 (2022/1214), 27 June 2023

[4] While we acknowledge that LLMs do not "understand" in the human cognitive sense, in this context we refer to their ability to recognize and process semantic patterns associated with the EU Taxonomy. The focus of the analysis lies on the model's capacity to extract meaning from textual data rather than relying on surface-level syntactic patterns.

[5] All documents can be found at the following link: https://finance.ec.europa.eu/sustainable-finance/tools-and-standards/eu-taxonomy-sustainable-activities_en [Accessed: July 2024].

(2023/2485), and 27 June 2023 (2023/2486), were included. These regulations provide detailed technical screening criteria for various economic activities, specifying the conditions under which they can be considered environmentally sustainable. They also outline the disclosure requirements for companies regarding their alignment with the EU Taxonomy.

- *TAXONOMY USER GUIDE*: This guide offers practical guidance on interpreting and applying the EU Taxonomy, clarifying technical terms and concepts.

In the RAG setup, Gemini will be prompted with specific questions about the expected activities described in the EU Taxonomy, and relevant sections from official EU Taxonomy documents, described above, will be retrieved and provided as context. The prompts will be designed to elicit Gemini's understanding of the definitions, technical screening criteria, and the overall structure of the Taxonomy. For the Long Context approach, the same official EU Taxonomy documents will be provided directly within the input context window of Gemini, and similar questions will be posed to evaluate its comprehension of the Taxonomy's requirements. The prompt engineering will involve providing clear and concise instructions, potentially including examples of activities and their corresponding Taxonomy criteria, to ensure Gemini accurately interprets the nuances of the regulatory framework.

Identifying Relevant Sections in Non-Financial Disclosure Documents

Identifying sections within company non-financial disclosure documents that are semantically similar to the activities described in the EU Taxonomy will be achieved using semantic similarity techniques in both approaches.

In the RAG framework, the company disclosure documents will be chunked, and each chunk will also be embedded into the same vector space. The EU Taxonomy activity descriptions will be embedded into the prompt. Semantic similarity search will then be performed to identify the document chunks that have the highest similarity scores with the Taxonomy activity embeddings. Once the relevant sections of the non-financial disclosure documents are identified, Gemini will be used to assess if and how the company is implementing the activities in accordance with the EU Taxonomy. Retrieved document chunks, along with the corresponding EU Taxonomy criteria, will be provided in the prompt, and Gemini will be asked to determine the level of implementation based on the information provided.

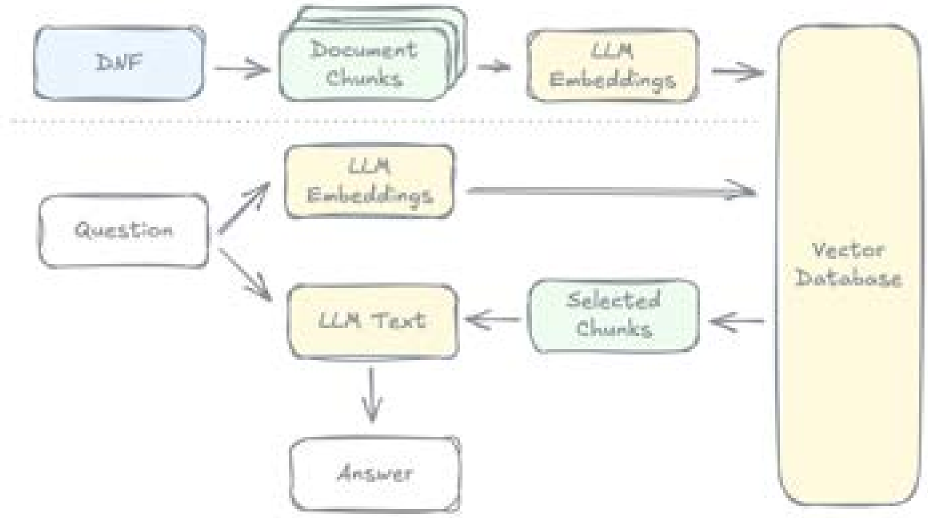

Figure 4: Diagram to represent RAG approach

For the Long Context method, the entire non-financial disclosure document will be fed into Gemini along with the EU Taxonomy activity descriptions. Gemini will be instructed to identify and extrapolate sections of the document that are semantically related to these activities. This might involve prompting Gemini to explain its reasoning for identifying certain sections and the degree of semantic closeness to the Taxonomy descriptions.

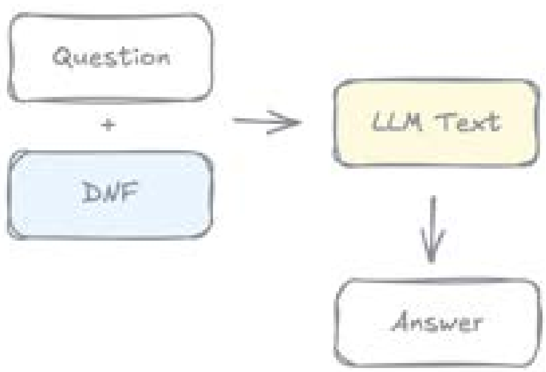

Figure 5: Diagram to represent Long Context approach

Prompt engineering

Prompt engineering, the art and science of crafting effective prompts to elicit desired responses from LLMs, is paramount in guiding these models to accurately interpret complex regulatory frameworks and extract relevant information from unstructured textual data. This section will elaborate on the iterative process of prompt development, the structural template employed, and the key components that contribute to a robust and reliable assessment methodology. The development of effective prompts is a predominantly iterative process (*Sahoo et al., 2024*). Several iterations were conducted, with each iteration building upon the insights gained from the previous one. This iterative approach allowed us to progressively enhance the prompt's clarity, precision, and ability to elicit the desired information from the LLM.

The initial prompts were designed to be relatively simple, focusing on extracting fundamental information such as activity descriptions and technical screening criteria. However, these early prompts often yielded incomplete or ambiguous results, highlighting the need for greater specificity and contextual guidance. Subsequent iterations incorporated more detailed instructions, constraints to steer the LLM towards more accurate and relevant outputs. Furthermore, the prompts were refined to explicitly request the LLM to provide justifications for its assessments, enhancing the transparency of the results.

To ensure consistency and facilitate modularity, the prompts were structured based on a predefined template. This template provided a clear and organized framework for conveying the necessary information to the LLM, enabling easier modification and adaptation for different assessment tasks. The template comprises several key sections, each serving a specific purpose in guiding the LLM's analysis.

The prompt template is designed to provide the LLM with all the necessary context and instructions to perform the EU Taxonomy alignment assessment effectively. The core components of the template are detailed below:

Table 3: Context parameter setting.

Context Setting	
Firm	Specifies the name of the company being assessed. This allows the LLM to focus its analysis on the relevant non-financial disclosure documents and tailor its responses accordingly.
Activity Sector	Indicates the specific economic sector to which the activity belongs (e.g., "Transport"). This helps the LLM to narrow down its search within the EU Taxonomy, as the criteria can vary significantly across sectors.
Activity Number	Provides the unique identification number of the activity within the EU Taxonomy. This ensures precise referencing and avoids any ambiguity in activity identification.

Activity Name	States the official name of the economic activity as defined in the EU Taxonomy (e.g., "Freight rail transport"). This provides a clear and concise label for the activity under consideration.
Activity Purpose	Specifies the environmental objective(s) that the activity is intended to contribute to (e.g., "Climate change mitigation," "Climate change adaptation"). This directs the LLM to focus on the relevant criteria associated with the specified objective(s).
Activity Details	
Activity Description	Presents the detailed description of the economic activity as outlined in the EU Taxonomy. This provides the LLM with the necessary information to understand the scope and boundaries of the activity.
Technical Screening Criteria	Includes the specific technical screening criteria (TSC) that the activity must meet to be considered environmentally sustainable. These criteria are crucial for assessing the company's alignment with the EU Taxonomy.
Do No Significant Harm (DNSH)	Specifies the DNSH criteria that the activity must comply with to ensure that it does not significantly harm any of the other environmental objectives. This section details the DNSH requirements for each of the six environmental objectives: Climate adaptation, Water, Circular economy, Pollution prevention, Biodiversity.
Appendices	
Appendices	The EU Taxonomy documents often include appendices that provide additional context, definitions, or clarifications related to specific activities or criteria. These appendices can be crucial for a comprehensive understanding of the requirements. The prompt template incorporates the relevant appendices when necessary.

Each component of the prompt template plays a vital role in guiding the LLM towards accurate and reliable EU Taxonomy alignment assessments: the context-setting components provide the necessary background information to focus the LLM's analysis; the activity details components define the specific requirements that the company's activities must meet; the appendices provide essential clarifications and contextual information. By combining these components in a structured and comprehensive manner, the prompt template enables the LLM to effectively navigate the complexities of the EU Taxonomy and extract relevant information from company disclosures.

The prompt developed through the iterative process and structured using the template is as follows:

```python
def generate_activity_prompt(company, reference_activityDefinition,
reference_eu_appendix):
    query = f"""You are a bot that is capable of retriveing Environmental, Social and
Governance (ESG) information
```

```
from integrated annual reports provided by a firm.

Your task is to summarize section of the document related to a
given activity from the EU taxonomy for sustainable activities.
You will get as input:
 - Firm name
 - Activity sector
 - Activity Number
 - Activity Name
 - Activity Purpose
 - Activity Description

If an Appendix is mentioned in the definition of the Activity key concepts, consieder
the description of the Appendix provided.

You will provide as output:
Description of what {company} is doing related to the activity
"{reference_activityDefinition['Activity']}"

Input

Firm: {company}
Activity sector: {reference_activityDefinition['Sector']}
Activity Number: {reference_activityDefinition['ActivityNumber']}
Activity Name: {reference_activityDefinition['Activity']}
Activity Purpose {reference_activityDefinition['Climate']}
Activity Description: {reference_activityDefinition['Description']}.
 - Technical screening criteria:
{reference_activityDefinition['SubstantialContributionCriteria']}
 - Do no significant harm (‘DNSH’): {reference_activityDefinition['DNSH on Climate
adaptation']}, {reference_activityDefinition['DNSH on Water']},
{reference_activityDefinition['DNSH on Circular economy']},
```

```
{reference_activityDefinition['DNSH on Pollution prevention']},
{reference_activityDefinition['DNSH on Biodiversity']}
"""
    query = query + ('Appendix A|B|C|D|E: ' +
(reference_eu_appendix[reference_eu_appendix['Appendix']=='Appendix'])['Description']
if query.find('Appendix A|B|C|D|E')) else ""
    query = query + 'Output:'
    return query
```

The challenges in achieving accurate semantic matching will be considered, given the potential for variations in terminology and reporting styles across different companies. Moreover, the inherent generality and high-level abstraction employed in the description of EU Taxonomy-aligned activities present a significant challenge to the efficacy of semantic matching. This characteristic may impede the accurate retrieval of pertinent information from company non-financial disclosures, where specific operational details often diverge from the broad, regulatory definitions. Moreover, the inherent generality and high-level abstraction employed in the description of EU Taxonomy-aligned activities present a significant challenge to the efficacy of semantic matching. This characteristic may impede the accurate retrieval of pertinent information from company non-financial disclosures, where specific operational details often diverge from the broad, regulatory definitions.

Comparison of RAG and Long Context Techniques

It is crucial to establish a reliable benchmark for evaluating the performance of the LLM-driven assessment methodologies. The comparison between the two methodologies, RAG and Long Context, is conducted by comparing the activities identified by each method with a benchmark set of activities. This benchmark is established through a thorough analysis of the same company's non-financial disclosure documents by a human expert possessing in-depth knowledge of ESG topics and the EU Taxonomy. The human expert's analysis serves as the "ground truth," providing a high-quality reference point for assessing the accuracy and effectiveness of the LLM-based approaches (*Mosqueira-Rey et al., 2023*). This comparative evaluation allows for a quantitative assessment of how well each methodology performs in correctly identifying EU Taxonomy-aligned activities.

The comparison of the two techniques will be based on several key metrics. Accuracy will be assessed by comparing Gemini's alignment assessments against a human expert's evaluation of the same documents. Efficiency will be measured in terms of the processing time required by each method and the computational resources utilized. The ability of each method to handle long and complex non-financial disclosure documents will be evaluated, considering factors like the number of documents processed and the length of the documents. Furthermore, the effectiveness of each technique in identifying information that is truly relevant to EU Taxonomy alignment will be analyzed. Discrepancies and similarities in the results obtained by the two methods across these metrics will be highlighted to provide a comprehensive comparison of their strengths and weaknesses in the context of EU Taxonomy compliance assessment.

The results of applying both RAG and Long Context methods to a sample set of company non-financial disclosure documents will be presented in this section. The accuracy of the alignment assessments produced by each method will be compared against a benchmark established by human expert evaluation.

3.2. Empirical Results

To empirically evaluate the efficacy of the proposed LLM-driven methodology for EU Taxonomy alignment assessment, this study focuses on two distinct multinational corporations: Mundys and Maersk. The selection of these firms was purposeful, driven by the desire to capture both similarities and differences in their business activities and reporting practices. Mundys operates in the field of infrastructure management, with a focus on motorways, airports, and other transportation-related assets. The company's activities are primarily centered around the development, maintenance, and operation of transportation networks. Maersk is a global leader in container shipping and logistics. The company's core business revolves around maritime transport, port operations, and integrated supply chain solutions.

While both Mundys and Maersk are significant players in the transportation sector, their specific activities and business models diverge. Mundys focuses on physical infrastructure, whereas Maersk specializes in the movement of goods. This distinction provides an opportunity to assess the LLM's ability to handle different types of disclosures and identify EU Taxonomy alignment across diverse operational contexts. Furthermore, both companies produce comprehensive non-financial disclosure documents, albeit with variations in structure and content, allowing for an evaluation of the methodology's robustness in processing heterogeneous reporting formats.

The empirical analysis within this study will concentrate on activities classified under the "Transport" sector according to the EU Taxonomy. This sector is of particular relevance due to its significant contribution to greenhouse gas emissions and its crucial role in the transition to a sustainable economy (*European Environment Agency, 2024; Schütze & Stede, 2024*). Focusing on a single sector allows for a more in-depth analysis and facilitates the identification of specific challenges and opportunities in applying LLM-driven methodologies to EU Taxonomy assessment.

It is important to note that an initial approach considering the use of NACE codes (Nomenclature statistique des activités économiques dans la Communauté européenne) to define the scope of the analysis was deemed inappropriate. NACE codes, while providing a standardized classification of economic activities, cannot be directly and unambiguously associated with each business activity relevant to the EU Taxonomy due to their inherent structural rigidity and limited flexibility in capturing nuanced or cross-cutting business practices (*Hoberg & Phillips, 2016*). The EU Taxonomy often defines activities based on specific technical criteria and environmental objectives, which may cut across or fall outside the boundaries of traditional NACE code classifications. Therefore, a sector-based approach, focusing on the "Transport" sector, was chosen to ensure a more accurate and relevant assessment of EU Taxonomy alignment.

Accuracy of the alignment assessments produced by each method was compared against expected activities mentioned in the non-financial disclosure documents and extrapolated by a human expert evaluation. For the purpose of this analysis, accuracy is defined as the binary outcome of whether an activity is identified by the LLM model. Given the promising results obtained, further research is suggested to strengthen these findings through an expanded empirical evaluation encompassing a larger and more diverse dataset of companies across various sectors.

In Tables 4 and 5, the alignment assessment results for Mundys and Maersk are reported.

Table 4: Maersk, summary of RAG Vs Long Context results

Maersk			
Activity #	**Description**	**RAG**	**Long Context**
6.00	Construction, modernisation, maintenance and operation of motorways, streets, roads, other vehicular and pedestrian ways, surface work on streets, roads, highways, bridges or tunnels and construction of airfield runways, including the provision of architectural services, engineering services, drafting services, building inspection services and surveying and mapping services and the like as well as the performance of physical, chemical and other analytical testing of all types of materials and products, and excludes the installation of street lighting and electrical signals. The economic activities in this category could be classified under several NACE codes, in particular F42.11, F42.13, M71.12 and M71.20 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	TRUE	TRUE
6.02	Purchase, financing, leasing, rental and operation of freight transport on mainline rail networks as well as short line freight railroads. The economic activities in this category could be associated with several NACE codes, in particular H49.20 and N77.39 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	FALSE	TRUE
6.06	Purchase, financing, leasing, rental and operation of vehicles designated as category N1, N2(269)As referred to in Article 4(1), point (b)(ii), of Regulation (EU) 2018/858. or N3(270)As referred to in Article 4(1), point (b)(iii), of Regulation (EU) 2018/858. falling under the scope of EURO VI(271)As set out in Regulation (EC) No 595/2009., step E or its successor, for freight transport services by road. The economic activities in this category could be associated with several NACE codes, in particular H49.4.1, H53.10, H53.20 and N77.12 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	TRUE	TRUE
6.10	Purchase, financing, chartering (with or without crew) and operation of vessels designed and equipped for transport of freight or for the combined transport of freight and passengers on sea or coastal waters, whether scheduled or not. Purchase, financing, renting and operation of vessels required for port operations and auxiliary activities, such as tugboats, mooring vessels, pilot vessels, salvage vessels and ice-breakers. The economic activities in this category could be associated with several NACE codes, in particular H50.2, H52.22 and N77.34 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	FALSE	TRUE
6.12	Retrofit and upgrade of vessels designed and equipped for the transport of freight or passengers on sea or coastal waters, and of vessels required for port operations and auxiliary activities, such as tugboats, mooring vessels, pilot vessels, salvage vessels and ice-breakers. The economic activities in this category could be associated with NACE codes H50.10, H50.2, H52.22, C33.15, N77.21 and N.77.34 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	TRUE	TRUE
6.16	Construction, modernisation, operation and maintenance of infrastructure that is required for zero tailpipe CO2 operation of vessels or the port's own operations, as well as infrastructure dedicated to transhipment and modal shift and service facilities, safety and traffic management systems. The economic activities in this category excludes dredging of waterways. The economic activities in this category could be associated with several NACE codes, in particular F42.91, M71.12 and M71.20 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	FALSE	TRUE

Table 5: Mundys*, summary of RAG Vs Long Context results*

Mundys			
Activity #	**Description**	**RAG**	**Long Context**
6.03	Purchase, financing, leasing, rental and operation of urban and suburban transport vehicles for passengers and road passenger transport. For motor vehicles, operation of vehicles designated as category M2 or M3, in accordance with Article 4(1) of Regulation (EU) 2018/858, for the provision of passenger transport. The economic activities in this category may include operation of different modes of land transport, such as by motor bus, tram, streetcar, trolley bus, underground and elevated railways. This also includes town-to-airport or town-to-station lines and operation of funicular railways and aerial cableways which are part of urban or suburban transit systems. The economic activities in this category also include scheduled long-distance bus services, charters, excursions and other occasional coach services, airport shuttles (including within airports), operation of school buses and buses for the transport. The economic activities in this category could be associated with several NACE codes, in particular H49.31, H49.3.9, N77.39 and N77.11 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	FALSE	TRUE
6.14	Construction, modernisation, operation and maintenance of railways and subways as well as bridges and tunnels, stations, terminals, rail service facilities(326)In accordance with Article 3, point (11), of Directive 34/2012/EU of the European Parliament and of the Council of 21 November 2012 establishing a single European railway area (OJ L 343, 14.12.2012, p. 32)., safety and traffic management systems including the provision of architectural services, engineering services, drafting services, building inspection services and surveying and mapping services and the like as well as the performance of physical, chemical and other analytical testing of all types of materials and products. Manufacture, installation, technical consulting, retrofitting, upgrade, repair, maintenance, repurposing of products, equipment, systems and software related to one of the following elements: assembled railway track fixtures; rail constituents detailed in Points 2.2 to 2.6 to Annex II of Directive (EU) 2016/797. The economic activities in this category could be associated with several NACE codes, in particular C25.99, C27.9, C30.20, F42.12, F42.13, M71.12, M71.20, F43.21, and H52.21 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	TRUE	FALSE
6.15	Construction, modernisation, maintenance and operation of infrastructure that is required for zero tailpipe CO2 operation of zero-emissions road transport, as well as infrastructure dedicated to transshipment, and infrastructure required for operating urban transport. The economic activities in this category could be associated with several NACE codes, in particular F42.11, F42.13, M71.12 and M71.20 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	TRUE	TRUE
6.17	Construction, modernisation, maintenance and operation of infrastructure that is required for zero tailpipe CO2 operation of aircraft or the airport's own operations, and for provision of fixed electrical ground power and preconditioned air to stationary aircraft as well as infrastructure dedicated to transhipment with rail and water transport. The economic activities in this category could be associated with several NACE codes, in particular F41.20 and F42.99 in accordance with the statistical classification of economic activities established by Regulation (EC) No 1893/2006.	FALSE	TRUE

Accuracy, latency, and average cost of information retrieval for each EU taxonomy activity are compared against baseline results. Latency refers to the average time the model takes to retrieve information for each activity, while cost represents the financial expenditure required for each information retrieval operation. These three factors—accuracy, latency, and cost—are paramount in identifying the optimal balance for analytical tool selection. The interplay of these elements can significantly influence the choice of a specific methodology. For instance, in an initial assessment phase, a RAG approach might be favored to minimize both time and cost, with a more in-depth analysis pursued only if the preliminary results warrant further investigation.

The baseline is established using the same Gemini LLM, but without any contextual documents, whether through a RAG or long-context method. The results for each company, detailed in Tables 4 and 5, are subsequently consolidated and summarized in Table 6. We can summarize results in the following table taking into account average latency and cost of the operation to retrieve information for each EU taxonomy activity:

Table 6: RAG and Long Context summary results compared to baseline.

Model	Accuracy	Latency	Cost
Baseline	40 % (4/10)	0.2 min	$ 0.01
RAG	60% (6/10)	0.2 min	$ 0.30
Long Context	90% (9/10)	2.3 min	$ 1.015

The RAG model achieves 60% accuracy, representing a 50% improvement over the baseline model's 40% accuracy, while maintaining the same latency of 0.2 minutes. However, this enhanced accuracy comes at a significantly higher cost, with the RAG model incurring a cost of $0.30, which is 30 times greater than the baseline's cost of $0.01. The Long Context model demonstrates the highest accuracy at 90%, a 125% improvement over the baseline. This superior accuracy is accompanied by a substantial increase in latency, reaching 2.3 minutes, more than 11 times that of the baseline and RAG models. Furthermore, the Long Context model is the most expensive, with a cost of $1.015, over 100 times higher than the baseline and approximately 3.4 times higher than the RAG model.

Strengths of the Analysis

The application of Gemini with RAG for EU Taxonomy assessment presents several potential strengths. The RAG approach can enhance accuracy by retrieving and grounding the assessment in specific external knowledge from the official EU Taxonomy documentation passed in the prompt. This ability to reference precise regulatory text can reduce the likelihood of the AI model misinterpreting the requirements. Furthermore, RAG offers enhanced transparency, as the retrieved documents that informed the assessment can be identified and cited, providing a clear audit trail of the reasoning process. This modularity allows for easier updates to the knowledge base as the EU Taxonomy evolves, ensuring the assessment remains current in a dynamic regulatory environment. For instance, when assessing alignment with specific technical screening criteria that require up-to-date thresholds or methodologies, RAG can be configured to retrieve the latest versions of the relevant delegated acts, ensuring the assessment is based on the most current information.

Conversely, utilizing Gemini with a Long Context window for EU Taxonomy assessment also offers distinct advantages. The ability to process entire non-financial disclosure documents at once allows the model to capture broader contextual information that might be missed by a chunk-based RAG approach. This can be particularly beneficial for understanding the overall narrative and interconnectedness of a company's sustainability efforts as presented in their report. Moreover, implementing a Long Context approach can be less complex compared to setting up and managing a full RAG pipeline with its retrieval and indexing components. The Long Context method also holds the potential to identify cross-document relationships within the non-financial disclosure, such as when different sections of the report refer to the same underlying activity or initiative, providing a more holistic view of the company's alignment efforts.

Weaknesses of the Analysis

Despite its strengths, using Gemini with RAG for EU Taxonomy assessment also presents certain weaknesses. The implementation of a RAG system can be complex, requiring careful design of the retrieval mechanism, the knowledge base, and the prompts used for generation. The accuracy of the assessment is heavily reliant on the quality and relevance of the information retrieved from the external knowledge base. If the retrieval process fails to identify the most pertinent documents or if the knowledge base is incomplete or outdated, the resulting assessment may be inaccurate or incomplete. Furthermore, while RAG aims to address the context window limitations of LLMs, these limitations might still pose challenges if the query requires retrieving and processing a large number of documents to adequately address the EU Taxonomy criteria. The specific structure and technical language of the EU Taxonomy can also make accurate retrieval challenging, as semantic search algorithms might struggle to match the precise terminology used in the regulation with the potentially more general language used in company reports.

Similarly, employing Gemini with a Long Context window for EU Taxonomy assessment also has its limitations. Processing very long non-financial disclosure documents can be computationally expensive, demanding significant computational resources. The potential for information overload and the "lost in the middle" phenomenon could affect the accuracy of information extraction, as the model might struggle to retain and process relevant details dispersed throughout a lengthy document. Pinpointing the exact source of information within a long document can be difficult, potentially hindering the transparency and verifiability of the assessment results. Additionally, the model might be distracted by irrelevant information present within the extensive context, leading to less focused and potentially less accurate assessments. The length and complexity of typical non-financial disclosure documents could exacerbate the "lost in the middle" problem, making it harder for the model to effectively utilize information from all parts of the report.

One important limitation of this analysis is the relatively narrow scope of the empirical evaluation, which focuses on only two companies (Mundys and Maersk) within a single sector (transportation). While this focused approach allowed for in-depth analysis and the identification of specific challenges and opportunities within the sector, it may affect the generalizability of the results to other industries or to a broader range of companies. Different sectors may have distinct reporting practices, regulatory nuances, and types of activities relevant to the EU Taxonomy, which could influence the performance and applicability of LLM-driven assessment methodologies. Therefore, the findings presented here should be interpreted within the context of the selected companies and sector, and further research across a wider array of industries and company sizes is needed to validate and extend these conclusions. However, we are confident that the same approach will work on other companies and sectors, thanks to the generalist nature of the Gemini model.

Comparison of Strengths and Weaknesses

Both RAG and Long Context approaches offer unique advantages and disadvantages for assessing EU Taxonomy alignment using Gemini. RAG excels in providing accuracy and transparency by grounding its analysis in external, verifiable knowledge from the Taxonomy documentation. Its modularity also allows for easier updates (*Lewis et al., 2020*). However, it introduces complexity in implementation and relies heavily on the quality of the retrieval process (*Gao et al., 2023*). Long Context models, on the other hand, offer simplicity by processing entire documents and potentially capturing broader context. However, they face challenges related to computational cost, information overload, and the "lost in the middle" phenomenon (*Liu et al., 2023*). The choice between these approaches might depend on the specific requirements of the assessment. For tasks requiring high accuracy and traceability, especially when dealing with specific technical criteria, RAG might be preferred. Conversely, for gaining a holistic understanding of a company's overall sustainability narrative and identifying cross-document relationships, Long Context models could be more suitable, provided the computational resources are available and strategies to mitigate information overload are employed.

3.3. Key takeaway

We explored the feasibility of using the Gemini 1.5 model (*Georgiev et al., 2024*) with two distinct techniques, RAG and Long Context processing, to assess the alignment of company non-financial disclosure documents with the EU Taxonomy for sustainable activities. This preliminary analysis suggests that both approaches hold potential for automating and enhancing the process of compliance assessment. RAG offers the strength of grounding its analysis in the official EU Taxonomy documentation, potentially leading to more accurate and transparent evaluations. Long Context models provide the advantage of processing entire documents, allowing for a more holistic understanding of a company's sustainability efforts. However, both methods also present challenges. RAG requires careful implementation and is susceptible to the quality of the retrieval process, while Long Context models can be computationally expensive and might suffer from information overload and the "lost in the middle" problem.

Based on these findings, several potential next steps can be identified for further research. The research could benefit from a more in-depth empirical evaluation of both techniques using a larger and more diverse set of company NFD documents across various sectors. Comparing the performance of these methods against a more extensive group of human expert assessments would provide valuable insights into their reliability and accuracy in real-world scenarios. Future research could also explore the application of these techniques to specific sectors or environmental objectives within the EU Taxonomy, allowing for a more granular understanding of their strengths and weaknesses in different contexts. Refining the evaluation metrics and methodologies used to assess the accuracy and reliability of AI-driven compliance assessments is also crucial. Investigating new techniques and methodologies, such as Cache-Augmented Generation (CAG) could also be a promising avenue (*Chan et al., 2024*). CAG aims to leverage on the advantages of Long Context and RAG. Finally, addressing potential biases inherent in using AI for regulatory compliance is essential to ensure fairness and accountability in these automated processes. Exploring the potential of using other GenAI models besides Gemini could also provide a broader perspective on the applicability of these techniques. Investigating methods for improving the transparency and explainability of AI-driven assessments is critical for building trust and facilitating human oversight in regulatory compliance.

Beyond these technical insights, the results of this study also carry important implications for both policymakers and corporate practitioners. From a policy perspective, the demonstrated potential of LLM-based tools in assessing taxonomy alignment suggests a promising path toward more scalable and consistent regulatory monitoring, especially in light of the growing volume and complexity of sustainability reporting (*Hummel & Jobst, 2024*). These techniques could support regulators in identifying gaps or intersections in disclosure and prioritizing areas for further scrutiny or guidance.

From a managerial standpoint, the automation of alignment assessments through GenAI models could significantly reduce the resource burden associated with sustainability compliance, particularly for SMEs that often lack the internal expertise to interpret and implement complex regulations like the EU Taxonomy (*Hummel & Bauernhofer, 2024; O'Reilly et al., 2024*). By offering accessible and interpretable assessments, these tools may also help firms better understand where they stand in their sustainability journey and benchmark their performance against competitors or industry leaders.

More broadly, integrating these approaches into ESG auditing and reporting frameworks could contribute to reducing information asymmetries between firms and stakeholders, thereby fostering more transparent, accountable, and comparable sustainability disclosures across the European market.

Looking ahead, the continuous evolution of GenAI models and tools necessitates a robust methodology for verifying their reliability and trustworthiness. It is crucial to establish a test framework based on a "golden dataset" of companies and their human-identified activities. This framework would allow for the systematic testing of new model versions and tools before their adoption, ensuring their accuracy and consistent performance in assessing sustainability disclosures.

Specialized frameworks simplify RAG evaluation by providing pre-built metrics and a structured way to test your system. They often use an LLM-as-a-judge approach, where a powerful LLM evaluates the output of your RAG system based on specific criteria. RAGAS (RAG Assessment) is an open-source framework specifically designed for RAG evaluation (Es et al., 2024). It quantifies the quality of generated responses by using LLM-generated synthetic data and key metrics. RAGAS is particularly useful for assessing faithfulness (i.e., *are the facts in the response supported by the retrieved context?*) and answer relevance (i.e., *is the answer relevant to the original question?*). DeepEval is a library that offers a variety of metrics for RAG evaluation, including faithfulness, answer relevance, and context relevance. It also allows for the creation of custom metrics, which is useful for highly specific use cases. TruLens is an open-source toolkit for LLM and RAG evaluation. It provides a way to track the entire RAG pipeline and offers a range of metrics to measure the quality of both retrieval and generation. Evidently AI is an open-source library that helps evaluate and monitor RAG systems, with or without code. It supports both reference-based and reference-free evaluation and can be used for stress testing and adversarial testing. LlamaIndex and LangChain are not solely evaluation frameworks, but they include built-in evaluation modules that allow you to test and iterate on different components of your RAG pipeline.

4. Generative AI to support ESG researchers: a case study

The preceding chapters have established the growing significance of non-financial disclosures (NFDs) in corporate sustainability, alongside the inherent complexities and challenges associated with their manual processing and analysis. Critically, in Chapter 3, we demonstrated the robust capability of an off-the-shelf Generative AI model, specifically Gemini 1.5 Pro, to effectively understand and extrapolate information pertaining to EU Taxonomy-aligned activities from complex corporate documents. This was achieved by providing the model with the appropriate context and through meticulous prompt engineering, highlighting the potential for these advanced Large Language Models (LLMs) to overcome the traditional limitations of unstructured data analysis.

Building upon this foundational demonstration of LLM proficiency, this chapter now transitions from the assessment of taxonomy-aligned activities to the practical application of GenAI in supporting ESG researchers and corporate employees. We posit that the inherent strengths of LLMs – their ability to automate data extraction, identify nuanced trends and patterns, and generate valuable insights from vast amounts of textual data – can fundamentally reshape how non-financial disclosures are both analyzed and potentially produced.

This chapter will present a detailed case study illustrating how researchers, or indeed company employees tasked with sustainability reporting, can leverage LLM-based tools to efficiently document existing ESG activities or extrapolate EU Taxonomy-related information. The case study will focus on two prominent Italian firms operating within the transport infrastructure sector: Mundys S.p.A. and Ferrovie dello Stato Italiane (FSI). By examining their non-financial disclosure documents, we aim to provide concrete examples of how GenAI can streamline the often labor-intensive process of ESG data management, enhance the quality and comparability of sustainability information, and ultimately foster more informed decision-making for both internal stakeholders and external investors.

We employed a paradigm-driven case study, which is particularly suited to exploring novel perspectives on previously underexplored phenomena (*Dyer & Wilkins, 1991*). The case study focuses on the corporate sustainability reports of two major Italian companies operating in the transport infrastructure sector, i.e. *Mundys S.p.A.*[6] (Mundys, hereafter) and Ferrovie dello Stato Italiane (FSI,

[6] Mundys S.p.A. (formerly Atlantia S.p.A.) is a global operator in the transport infrastructure sector, with activities spanning motorways, airports, and intelligent transport systems. Ferrovie dello Stato Italiane is Italy's state-owned railway holding company, operating in passenger and freight transport as well as infrastructure management.

hereafter). This sector was selected due to the fact that transportation is responsible for approximately one-tenth of total greenhouse gas emissions in the European Union, making it one of the most environmentally impactful industries (*Schütze & Stede, 2024*). Moreover, transport remains the only major sector in the EU where emissions were still rising until the Covid-19 pandemic and are still above their 1990 levels (*European Environment Agency, 2024*). Moreover, Mundys and FSI were selected for their strategic centrality and national relevance, as both companies operate extensively across the entire Italian territory. In addition to their geographic and infrastructural significance, they are also highly relevant from an environmental impact perspective. As such, the EU Taxonomy is particularly pertinent to their operations and sustainability reporting, making them especially suitable for analyzing the application of sustainable finance regulations in the transport infrastructure sector.

We leverage two major tools:

- *Gemini Pro 1.5*, a cutting-edge language model developed by Google AI, known for its impressive capabilities in natural language understanding, generation, and reasoning (*Georgiev et al., 2024*). This model can be used to analyze textual data from corporate sustainability reports, identify specific activities and initiatives, and evaluate their alignment with the criteria defined in the EU taxonomy. Furthermore, Gemini Pro 1.5 can generate summaries and insights, thereby providing valuable assistance in understanding complex sustainability reports and identifying key areas for further investigation. Its ability to discern nuances in language and context can help to accurately assess the degree to which companies are adhering to the EU taxonomy and identify potential areas for improvement.
- *NotebookLM* (with Gemini 1.5 Pro), an AI-powered note-taking application developed by Google, designed to streamline research and learning processes. Integrated with the Gemini 1.5 Pro model, NotebookLM can summarize complex information, generate insights, and assist in understanding technical jargon or nuanced concepts. By utilizing this tool, analysts can efficiently extract relevant information from corporate sustainability reports and identify key activities that are consistent with the EU taxonomy. Additionally, NotebookLM can help in analyzing and summarizing large volumes of text data, thereby facilitating a comprehensive evaluation of corporate sustainability practices.

Drawing upon the results of the previous chapter, NotebookLM was provided with the latest available version of the EU Taxonomy at the time of the analysis to serve as the knowledge base. The LLM was then prompted to perform three distinct analytical tasks related to the assessment of NFDs. First, for each identified Key Concept (as defined in the previous chapter) the LLM generated a set of five potential examples illustrating how a company might implement the concept within a specific sector, activity, and climate impact category. These examples (150, approximately) were subsequently consolidated into a refined list of the 20 most relevant and representative activities. Second, the model was applied to the NFD of the company selected for the use-case, i.e. Mundys, to determine which of the refined activities had actually been implemented. Finally, a comparative analysis was conducted to identify possible points of convergence between the information disclosed by Mundys and that provided by FSI, which was used as a benchmark case.

The adoption of NotebookLM in this study carries two key methodological implications. First, NotebookLM combines a RAG approach with a long-context approach, allowing the tool to leverage the perks of both strategies simultaneously. This hybrid configuration extends the LLM's access to domain-specific information beyond its pre-trained knowledge, thereby reducing the risk of hallucinations and improving the relevance and accuracy of generated outputs (*Li et al., 2022*). Moreover, each step of our analysis builds upon the outcomes of the previous phase, mimicking a chain-of-thought approach, which is increasingly demonstrating strong potential in the analysis of corporate reports (*Föhr et al., 2023; Gu et al., 2024*). Second, NotebookLM does not require any prior programming skills or technical understanding of LLM functioning, and it operates through a user-friendly graphical interface comparable to that of ChatGPT. This feature makes it accessible to a wide range of users, regardless of their technical background. Consequently, the method adopted here reflects a plausible real-world use case in which an external analyst relies on LLM tools to analyze NFDs, while also ensuring the transparency and replicability of the proposed approach.

4.1. Activity example Generation and consolidation

The EU Taxonomy, while representing a critical regulatory framework to guide the transition toward a low-carbon economy, is necessarily technical in nature. This level of technicality ensures precision and consistency in defining sustainable economic activities. However, it also presents challenges for stakeholders who lack specialized training in sustainability measurement and reporting, potentially making it difficult to interpret how specific business activities qualify as eligible, or aligned, under the Taxonomy (*Cho et al., 2020; O'Reilly et al., 2024*). For example, the activity "Freight rail transport" is briefly defined and includes a key concept such as: "Measures are in place to manage waste in accordance with the waste hierarchy, in particular during maintenance." Yet, it does not offer concrete illustrations of how companies might operationalize such requirements in practice. To address this gap, in the first phase of the analysis, we leveraged Gemini 1.5 to generate practical implementation examples that could make the Taxonomy more accessible and actionable. Specifically, Gemini 1.5 was prompted with the key concept associated with each activity as defined in the EU Taxonomy, and asked to produce five diverse examples per activity. The base prompt was as follows:

Python
```python
"As per the EU Taxonomy, for the Transport sector, to have '$IMPACT_TYPE' on the '$CLIMATE_MITIGATION' for the '$ACTIVITY' a company has to implement the key concept below. List of 5 activities a EU company could implement.

Key Concept: $KEY_CONCEPT
Concise list of activities: "
```

Then, each prompt was customized based on the specific sector, activity, and climate impact category. An example of how a prompt was provided to the LLM is the following:

None
```
As per the EU Taxonomy, for the Transport sector, to have 'Do not significant harm Circular economy' on the 'Climate Adaptation' for the 'Freight rail transport' a company has to implement the key concept below. List of 5 activities a EU company could implement.

Key Concept: Measures are in place to manage waste in accordance with the waste hierarchy, in particular during maintenance.
Concise list of activities:
```

At this point, the LLM provided us with the following output:

```None
1. Prioritize waste prevention and reduction during maintenance activities.
2. Implement robust recycling and reuse programs for maintenance materials.
3. Partner with specialized waste management companies for proper disposal.
4. Train staff on waste management procedures and the waste hierarchy.
5. Regularly monitor and report on waste generation and management efforts.
```

This process, repeated across all 30 Transport activities, yielded a comprehensive set of 150 implementation examples.

To refine this further, we again utilized Gemini, this time for its ability to analyze and consolidate information. We tasked Gemini with identifying the 20 most significant examples from the generated list. Below the prompt used for the task:

```Python
"From the following separate list, can you make a comprehensive summary list of the 20 most important one?

Original lists:
1. Prioritize waste prevention and reduction during maintenance activities.
2. Implement robust recycling and reuse programs for maintenance materials.
3. ..."
```

The output provided by Gemini was the following:

```Python
"Here's a comprehensive summary of the 20 most important activities from the provided lists, aiming to encompass a broad range of climate adaptation and mitigation actions across different transport sectors:
```

Infrastructure Adaptation & Resilience

- Elevate and reinforce critical infrastructure in flood-prone areas.
- Strengthen infrastructure against extreme winds and other weather events.
- Adapt roads and infrastructure for climate resilience (heat-resistant materials, drainage, etc.).
- Retrofit existing infrastructure for extreme weather resilience.
- Implement nature-based solutions (green roofs, bioswales, vegetation) for climate adaptation and mitigation benefits.
- Develop and implement Climate Risk Management and Adaptation Plans.
- Monitor and maintain infrastructure based on performance indicators.

Sustainable Transport & Operations

- Transition to electric or hydrogen-powered vehicles for passenger and freight transport.
- Invest in and utilize rolling stock (trains, railcars, vessels) designed for non-fossil fuel goods.
- Promote intermodal transport solutions for efficient and low-carbon freight movement.
- Optimize routes and schedules to minimize fuel consumption and emissions.
- Implement eco-driving techniques and provide training to drivers.

Emissions Reduction & Energy Efficiency

- Retrofit or replace existing engines to meet or exceed emission limits.
- Purchase new vehicles and equipment compliant with the latest emission standards.
- Regularly maintain and inspect engines to ensure optimal performance and emissions control.
- Explore and adopt alternative, cleaner propulsion technologies (e.g., electric or hydrogen).
- Utilize renewable energy sources to power transport infrastructure and operations.

Waste Management & Circular Economy

- Prioritize waste prevention and reduction during maintenance and operations.
- Implement robust recycling and reuse programs for various materials and components.

```
- Partner with specialized waste management companies for proper disposal and
treatment.

This list prioritizes actions that directly address climate risks, reduce greenhouse
gas emissions, and promote resource efficiency in the transport sector. It considers
both physical and non-physical solutions, emphasizes nature-based approaches where
feasible, and highlights the importance of planning, monitoring, and collaboration."
```

This generated list served as the basis for the following step in which the actual NFD of Mundys is analyzed.

4.2. Company-Specific Activities Identification

In the second step of the analysis, the objective is to assess the potential of our approach in analyzing a real-world scenario. To this end, we examine Mundys' 2022 NFD document in order to identify which of the 20 example activities, generated in the first step, have been implemented by the company. More specifically, we tasked the LLM with assigning relevance scores to each of the identified activities and providing justifications based on the textual evidence extracted from the document. This step illustrates how an external analyst could use LLMs to evaluate the extent to which a company is implementing the principles outlined in the EU Taxonomy[7].

Consequently, we provided NotebookLM with the following prompt:

```python
"Does Mundys implement the following activities?


Infrastructure Adaptation & Resilience
- Elevate and reinforce critical infrastructure in flood-prone areas.
- Strengthen infrastructure against extreme winds and other weather events.
```

[7] We note how the primary objective of this research is to investigate the potential of LLMs in extracting relevant information from corporate reports and not to audit the completeness of the NFD of the focal firm analysed. Therefore, our analysis does not carry any implications for the quality of the disclosure itself.

- Adapt roads and infrastructure for climate resilience (heat-resistant materials, drainage, etc.).
- Retrofit existing infrastructure for extreme weather resilience.
- Implement nature-based solutions (green roofs, bioswales, vegetation) for climate adaptation and mitigation benefits.
- Develop and implement Climate Risk Management and Adaptation Plans.
- Monitor and maintain infrastructure based on performance indicators.

Sustainable Transport & Operations
- Transition to electric or hydrogen-powered vehicles for passenger and freight transport.
- Invest in and utilize rolling stock (trains, railcars, vessels) designed for non-fossil fuel goods.
- Promote intermodal transport solutions for efficient and low-carbon freight movement.
- Optimize routes and schedules to minimize fuel consumption and emissions.
- Implement eco-driving techniques and provide training to drivers.

Emissions Reduction & Energy Efficiency
- Retrofit or replace existing engines to meet or exceed emission limits.
- Purchase new vehicles and equipment compliant with the latest emission standards.
- Regularly maintain and inspect engines to ensure optimal performance and emissions control.
- Explore and adopt alternative, cleaner propulsion technologies (e.g., electric or hydrogen).
- Utilize renewable energy sources to power transport infrastructure and operations.

Waste Management & Circular Economy
- Prioritize waste prevention and reduction during maintenance and operations.
- Implement robust recycling and reuse programs for various materials and components.
- Partner with specialized waste management companies for proper disposal and treatment."

As a response to the above input, NotebookLM provided the following output table:

Table 7: EU Taxonomy activities implemented by Mundys

Activity	**Implementation by Mundys**	**Evidence from the Text**
	Infrastructure Adaptation & Resilience	
Elevate and reinforce critical infrastructure in flood-prone areas.	Likely	The text mentions Mundys' focus on managing extensive motorway infrastructure, and flood resilience is crucial for such infrastructure.
Strengthen infrastructure against extreme winds and other weather events.	Likely	The text states Mundys likely incorporates designs and retrofits to enhance resilience against various weather-related risks.
Adapt roads and infrastructure for climate resilience (heat-resistant materials, drainage, etc.).	Highly Likely	The text suggests that using climate-resilient materials and design features is becoming standard practice, especially for new construction and major renovations undertaken by Mundys.
Retrofit existing infrastructure for extreme weather resilience.	Likely	The text indicates that retrofitting existing infrastructure is likely a key part of Mundys' climate adaptation strategy, given the age of some Italian infrastructure.
Implement nature-based solutions (green roofs, bioswales, vegetation) for climate adaptation and mitigation benefits.	Partially	Mundys' sustainability reports mention efforts to increase green spaces and biodiversity along their motorways, suggesting some implementation of nature-based solutions. The extent and specific examples may vary.
Develop and implement Climate Risk Management and Adaptation Plans.	Yes	Mundys has a publicly available "Climate Action Plan" which strongly indicates they are actively engaged in assessing and managing climate risks.
Monitor and maintain infrastructure based on performance indicators.	Likely	The text suggests Mundys likely has systems in place to monitor infrastructure performance and condition, including assessing climate-related impacts.
	Sustainable Transport & Operations	
Transition to electric or hydrogen-powered vehicles for passenger and freight transport.	Partially	Mundys is actively developing electric vehicle charging infrastructure, but their involvement in hydrogen-powered vehicles is less clear.

Activity	Implementation by Mundys	Evidence from the Text
Invest in and utilize rolling stock (trains, railcars, vessels) designed for non-fossil fuel goods.	Not Applicable	Mundys primarily manages road infrastructure and airports, not rail or maritime transport.
Promote intermodal transport solutions for efficient and low-carbon freight movement.	Possibly, indirectly	Mundys might support intermodal transport through infrastructure planning and development, facilitating connections between different modes of transport.
Optimize routes and schedules to minimize fuel consumption and emissions.	Partially	This is relevant to their airport operations where efficient ground handling and aircraft movements can reduce emissions, but less directly applicable to motorway operations.
Implement eco-driving techniques and provide training to drivers.	Not directly applicable	Mundys manages infrastructure, not vehicle fleets, but they may promote eco-driving through campaigns or collaborations.
Emissions Reduction & Energy Efficiency		
Retrofit or replace existing engines to meet or exceed emission limits.	Not directly applicable	Mundys doesn't own or operate vehicle fleets.
Purchase new vehicles and equipment compliant with the latest emission standards.	Likely	This applies to vehicles and equipment used in their own operations, and they would likely prioritize low-emission options.
Regularly maintain and inspect engines to ensure optimal performance and emissions control.	Likely	Applicable to their own fleet of vehicles and equipment.
Explore and adopt alternative, cleaner propulsion technologies (e.g., electric or hydrogen).	Likely	Relevant to their own fleet and potentially supported through infrastructure development (e.g., charging stations).
Utilize renewable energy sources to power transport infrastructure and operations.	Yes	Mundys is committed to increasing the use of renewable energy, including solar panels on service areas.
Waste Management & Circular Economy		
Prioritize waste prevention and reduction during maintenance and operations.	Likely	Mundys likely has waste management programs in place.
Implement robust recycling and reuse programs for various materials and components.	Likely	Recycling and reuse are likely part of their waste management strategy.

Activity	Implementation by Mundys	Evidence from the Text
Partner with specialized waste management companies for proper disposal and treatment.	Likely	Probable, especially for hazardous or specialized waste.

4.3. Cross company comparison

To further illustrate the potential of LLMs in analyzing corporate sustainability efforts, we performed a cross-company comparison between Mundys and FSI. To do so, we add to the knowledge base of NotebookLM the 2022 NFD document of FSI and ask the tool for a comparison of how the EU Taxonomy indications were implemented by the two companies. By tasking the LLM to identify activities implemented by FSI but not by Mundys, we aimed to uncover potential opportunities for Mundys to adopt and adapt successful sustainability practices from another industry leader. While the companies examined in this case study are large corporations that arguably possess the internal resources necessary to develop their own taxonomy-aligned strategies without external reference, this use case becomes particularly relevant for SMEs. SMEs, in fact, often lack the internal capacity or technical expertise to independently design and implement taxonomy-compliant practices (*Kinderman, 2020*). In this sense, our exercise reflects how an SME, pressured by the demands of larger firms within its value chain, might leverage LLM tools to understand and replicate sustainability practices. This becomes particularly relevant in light of the findings by *O'Reilly et al. (2024)*, who highlight the role of large corporations in accelerating the low-carbon transition by exerting upward and downward influence across their networks.

In this third step, the prompt provided to NotebookLM was as follows:

```python
"In the context of the EU taxonomy, make a table of the list of activities below. For each activity list if Mundys and Feerovie dello stato implemented or not with a short description.

Given the analisys of the table, create a new table suggesting activities that Ferrovie dello stato is implementing and you would suggest to implement to Mundys. Exclude suggestions that would change the business and business focus or model of Mundys. Add a column with a short description on why"
```

The output provided by NotebookLM in this third phase was as follows:

Table 8: EU Taxonomy activities implemented comparison between Mundys and FSI

Activity	**Mundys**	**FSI**	**Short Description**
Renewable Energy Integration in Operations	Implemented	Implemented	Both entities are actively involved in incorporating renewable energy sources into their operations. Mundys focuses on solar and wind energy, while Ferrovie dello Stato is exploring solar energy at stations and potentially using renewable sources to power trains.
Promotion of Sustainable Urban and Suburban Transport Systems	Implemented	Implemented	Both entities are engaged in promoting sustainable urban and suburban transport. Mundys focuses on developing and operating intermodal hubs and intelligent transport systems, while Ferrovie dello Stato actively develops and operates sustainable urban and suburban transport systems like trams, metros, and buses.
Implementation of Circular Economy Practices	Implemented	Implemented	Both entities are actively implementing circular economy principles. Mundys focuses on waste recycling and reuse in construction and promoting the use of recycled materials in road construction. Ferrovie dello Stato is also engaged in recycling and reuse, and promotes sustainable procurement practices.
Biodiversity Conservation Efforts	Implemented	Implemented	Both entities are involved in biodiversity conservation. Mundys focuses on minimizing land consumption and restoring ecosystems

Activity	Mundys	FSI	Short Description
			affected by infrastructure development. Ferrovie dello Stato is likely undertaking initiatives to conserve biodiversity along its extensive network and land holdings.
Development of Sustainable Passenger Rail Transport	Not Implemented	Implemented	Ferrovie dello Stato, as a passenger rail operator, likely has a significant portion of its activities aligned with sustainable passenger transport criteria. Mundys, primarily focused on infrastructure development, may not be directly involved in passenger rail operations.
Integration of Renewable Energy in Passenger Rail Transport	Not Implemented	Implemented	Ferrovie dello Stato is likely implementing measures to increase renewable energy use in its passenger rail operations, such as solar panels on stations or using renewable energy to power trains. This is not within Mundys' current business scope.
Promotion of Modal Shift towards Rail	Not Implemented	Implemented	Ferrovie dello Stato actively promotes shifting passenger and freight transport from road to rail to reduce emissions. Mundys, as an infrastructure developer, may not be directly involved in promoting modal shift campaigns.

This comparative analysis not only demonstrates the versatility of LLMs in identifying divergent implementation strategies but also offers valuable insights for companies seeking to benchmark their sustainability performance against industry leaders. The LLM's ability to pinpoint discrepancies in implementation across different companies highlights its potential as a powerful tool for fostering knowledge exchange and driving continuous improvement within the sustainability landscape. Furthermore, this analysis showcases how LLMs can help companies identify untapped opportunities for the implementation of sustainable practices by leveraging the successful strategies of other firms or competitors. By embracing such insights, companies can accelerate their progress towards a more sustainable future while simultaneously enhancing their competitive advantage.

4.4. Key Takeaways

The emergence of GenAI is bringing a paradigm shift in how companies integrate AI technologies into their daily operations. Unlike previous generations of AI tools, which were typically designed for narrow, task-specific applications (*Ashta & Herrmann, 2021*), GenAI, and LLMs in particular, offers a versatile, adaptive approach to handling complex, unstructured information. This evolution, facilitated by the introduction of ready-to-use web-based tools, is already reshaping business practices across sectors and will probably influence the way firms manage sustainability-related data in the future. In particular, LLMs present a transformative opportunity for how companies collect, process, and disclose non-financial information. Given their ability to interpret regulatory frameworks, analyze large volumes of heterogeneous documents, and generate structured outputs, LLMs may become central tools in enhancing both the efficiency and quality of corporate sustainability reporting (*Föhr et al., 2023; Zhuang & Wu, 2023*). As such, their integration could mark a turning point in advancing transparency and alignment with emerging regulatory standards like the EU Taxonomy.

Despite the potential of these new AI models, the consequences of such a shift still need to be understood. By exploiting a paradigmatic case study approach, this research aims at understanding how LLMs can be exploited to assess firms' sustainability efforts as provided through their NFDs. The assessment of LLM comprehension of the EU Taxonomy revealed a nuanced understanding of the technical screening criteria across various sectors. While occasional minor discrepancies were observed, the overall accuracy and alignment of the LLM-generated responses with the official EU Taxonomy documents were noteworthy. This suggests that LLMs, even without extensive fine-tuning, can effectively grasp the complex regulatory requirements and environmental objectives outlined in the taxonomy. More specifically, our three-phase approach highlighted the potential usefulness of LLMs in three distinct but intertwined steps in the analysis of firms' NFDS. First, the "generation and consolidation of implementation examples" phase showed the versatility of LLMs in bridging the gap between the broad outlines provided by the EU Taxonomy and the specific actions companies can take to align with its objectives. The LLM-generated examples were diverse, contextually relevant, and offered practical insights into how companies could implement sustainable practices across different sectors. Second, the company-specific analysis on Mundys showcased the practical applicability of the LLM-driven methodology in evaluating a company's alignment with the EU Taxonomy. The LLM effectively identified and ranked Mundys' implementation of various sustainable activities, providing a comprehensive overview of the company's sustainability efforts. Third, the cross-company comparison illustrated how LLMs can generate insights from a diverse range of sources and pinpoint potential areas of improvements for firms.

The findings of this study carry particular relevance when considered in light of the recent "Stop the Clock" directive, which postpones the enforcement of corporate obligations related to sustainability reporting and due diligence. This regulatory delay implies that a significant number of companies will be temporarily exempt from environmental disclosure requirements. As a result, there is a risk of creating a two-speed Europe in the field of sustainability disclosure, where only a subset of firms will be subject to formal reporting obligations. This divergence may further exacerbate the already existing heterogeneity in non-financial disclosures across firms. In this context, LLMs offer a promising tool for mitigating such inconsistencies. By enabling structured analysis of diverse and unstandardized disclosures, LLMs can support the development of a shared interpretative framework for evaluating corporate sustainability efforts. This common vocabulary can serve as a reference point for external analysts, and potentially for the companies themselves, when assessing the alignment of sustainability practices with regulatory expectations and broader environmental objectives.

The adoption of Generative AI in sustainability assessment and reporting is not just a technological and regulatory matter, but represents a significant managerial opportunity and challenge. Integrating GenAI tools is a driver of organizational change that requires new skills, adapted processes, and a reconsideration of information management strategies.

In line with the Technology Acceptance Model (Davis, 1989), the success of these solutions will depend on their technical performance and the perception of added value in terms of sustainable competitive advantage and improved corporate legitimization.

This integration allows companies to develop more sophisticated and data-driven Corporate Social Responsibility strategies, increasing transparency and credibility with stakeholders and supporting the transition towards a more resilient and long-term oriented business model (Strategic CSR).

This study actively supports companies in their journey of organizational change towards greater sustainability, enabled by technologies. GenAI emerges as a key tool not only for formulating and implementing more effective CSR strategies, but also for proactive risk management related to regulatory non-compliance, offering a framework for its strategic integration into corporate sustainability practices.

As with any study, this work is not without limitations. First, while the LLMs demonstrated a good understanding of the EU Taxonomy based on the provided reference documents, their performance could be further enhanced by incorporating additional context from a wider range of sources. However, this approach would be constrained by the information that companies choose to disclose in their sustainability reports and other publicly available documents. The comprehensiveness and transparency of corporate disclosures would directly impact the LLM's ability to accurately assess a company's alignment with the EU Taxonomy. Second, this study is limited to a single use case, focusing on the transport sector and a comparison between two companies. While this provides valuable insights into the potential of LLMs in this specific context, further research is needed to explore the generalizability of these findings to other sectors and a wider range of companies. Expanding the scope of the study would provide a more comprehensive understanding of the capabilities and limitations of LLMs in analyzing corporate sustainability disclosures and their alignment with the EU Taxonomy. Finally, we only relied on an out-of-the-box LLM tool. Our experimental setting, thus, does not exploit the full potential of LLM in analyzing large corpora of documents, since AI models tend to perform better when they are fine-tuned for a specific task (*Bommasani et al., 2021*). So, while this approach renders the analysis more replicable also for non-expert, our results must be interpreted as a lower bound for how the LLMs might behave in the context of corporate sustainability disclosure.

In conclusion, we believe that this case study will serve as a starting point for further, more systematic research into the transformative potential of LLMs in the realm of corporate sustainability disclosures. As these models continue to evolve and mature, their ability to process and analyze vast amounts of unstructured data could revolutionize the way companies report on their sustainability performance and how investors assess and compare corporate sustainability efforts. By harnessing the power of LLMs, we can move towards a more transparent, accountable, and sustainable business environment.

5. Conclusion

This research has evaluated the transformative potential of GenAI, and in particular LLMs such as Gemini 1.5 Pro, in advancing the analysis and interpretation of corporate environmental sustainability disclosures. By integrating GenAI tools across all phases of the study, the research not only assessed their technical capabilities but also reflected on their broader implications for internal and external sustainability stakeholders. This analysis primarily utilized the Gemini model; however, similar outcomes are expected with other prominent LLMs. As previously noted, the adoption of alternative models or tools necessitates a robust test framework to ensure reliability and consistent performance.

The investigation began with a comprehensive literature review on the adoption of GenAI in the context of environmental sustainability. Chapter 2, *Generative AI & corporate environmental sustainability: a literature review"*, systematically categorizes existing research into three thematic clusters:

- Applications of Generative AI for Improving Sustainability of Business Operations: This cluster highlights GenAI's use in optimizing product and process design (e.g., Life Cycle Assessment), enhancing supply chain management for sustainability, and improving resource utilization through accurate predictions.
- Generative AI for Sustainability Reporting and ESG Analysis: This section focuses on how LLMs can improve the quality of sustainability reporting by identifying intersections between various environmental accounting methods and standardizing information. It also covers the design of AI-powered tools for ESG assessment and risk evaluation, while noting challenges like greenwashing and data biases.
- Addressing Environmental Impact of Generative AI: This cluster addresses the energy-intensive nature of the AI industry, particularly GenAI, and the need for a multi-stakeholder approach to developing responsible AI that balances environmental benefits with its own carbon footprint.

Secondly, we assessed how GenAI techniques can be leveraged to analyze and process non-financial disclosure documents. Chapter 3, Generative AI to assess EU Taxonomy alignment in Non Financial Disclosures documents, explored the feasibility of using Gemini 1.5 with two distinct techniques: RAG and Long Context processing. The findings indicate that both approaches hold potential for automating and enhancing compliance assessment with the EU Taxonomy. RAG offers accuracy and transparency by grounding analysis in official documentation, while Long Context models provide a holistic understanding by processing entire documents. However, both methods have limitations, such as implementation complexity for RAG and computational expense and potential information overload for Long Context models. The empirical results from analyzing documents of Mundys and Maersk showed that Long Context achieved higher accuracy (90%) compared to RAG (60%) and a baseline (40%), albeit at a higher cost and latency.

Based on the demonstration that out-of-the-box models, with the right context, have a good understanding of the EU Taxonomy and the ability to extrapolate information, we conducted a case study to show how these tools can be leveraged to extrapolate information, identify similarities between firms, and suggest new activities. Chapter 4, *Generative AI to support ESG researchers: a case study*, focused on the corporate sustainability reports of Mundys S.p.A. and Ferrovie dello Stato Italiane (FSI) in the transport infrastructure sector. The case study highlighted GenAI's proficiency in three main tasks:

- Automating Data Extraction: LLMs can efficiently extract key data points, such as ESG metrics, risk factors, and regulatory compliance data, from non-financial documents.
- Identifying Trends and Patterns: LLMs can detect consistent improvements in ESG performance, highlight sectoral challenges, and identify emerging themes like biodiversity, supporting informed investment decisions.
- Generating Insights: LLMs can identify relationships between disparate pieces of information, assess the impact of ESG factors on financial performance, and identify best practices in ESG reporting. The three-phase approach demonstrated the versatility of LLMs in bridging the gap between broad EU Taxonomy outlines and specific company actions, evaluating company alignment, and generating insights through cross-company comparisons.

Together, these chapters demonstrate that GenAI is not merely a technical facilitator, but a foundational enabler of a new paradigm in corporate sustainability analysis. By bridging the gap between regulatory frameworks like the EU Taxonomy and the unstructured, heterogeneous disclosures produced by firms, GenAI tools can support more transparent, accountable, and data-driven sustainability ecosystems. Nonetheless, several limitations must be acknowledged. First, there remains a pressing need for standardized validation frameworks, such as benchmarking against “golden datasets", to assess the reliability and replicability of AI-generated outputs. Second, while Generative AI offers significant advantages, the continuous evolution of this technology necessitates a strategic approach: creating bespoke, dedicated models for ESG analysis may not be cost-effective given the rapid pace of innovation and the constant emergence of new, powerful out-of-the-box solutions. Instead, grounding existing highly-capable LLMs with specific ESG information appears to be a more viable and future-proof strategy, allowing continuous leveraging of their advanced features. The ability to ground information with the source of the answer is also crucial for maintaining transparency and trustworthiness in AI-generated insights.

Looking ahead, the evolution of GenAI presents compelling avenues for both academic research and practical implementation. Future studies could explore the use of multimodal GenAI systems capable of analyzing visual content, such as images and charts, embedded within sustainability reports, thereby enriching the scope and depth of ESG analysis. Ethical concerns, particularly around AI-enabled greenwashing, warrant urgent attention, and the development of detection tools to flag manipulative or misleading disclosures will be critical. Furthermore, real-time AI-based compliance monitoring systems could help businesses and regulators keep pace with shifting regulatory expectations, while expanding GenAI's role beyond analysis represents a major frontier for innovation.

In conclusion, this work contributes to an emerging field at the intersection of sustainability, regulatory compliance, and artificial intelligence. It highlights the potential for GenAI to serve not only as a reporting aid but as a strategic asset for sustainability-oriented decision-making, offering a practical and forward-looking roadmap for harnessing the next generation of AI technologies in service of environmental and corporate responsibility.

References

Abulibdeh, A., Zaidan, E., & Abulibdeh, R. (2024). Navigating the confluence of artificial intelligence and education for sustainable development in the era of industry 4.0: Challenges, opportunities, and ethical dimensions. Journal of Cleaner Production, 437, 140527. https://doi.org/10.1016/j.jclepro.2023.140527

Agathokleous, E., Saitanis, C. J., Fang, C., & Yu, Z. (2023). Use of ChatGPT: What does it mean for biology and environmental science? Science of The Total Environment, 888, 164154. https://doi.org/10.1016/j.scitotenv.2023.164154

Akhtar, P., Ghouri, A. M., Ashraf, A., Lim, J. J., Khan, N. R., & Ma, S. (2024). Smart product platforming powered by AI and generative AI: Personalization for the circular economy. International Journal of Production Economics, 273, 109283. https://doi.org/10.1016/j.ijpe.2024.109283

Alkire, L., Bilgihan, A., Bui, M. (Myla), Buoye, A. J., Dogan, S., & Kim, S. (2024). RAISE: Leveraging responsible AI for service excellence. Journal of Service Management, 35(4), 490–511. https://doi.org/10.1108/JOSM-11-2023-0448

Alzoubi, Y. I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. Journal of Cleaner Production, 468, 143090. https://doi.org/10.1016/j.jclepro.2024.143090

Ashta, A., & Herrmann, H. (2021). Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance. Strategic Change, 30(3), 211–222. https://doi.org/10.1002/jsc.2404

Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. Finance Research Letters, 47, 102776. https://doi.org/10.1016/j.frl.2022.102776

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2108.07258

Breijer, R., Erkens, M. H. R., Orij, R. P., & Vergoossen, R. G. A. (2025). Mandatory versus voluntary non-financial reporting: Reporting practices and economic consequences. Accounting Forum, 49(2), 303–335. https://doi.org/10.1080/01559982.2024.2326334

Brühl, V. (2023). The Green Asset Ratio (GAR): a new key performance indicator for credit institutions. Eurasian Economic Review, 13(1), 57-83. https://doi.org/10.1007/s40822-023-00224-0

Burger, B., Kanbach, D. K., Kraus, S., Breier, M., & Corvello, V. (2023). On the use of AI-based tools like ChatGPT to support management research. European Journal of Innovation Management, 26(7), 233–241. https://doi.org/10.1108/EJIM-02-2023-0156

Büyüközkan, G., & Karabulut, Y. (2018). Sustainability performance evaluation: Literature review and future directions. Journal of Environmental Management, 217, 253–267. https://doi.org/10.1016/j.jenvman.2018.03.064

Cao, Y., & Zhai, J. (2023). Bridging the gap – the impact of ChatGPT on financial research. Journal of Chinese Economic and Business Studies, 21(2), 177–191. https://doi.org/10.1080/14765284.2023.2212434

Cao, Y., & Zhai, J. (2023). Bridging the gap – the impact of ChatGPT on financial research. Journal of Chinese Economic and Business Studies, 21(2), 177–191. https://doi.org/10.1080/14765284.2023.2212434

Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: The growing role of generative AI in business and finance. Journal of Chinese Economic and Business Studies, 21(4), 471–496. https://doi.org/10.1080/14765284.2023.2245279

Chen, C.-T., Chen, S.-C., Khan, A., Lim, M. K., & Tseng, M.-L. (2024). Antecedents of big data analytics and artificial intelligence adoption on operational performance: The ChatGPT platform. Industrial Management & Data Systems, 124(7), 2388–2413. https://doi.org/10.1108/IMDS-10-2023-0778

Cho, C. H., Kim, A., Rodrigue, M., & Schneider, T. (2020). Towards a better understanding of sustainability accounting and management research and teaching in North America: A look at the community. Sustainability Accounting, Management and Policy Journal, 11(6), 985–1007. https://doi.org/10.1108/SAMPJ-08-2019-0311

Cornago, S., Ramakrishna, S., & Low, J. S. C. (2023). How can Transformers and large language models like ChatGPT help LCA practitioners? Resources, Conservation and Recycling, 196, 107062. https://doi.org/10.1016/j.resconrec.2023.107062

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. Management Information Systems Quarterly, 13(3), 319-340, https://doi.org/10.2307/249008

De Villiers, C., Dimes, R., & Molinari, M. (2024). How will AI text generation and processing impact sustainability reporting? Critical analysis, a conceptual framework and avenues for future research. Sustainability Accounting, Management and Policy Journal, 15(1), 96–118. https://doi.org/10.1108/SAMPJ-02-2023-0097

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Dyer, W. G., & Wilkins, A. L. (1991). Better Stories, Not Better Constructs, to Generate Better Theory: A Rejoinder to Eisenhardt. The Academy of Management Review, 16(3), 613. https://doi.org/10.2307/258920

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150-158. https://doi.org/10.18653/v1/2024.eacl-demo.16

European Environment Agency. (2024). Sustainability of Europe's mobility systems. Publications Office. https://data.europa.eu/doi/10.2800/8560026

Feng, J., Ning, Y., Wang, Z., Li, G., & Xiu Xu, S. (2024). ChatGPT-enabled two-stage auctions for electric vehicle battery recycling. Transportation Research Part E: Logistics and Transportation Review, 183, 103453. https://doi.org/10.1016/j.tre.2024.103453

Föhr, T. L., Schreyer, M., Juppe, T. A., & Marten, K.-U. (2023). Assuring Sustainable Futures: Auditing Sustainability Reports using AI Foundation Models [Working Paper]. https://www.ssrn.com/abstract=4502549

Fosso Wamba, S., Guthrie, C., Queiroz, M. M., & Minner, S. (2024). ChatGPT and generative artificial intelligence: An exploratory study of key benefits and challenges in operations and supply chain management. International Journal of Production Research, 62(16), 5676–5696. https://doi.org/10.1080/00207543.2023.2294116

Galletta, S., Mazzù, S., Naciti, V., & Paltrinieri, A. (2024). A PRISMA systematic review of greenwashing in the banking industry: A call for action. Research in International Business and Finance, 69, 102262. https://doi.org/10.1016/j.ribaf.2024.102262

Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., Tacchetti, A., … Vinyals, O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2403.05530

Ghobakhloo, M., Fathi, M., Iranmanesh, M., Vilkas, M., Grybauskas, A., & Amran, A. (2024). Generative artificial intelligence in manufacturing: Opportunities for actualizing Industry 5.0 sustainability goals. Journal of Manufacturing Technology Management, 35(9), 94–121. https://doi.org/10.1108/JMTM-12-2023-0530

Gu, H., Schreyer, M., Moffitt, K., & Vasarhelyi, M. (2024). Artificial intelligence co-piloted auditing. International Journal of Accounting Information Systems, 54, 100698. https://doi.org/10.1016/j.accinf.2024.100698

Guo, M., Zong, X., Guo, L., & Lei, Y. (2024). Does haze-related sentiment affect income inequality in China? International Review of Economics & Finance, 94, 103371. https://doi.org/10.1016/j.iref.2024.05.050

Haddud, A. (2024). ChatGPT in supply chains: Exploring potential applications, benefits and challenges. Journal of Manufacturing Technology Management. https://doi.org/10.1108/JMTM-02-2024-0075

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., … Zhu, J. (2021).

Pre-trained models: Past, present and future. AI Open, 2, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

Hope, O. (2003). Disclosure Practices, Enforcement of Accounting Standards, and Analysts' Forecast Accuracy: An International Study. Journal of Accounting Research, 41(2), 235–272. https://doi.org/10.1111/1475-679X.00102

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research, 40(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Hummel, K., & Jobst, D. (2024). An Overview of Corporate Sustainability Reporting Legislation in the European Union. Accounting in Europe, 21(3), 320–355. https://doi.org/10.1080/17449480.2024.2312145

IPCC. (2018). Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty (1a ed.). Cambridge University Press. https://doi.org/10.1017/9781009157940

Kar, A. K., Varsha, P. S., & Rajan, S. (2023). Unravelling the Impact of Generative Artificial Intelligence (GAI) in Industrial Applications: A Review of Scientific and Grey Literature. Global Journal of Flexible Systems Management, 24(4), 659–689. https://doi.org/10.1007/s40171-023-00356-x

Kinderman, D. (2020). The challenges of upward regulatory harmonization: The case of sustainability reporting in the European Union. Regulation & Governance, 14(4), 674–697. https://doi.org/10.1111/rego.12240

Klarin, A. (2024). How to conduct a bibliometric content analysis: Guidelines and contributions of content co‐occurrence or co‐word literature reviews. International Journal of Consumer Studies, 48(2), e13031. https://doi.org/10.1111/ijcs.13031

Krippendorff, K. (2019). Content analysis: An introduction to its methodology (Fourth edition). SAGE.

Leippold, M. (2023). Thus spoke GPT-3: Interviewing a large-language model on climate finance. Finance Research Letters, 53, 103617. https://doi.org/10.1016/j.frl.2022.103617

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & others. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A Survey on Retrieval-Augmented Text Generation (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2202.01110

Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A Survey on Retrieval-Augmented Text Generation (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2202.01110

Loughran, T., & McDonald, B. (2020). Textual Analysis in Finance. Annual Review of Financial Economics, 12(1), 357–375. https://doi.org/10.1146/annurev-financial-012820-032249

Maibaum, F., Kriebel, J., & Foege, J. N. (2024). Selecting textual analysis tools to classify sustainability information in corporate reporting. Decision Support Systems, 183, 114269. https://doi.org/10.1016/j.dss.2024.114269

Mariani, M., & Dwivedi, Y. K. (2024). Generative artificial intelligence in innovation management: A preview of future research developments. Journal of Business Research, 175, 114542. https://doi.org/10.1016/j.jbusres.2024.114542

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. Journal of Informetrics, 12(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Moreno, A.-I., & Caminero, T. (2024). Assessing climate-related disclosures of European banks through text mining. Review of World Economics. https://doi.org/10.1007/s10290-024-00531-x

Moreno, A.-I., & Caminero, T. (2024). Assessing climate-related disclosures of European banks through text mining. Review of World Economics. https://doi.org/10.1007/s10290-024-00531-x

Nandiraju, A., & Kanthi, S. (2024). Leveraging Multimodal Large Language Models and Natural Language Processing Techniques for Comprehensive ESG Risk Score Prediction: Proceedings of the 6th International Conference on Finance, Economics, Management and IT Business, 69–78. https://doi.org/10.5220/0012725700003717

O'Reilly, S., Gorman, L., Mac An Bhaird, C., & Brennan, N. M. (2024). Implementing the European Union Green Taxonomy: Implications for small- and medium-sized enterprises. Accounting Forum, 48(3), 401–426. https://doi.org/10.1080/01559982.2023.2272394

Ong, K., Mao, R., Satapathy, R., Filho, R. S., Cambria, E., Sulaeman, J., & Mengaldo, G. (2025). Explainable natural language processing for corporate sustainability analysis. Information Fusion, 115, 102726. https://doi.org/10.1016/j.inffus.2024.102726

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and examples for reporting systematic reviews. BMJ, n160. https://doi.org/10.1136/bmj.n160

Polyportis, A., & Pahos, N. (2024). Navigating the perils of artificial intelligence: A focused review on ChatGPT and responsible research and innovation. Humanities and Social Sciences Communications, 11(1), 107. https://doi.org/10.1057/s41599-023-02464-6

Preuss, N., Alshehri, A. S., & You, F. (2024). Large language models for life cycle assessments: Opportunities, challenges, and risks. Journal of Cleaner Production, 466, 142824. https://doi.org/10.1016/j.jclepro.2024.142824

Ramautar, V., Ritfeld, N., Brinkkemper, S., & España, S. (2024). Optimising Sustainability Accounting: Using Language Models to Match and Merge Survey Indicators. In J. Araújo, J. L. De La Vara, M. Y. Santos, & S. Assar (A c. Di), International Conference on Research Challenges in Information Science (Vol. 513, pp. 338–354). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-59465-6_21

Rau, P. R., & Yu, T. (2024). A survey on ESG: Investors, institutions and firms. China Finance Review International, 14(1), 3–33. https://doi.org/10.1108/CFRI-12-2022-0260

Rillig, M. C., & Kasirzadeh, A. (2024). AI Personal Assistants and Sustainability: Risks and Opportunities. Environmental Science & Technology, 58(17), 7237–7239. https://doi.org/10.1021/acs.est.4c03300

Scartozzi, G., Delladio, S., Rosati, F., Nikiforou, A. I., & Caputo, A. (2024). The social and environmental impact of entrepreneurship: A review and future research agenda. Review of Managerial Science. https://doi.org/10.1007/s11846-024-00783-9

Schütze, F., & Stede, J. (2024). The EU sustainable finance taxonomy and its contribution to climate neutrality. Journal of Sustainable Finance & Investment, 14(1), 128–160. https://doi.org/10.1080/20430795.2021.2006129

Sedkaoui, S., & Benaichouba, R. (2024). Generative AI as a transformative force for innovation: A review of opportunities, applications and challenges. European Journal of Innovation Management. https://doi.org/10.1108/EJIM-02-2024-0129

Sedkaoui, S., & Benaichouba, R. (2024). Generative AI as a transformative force for innovation: A review of opportunities, applications and challenges. European Journal of Innovation Management. https://doi.org/10.1108/EJIM-02-2024-0129

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Sommer, B., & Von Querfurth, S. (2024). "In the end, the story of climate change was one of hope and redemption": ChatGPT's narrative on global warming. Ambio, 53(7), 951–959. https://doi.org/10.1007/s13280-024-01997-7

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. International Journal of Information Management, 74, 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Susithra, S., & Vasantha, S. (2024). The Adoption of Green Supply Chain Practices using Artificial Intelligence (AI) for Smart Global Value Chain. 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 1–8. https://doi.org/10.1109/ICONSTEM60960.2024.10568808

Toetzke, M., Probst, B., & Feuerriegel, S. (2023). Leveraging large language models to monitor climate technology innovation. Environmental Research Letters, 18(9), 091004. https://doi.org/10.1088/1748-9326/acf233

Trajanov, D., Lazarev, G., Chitkushev, L., & Vodenska, I. (2023). Comparing the performance of ChatGPT and state-of-the-art climate NLP models on climate-related text classification tasks. E3S Web of Conferences, 436, 02004. https://doi.org/10.1051/e3sconf/202343602004

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. Scientometrics, 81(2), 587–600. https://doi.org/10.1007/s11192-009-2178-0

Wang, N., Pan, H., Feng, Y., & Du, S. (2023). How do ESG practices create value for businesses? Research review and prospects. Sustainability Accounting, Management and Policy Journal. https://doi.org/10.1108/SAMPJ-12-2021-0515

Wang, S., Liu, C., & Zhou, Z. (2024). Government-enterprise green collaborative governance and urban carbon emission reduction: Empirical evidence from green PPP programs. Environmental Research, 257, 119335. https://doi.org/10.1016/j.envres.2024.119335

Xu, Y., Shao, X., & Tanasescu, C. (2024). How are artificial intelligence, carbon market, and energy sector connected? A systematic analysis of time-frequency spillovers. Energy Economics, 132, 107477. https://doi.org/10.1016/j.eneco.2024.107477

Yang, H., & Zhou, X. (2023). Study on the economic benefits of carbon-neutral digital platforms for sustainable development based on the GPT-QRCNN model. Frontiers in Ecology and Evolution, 11, 1263799. https://doi.org/10.3389/fevo.2023.1263799

Yang, T., Mei, Y., Xu, L., Yu, H., & Chen, Y. (2024). Application of question answering systems for intelligent agriculture production and sustainable management: A review. Resources, Conservation and Recycling, 204, 107497. https://doi.org/10.1016/j.resconrec.2024.107497

Zhuang, X., & Wu, Y. (2023). ChatGPT: How it can support corporate social responsibility. Journal of Business Strategy. https://doi.org/10.1108/JBS-05-2023-0089

Zhuang, X., & Wu, Y. (2023). ChatGPT: How it can support corporate social responsibility. Journal of Business Strategy. https://doi.org/10.1108/JBS-05-2023-0089

6. Appendix

This Appendix presents the full output produced by NotebookLM, which was used as a companion for the development of the analysis phase. Specifically, the prompt used to ask the LLM to produce thematic clusters of similar papers, as presented in Section 3, was as follows: "You are a researcher in the economics and business field. You are doing a systematic literature review about the implications of the introduction of Generative AI for the environmental sustainability of firms. You want to cluster all the uploaded documents into different groups depending on the major topic discussed in each document so that you can subsequently make more in-depth analysis. How many clusters can you identify? Make the clusters as comprehensive and specific as possible. Justify your answer. I suggest making as many clusters as possible and then merge the clusters that are too similar to be stand-alone clusters."
The final output of NotebookLM was the following:

Following your suggested approach, a comprehensive analysis of the provided sources reveals seven potential clusters, some of which can be merged due to their close relationship:

1. **Generative AI for Enhanced Sustainability Reporting:** This cluster encompasses the use of generative AI for improving the quality, efficiency, and scope of sustainability reporting in firms. It includes:
    - 1.1. Leveraging AI for data collection and analysis from various sources, including financial statements, sustainability reports, and ESG ratings.
    - 1.2. Automating report generation and ensuring the accuracy and reliability of reported information, potentially reducing the cost and time involved in traditional reporting processes.
    - 1.3. Addressing the potential of AI to perpetuate or even exacerbate biases in reporting and the need for transparency and accountability in AI-driven reporting processes.
2. **Combating Greenwashing with Generative AI:** Due to its close connection to sustainability reporting, this cluster can be merged with Cluster 1. It focuses on the risk of firms using generative AI to fabricate or exaggerate their sustainability performance, misleading stakeholders.
3. **Generative AI for Sustainable Manufacturing and Operations:** This cluster centers on applying generative AI to advance sustainability within Industry 5.0, which emphasizes human-centric and sustainable manufacturing. It includes:
    - 3.1. Optimizing production processes, minimizing waste, and enhancing resource efficiency through data-driven insights provided by AI.

    3.2. Improving supply chain resilience, transparency, and sustainability by leveraging AI for demand forecasting, inventory management, and risk mitigation.
4. **Generative AI for Sustainable Product Design and Lifecycle Assessment:** This distinct cluster emphasizes using generative AI to design environmentally friendly products and assess their environmental impact throughout their lifecycle. It involves:
    4.1. Leveraging AI for personalized product design and manufacturing that incorporates circular economy principles, such as using recycled materials and minimizing waste.
    4.2. Employing AI in life cycle assessment (LCA) to evaluate the environmental impacts of products and services across their entire life cycle, from raw material extraction to disposal.
5. **Green AI Initiatives and Tools:** This cluster focuses on developing and implementing AI technologies that are inherently environmentally friendly. This includes:
    5.1. Optimizing AI algorithms and models for energy efficiency to reduce the carbon footprint of AI itself.
    5.2. Leveraging cloud optimization strategies and promoting the use of renewable energy sources in data centers that power AI applications.
    5.3. Developing specialized AI solutions to address specific environmental challenges, such as climate change mitigation, biodiversity conservation, and pollution reduction.
6. **Generative AI for ESG Risk Assessment and Sustainable Finance:** This cluster examines how generative AI can analyze large datasets to evaluate a company's ESG performance and potential risks. It includes:
    6.1. Developing AI-powered tools to assess and predict ESG risks, enabling investors to make more informed and sustainable investment decisions.
    6.2. Analyzing corporate disclosures, news articles, social media sentiment, and other relevant data to generate insights into companies' ESG practices.
    6.3. Supporting the growth of sustainable finance initiatives by identifying investment opportunities that align with ESG principles.
7. **Generative AI and Responsible AI Development for Sustainability:** Given its significant overlap with Cluster 5 (Green AI Initiatives), this cluster can be merged. It emphasizes the ethical considerations and potential risks associated with developing and deploying AI for sustainability. This includes:
    7.1. Ensuring the responsible development and use of AI, addressing concerns related to bias, fairness, transparency, and accountability.
    7.2. Promoting a multi-stakeholder approach that includes input from policymakers, industry experts, researchers, and civil society to guide the ethical development and deployment of AI for sustainability.

```
In summary, the sources point to five overarching clusters:
● Generative AI for Enhanced Sustainability Reporting (including combating greenwashing).
● Generative AI for Sustainable Manufacturing and Operations.
● Generative AI for Sustainable Product Design and Lifecycle Assessment.
● Green AI Initiatives and Tools (including responsible AI development for
  sustainability)
● Generative AI for ESG Risk Assessment and Sustainable Finance.

These clusters provide a framework for conducting a systematic literature review by
categorizing the different ways generative AI can impact the environmental sustainability of
firms.
```

Additionally, the LLM was also used in the second step of the analysis to suggest future developments of the literature based on the thematic clusters created in the first phase. In this phase, we used the polished clusters which were created as a refinement of the initial clusters provided by NotebookLM. Specifically, the prompt used to ask to the LLM to produce future research directions, as presented in Section 4, was as follows: “You are a researcher in the economics and business field. You are doing a systematic literature review about the implications of the introduction of Generative AI for the environmental sustainability of firms. You want to identify the future research directions of the literature based on the three following clusters: "Cluster 1: Applications of Generative AI for improving Sustainability of Business Operations", "Cluster 2: Generative AI for Sustainability Reporting and ESG Analysis", "Cluster 3: Addressing Environmental Impact of Generative AI". Justify your answers.”

The final output of NotebookLM was the following:

```
Here are some potential future research directions for the literature on Generative AI and
environmental sustainability of firms, based on the three clusters:
Future Research Directions on Generative AI and Environmental Sustainability
Cluster 1: Applications of Generative AI for Improving Sustainability of Business Operations
● Investigate the use of Generative AI for designing eco-friendly products and processes.
  This could involve using Generative AI to optimize product designs for minimal material
  use, generate novel sustainable materials or develop more efficient manufacturing
  processes. For example, future research could expand on the application of ChatGPT to
  generate optimized Gcode in 3D printing.
● Explore how Generative AI can be used to optimize supply chains for sustainability.
  This could involve using Generative AI to develop circular economy models, optimize
```

logistics for minimal environmental impact, or improve supply chain transparency and traceability. For instance, research could focus on how to use AI to analyze consumer consumption patterns and predict resource usage in production planning within a circular economy.

- **Examine the potential of Generative AI to enhance resource efficiency in various business operations.** This could include using Generative AI to optimize energy consumption in buildings, reduce water usage in manufacturing processes, or minimize waste generation in supply chains.

Cluster 2: Generative AI for Sustainability Reporting and ESG Analysis

- **Examine the potential of Generative AI to automate the process of sustainability reporting**, making it more efficient and cost-effective for firms. This could involve developing AI algorithms that can extract relevant data from company documents, generate reports based on predefined templates, or even draft customized narratives based on specific user queries.
- **Investigate how Generative AI can enhance the quality and comparability of sustainability reports across different firms.** This could involve developing standardized frameworks for using Generative AI in sustainability reporting, establishing guidelines for data collection and analysis or creating benchmarks for evaluating the quality of AI-generated reports.
- **Address the ethical and practical challenges associated with using Generative AI in sustainability reporting.** This includes mitigating the risk of bias in AI algorithms, ensuring data privacy and security, and maintaining transparency and accountability in the reporting process.
- **Explore the use of Generative AI for ESG analysis and investment.** Research could investigate if measures of corporate culture or ESG scores produced by generative AI through textual analysis are consistent with or even superior to existing measures. For example, AI could be used to analyze corporate sentiments towards environmental policy by inputting text extracted from corporate financial statements.

Cluster 3: Addressing Environmental Impact of Generative AI

- **Develop a more comprehensive understanding of the carbon footprint of Generative AI.** This could include quantifying the energy consumption and emissions associated with different stages of Generative AI lifecycle - development, training, deployment and use. Researchers could investigate how AI emissions impact climate change and the environment.
- **Explore strategies to mitigate the environmental impact of Generative AI.** This could involve developing more energy-efficient AI algorithms, optimizing hardware and infrastructure for sustainability, or promoting responsible AI development and use practices within organizations. For example, research could focus on developing a comprehensive framework for the use of AI in non-financial reporting, including

guidelines for data collection, analysis, presentation and communication that minimizes environmental impact.

- **Investigate the use of Generative AI itself to address its own environmental impact.** This could involve training Generative AI models to optimize energy consumption in data centers, develop renewable energy sources or design more sustainable hardware for AI applications.

The sources highlight the emergence of "green AI" as a response to the environmental impact of AI. This field focuses on developing AI tools and techniques that minimize environmental impact and encourage the use of renewable energy sources. Future research could investigate the effectiveness and limitations of various green AI initiatives, as well as their contribution to achieving specific Sustainable Development Goals (SDGs). Additionally, research is needed to validate the discussions around green AI and explore new initiatives in this rapidly evolving field.