Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Information and Service Management

Evaluating AI's Ability to Analyse Sustainability Reports

A Case Study of India and Recommendations for CSRD Development

---

Pauliina Penttilä & Nette Poutiainen

Master's thesis
2025

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Pauliina Penttilä and Nette Poutiainen

Title of thesis Evaluating AI's ability to Analyse Sustainability Reports – A Case Study of India and Recommendations for CSRD development

Programme Master of Science in Economics and Business Administration

Major Information and Service Management

Thesis supervisor Prof. Esko Penttinen

Thesis advisor(s) Prof. Esko Penttinen

Date 12.06.2025 **Number of pages** 79 **Language** English

Abstract

Sustainability reporting has become an increasingly important part of corporate reporting. The assessment of narrative data however is not straightforward as rich descriptions can make it difficult to evaluate the actual situation. Fortunately, Natural Language Processing (NLP) and Artificial Intelligence (AI) are tools for more efficient analysis of data. By accurately assessing reports, companies can be evaluated on a more equal basis and their actual actions can be identified. As the Corporate Sustainability Reporting Directive (CSRD) aims to strengthen corporate responsibility for sustainability, the evaluation of reporting is crucial to assess real actions and their shortcomings. Thus, various stakeholders assessing companies' sustainability from different perspectives benefit from AI-driven insights and the identification of deficiencies in sustainability reporting.

In this thesis we focus to examine the benefits and limitations of machine-readable XBRL data when processed by AI, in comparison to PDF-based data. This study is done by looking at Indias Business Responsibility and Sustainability Reports (BRSR) in different file formats with OpenAI's ChatGPT 4o. GPT is asked to perform various tasks, and the accuracy of answers are then analysed. Based on results the best success rate for analysing multiple reports and data appear when using XBRL-XML format with the taxonomy file provided. Use of XBRL-JSON is not recommended in this context. Additionally in terms of context of the study the best success rate for analysing a single report was achieved when using PDF. However, there is a lack of transparency when utilizing AI and the possible hallucinations decreases the reliability. As a conclusion, the result of the study indicates potential in leveraging AI when analysing sustainability data.

Keywords Sustainability reporting, Sustainability, AI, LLM, GPT, BRSR, CSRD

Aalto-yliopisto
Aalto-universitetet
Aalto University

Tekijä Pauliina Penttilä ja Nette Poutiainen

Työn nimi Tekoälyn kyvyn arviointi vastuullisuusraporttien analysoinnissa – Tapausanalyysi Intiasta ja suosituksia CSRD-vaatimusten kehittämiseen.

Koulutusohjelma Kauppatieteiden Maisteri

Pääaine Tieto- ja palvelujohtaminen

Vastuuopettaja/valvoja/ohjaaja Prof. Esko Penttinen

Päivämäärä 12.06.2025 **Sivumäärä** 79 **Kieli** Englanti

Tiivistelmä

Vastuullisuusraportoinnista on tullut yhä tärkeämpi osa yritysraportointia. Narratiivisen datan analysointi ei kuitenkaan ole yksinkertaista, sillä selittävät kuvaukset voivat vaikeuttaa todellisen tilanteen arviointia. Onneksi luonnollisen kielen käsittely (Natural Language Processing, NLP) ja tekoäly (AI) tarjoavat tehokkaampia työkaluja datan analysointiin. Raporttien tarkka arviointi mahdollistaa yritysten arvioimisen tasavertaisemmin ja todellisten toimien tunnistamisen. Koska yritysten kestävyysraportointidirektiivin (CSRD) tavoitteena on vahvistaa yritysvastuuta kestävyyden näkökulmasta, raportoinnin arviointi on keskeistä todellisten toimien ja niiden puutteiden tunnistamisessa. Tämän vuoksi eri sidosryhmät, jotka arvioivat yritysten vastuullisuutta eri näkökulmista, hyötyvät tekoälypohjaisista havainnoista ja vastuullisuusraportoinnin puutteiden tunnistamisesta.

Tässä opinnäytetyössä keskitytään koneluettavan XBRL-datan hyötyjen ja rajoitteiden tarkasteluun verrattuna PDF-muotoiseen dataan, tekoälyä hyödynnettäessä. Tutkimus toteutetaan tarkastelemalla Intian vastuullisuus raportteja (BRSR) eri tiedostomuodoissa ja hyödynnetään OpenAI:n ChatGPT 4o –mallia. GPT-mallilla suoritetaan erilaisia tehtäviä, ja vastausten oikeinmukaisuutta analysoidaan. Tulosten perusteella paras onnistumisaste analysoitaessa yhtä aikaa useaa raporttia saavutettiin käytettäessä XBRL-XML-muotoa yhdessä taksonomiatiedoston kanssa. XBRL-JSON-muodon käyttöä ei suositella tässä kontekstissa. Lisäksi tutkimuksen kontekstissa parhain onnistumisaste yksittäisen raportin analysoinnissa saavutettiin käytettäessä PDF-muotoa. Tekoälyn käytössä esiintyvä läpinäkyvyyden puute sekä mahdolliset “hallusinaatiot” heikentävät kuitenkin luotettavuutta. Yhteenvetona tutkimuksen tulokset osoittavat potentiaalia tekoälyn hyödyntämisessä vastuullisuus dataa analysoitaessa.

Avainsanat Vastuullisuusraportointi, Kestävyys, AI, LLM, GPT, BRSR, CSRD

Table of contents

Abbreviations

Sustainability Abbreviations

CSDR	Corporate Sustainability Reporting Directive
ESRS	European Sustainability Reporting Standards
BRSR	Business Responsibility and Sustainability Reporting
NGRBC	National Guidelines for Responsible Business Conduct

Data Abbreviations

XBRL	eXtensible Business Reporting Language
PDF	Portable Document Format
JSON	JavaScript Object Notation
XML	Extensible markup language

AI Abbreviations

LLM	Large Language Model
Gen AI	Generative Artificial Intelligence
GPT	Generative Pre-training Transformer
NLP	Natural language processing

1 Introduction

1.1 Background and Motivation

Sustainability reporting started to attract interest in the 1970s and has become an increasingly important part of corporate reporting year by year (Kolk et al., 2005). Now, in the 2020s, sustainability reporting is even more relevant as new obligations for companies operating in Europe are being introduced with the Corporate Sustainability Reporting Directive (CSRD), which reshapes the reporting of business activities' sustainability by increasing its transparency (Baumüller & Grbenic, 2021).

However, the assessment of narrative data is not straightforward as rich descriptions can make it difficult to evaluate the actual situation and may thus contribute to greenwashing (Ramanan, 2024a). Fortunately, Natural Language Processing (NLP) enables the analysis of narrative text (Wong et al., 2018), and therefore, advancements in AI and its growing application possibilities can also enhance the analysis of sustainability reports.

By accurately assessing reports, companies can be evaluated on a more equal basis and their actual actions can be identified from rich descriptive text. As the CSRD aims to strengthen corporate responsibility for sustainability (Baumüller & Grbenic, 2021), the evaluation of reporting is crucial to assess real actions and their shortcomings enabling further development. Thus, various stakeholders assessing companies' sustainability from different perspectives benefit from AI-driven insights and the identification of deficiencies in sustainability reporting.

Moreover, timeliness of the topic makes it a particularly interesting. In order for business operations to shift toward more sustainable actions, companies' activities must be evaluated accurately. The utilization of AI creates new opportunities for this as it can enhance the efficiency and accuracy of sustainability assessments.

There is previous research on parts of the subject discussed in this thesis. Hillebrand et al. (2023) studied an AI context-aware recommender system that can outperform traditional report analysis methods. Ramanan (2024b) looks into LLM's reading PDF vs. XBRL reports and compares how the two forms perform differently. PwC (2020) also sees the value in XBRL machine-readable forms in reporting to lower cost and labor time. The opportunities and challenges of sustainability reporting have also been researched (Di Tullio et al., 2023; Ahmad et al., 2023).

Though this subject has been studied from various aspects it still needs more quantitative studies assessing the different forms read by AI. Objective research in XBRL and PDF comparisons is lacking, as the research is largely done by the XBRL institution. De Villiers et al. (2024, p.26) also recognizes the need for research on the "Impact of AI on different reporting types and the emergence of new types of reporting". Ahmad et al. (2023, p. 648) also recognize that "there is a lack of empirical research on the actual impact of AI on sustainability reporting in Accounting". There is also no industry-specific analysis in the energy sector on the subject in this thesis.

1.2 Research Objectives

There is a lack of holistic approaches to evaluate structured and unstructured data from company reports and making insights in AI-driven analysis. This gap limits the ability to use sustainability information in strategic decision-making and creates obstacles to equal compliance and evaluation of reporting standards. The objective is to analyse how correctly AI can analyse sustainability reports looking at different task types, as well as different file formats with and without taxonomy files, explaining the machine-readable tags. Afterwards the objective is to develop some practical recommendations for CSRD and companies on the use of AI in this topic.

1.3 Research Questions

The theoretical purpose of this study is to examine the benefits and limitations of machine-readable XBRL data when processed by AI, in comparison to PDF format data. The literature review aims to refine knowledge on various aspects of sustainability reporting in the EU and India, and assess the current understanding of AI's role, potential and risks in the field of sustainability reporting. Additionally, the overview of XBRL data is presented.

This leads to the research questions of this study:

1. How well can ChatGPT analyse sustainability reports in different file formats (PDF, XBRL-XML and XBRL-JSON)

2. How does the performance compare between different task types such as analysing data and text.

This study is executed by looking at Indias Business Responsibility and Sustainability Reports (BRSR) in different file formats with OpenAI's ChatGPT 4o. GPT is asked to perform various tasks, and the accuracy of answers are then analysed. The study examines different formats (PDF, XBRL-XML with

and without a taxonomy file, and XBRL-JSON with and without a taxonomy file) in which the questions are implemented. There are five tasks being studied, each of which is repeated three times in every project. In addition, the study is conducted using both the ChatGPT Plus and Pro versions. As a result, the research analyses a total of 150 responses generated by ChatGPT, which serve as the basis for compiling the results.

This thesis is structured as follows. Section 2 provides a review of the relevant literature, covering key theories in information systems as well as characteristics of structured and unstructured data. Also, the value of information and applications of artificial intelligence (AI), including an introduction to generative AI is presented. It introduces an overview of the XBRL format and describes the main characteristics of sustainability reporting. Section 3 outlines the methodology of the study, including the research approach, data selection, execution of the analysis and key limitations.

Further, section 4 presents the results of the analysis, which includes both numerical and textual data extracted by ChatGPT, along with summary of all results. Section 5 provides discussion of the findings and reflects it to both theoretical and practical implications. The fifth part also acknowledges limitations and proposes directions for future research. Finally, Section 6 concludes the thesis by summarizing the key findings and reflecting on their broader relevance for sustainability reporting and the use of AI in data analysis.

2 Literature review

In this chapter, we establish the theoretical and contextual foundation for the study. Section 2.1 introduces digital information systems and examines their role in modern society, with a focus on the distinction between structured and unstructured data. Section 2.2 explores the eXtensible Business Reporting Language (XBRL). This includes the key characteristics and the role of taxonomies in XBRL reporting. In Section 2.3 artificial intelligence is discussed, looking at relevant concepts such as large language models, generative AI, and Natural Language Processing (NLP). Additionally, their leveraging in various business cases is investigated. Finally, in Section 2.4 the core sustainability reporting frameworks relevant to this thesis are presented: India's Business Responsibility and Sustainability Report (BRSR) and the European Union's Corporate Sustainability Reporting Directive (CSRD).

2.1 Digital Information Systems and the Nature of Data

In today's organizations, operating without information systems is almost impossible. We are a "Hostage to information systems" as stated in the book Introduction to information systems (p. 9) by Rainer, et al. (2020). Information systems are truly seen as an essential part of the modern business environment, and for a good reason.

For example, as Cusumano et al. (2020) mentions, platform-based business models that build their operations on knowledge-based business and information systems create a very valuable and growing market today. They also state that many of the largest companies are built on digital platforms, such as Meta, Microsoft, and Amazon. One of the important keys to their success is the utilization of information flows (Cusumano et al., 2020).

Additionally, inside every organization, information connects to strategy and power as a large amount of data creates a new perspective on the use of information. Bhimani (2015) highlights in his article how arguments based on big data presented in support of decision-making can greatly affect organizational decision-making and dynamics in the exercise of power. Thus, according to him, one might also argue that the ability to utilize information obtained through big data enables power and influence in the organization, even creating policy-related consequences which can also be seen as changes in the relationships between stakeholders.

However, Müller et al., (2016) points out that applying big data analytics does not necessarily lead to more accurate decisions and higher business value, as there are many challenges involved in creating new insights. In this regard,

as Abbasi et al. (2016) bring up, companies see big data as one of the most important business advantages and are collecting it more than ever before. Therefore, big data modeling formalisms and integration creation are needed. The data produced today is rich in, for example, knowledge, emotions and geographical information and present opportunities for new prediction creation (Abbasi et al., 2016). However, as a result of the increasing use of data, organizations also need the ability to assess the value that can be obtained from large amounts of diverse data, taking into account the costs that arise from managing its reliability (Abbasi et al., 2016).

Luckily, as rich, unstructured data is complicated to handle, applying NLP can help to understand textual information better (Gharehchopogh & Khalifelu, 2011). Hence, it can be concluded that utilizing artificial intelligence in analysing collected data has a great potential to move towards more efficient decision-making. Nowadays, artificial intelligence is rapidly becoming a part of the business world. Generative Artificial Intelligence creates content using neural network-based logic for data processing (Susarla et al., 2023).

The release of ChatGPT made accessible the wide range of functionalities offered by generative AI models, which are easy for the user to use on a simple browser-based platform (Susarla et al., 2023). Furthermore, as AI technology has become an important part of the organizations' business models attracting academic interest, recent studies have been more interested in the impacts and consequences of AI technology than its performance (Dwivedi et al., 2021). Therefore, it is beneficial to examine the current performance of AI and how it can be utilized within its current technical capabilities, which is the focus of this study.

Moreover, a notable characteristic feature is that, when using models such as ChatGPT new tasks can be performed without coding skills as it can be instructed by using natural language (Susarla et al., 2023). Unlike traditional computer programs, machine learning (ML) algorithms do not need precise instructions, but are able to infer them based on examples. For example, when ML systems are given sets of images, the system can learn to recognize a person without being given precise characteristics (Brynjolfsson et al., 2025). This can be seen as revolutionizing the use of information systems and the creation of new information today, as fewer technical skills can produce increasingly complex outputs.

However, data itself does not create value until it is transformed into information. As Ackoff (1989) states, data can be defined as symbols that describe the properties of some events without explaining them in more detail. He also explains that, through processing, data is made more usable, and it becomes information. Thus, information differs from data in terms of its functionality

rather than its structure as the information that is formed can be thought of as answering questions such as what, when and where (Ackoff, 1989). This demonstrates that leveraging AI to more efficient information creation is seen as a groundbreaking advancement.

Moreover, collecting and owning large and complex data sets can benefit companies, increasing their performance and effectiveness (Müller et al., 2018). However, it is not only enough to have this big data, but it is crucial to use it correctly, investing in state-of-art tools among others in order to get the benefits (Grover et al., 2018). Since data can exist in various formats, leveraging it with tools is not entirely straightforward. Usually, data can be divers as it contains different amounts of structured data.

As Bouquet et al. (2024) gives an example in their article, fintech data may be diverse including unstructured data like news articles, semi-structured data, such as financial reports, and structured data such as transaction records. Additionally, it is complex integrating these different data types and therefore efficient and accurate models for large volume of data is needed to improve analysis and decision-making (Bouquet et al., 2024).

As defined, unstructured data does not have a predetermined structure, including text documents, emails or images, which presents unique processing challenges in data analyses (Bouquet et al., 2024). A large portion of companies collected data is in an unstructured textual format like documents, webpages, emails or social media content (Chen et al., 2012). As it is difficult to draw direct conclusions from unstructured data using existing strategic methods (Bhimani, 2015) large language models (LLM) can be seen as a great addition in data analysis (Bouquet et al., 2024). Additionally, making unstructured data easier to interpret, with a help of information extraction these types of documents can be converted into structured data (Bouquet et al., 2024).

In addition, as Aaltonen and Penttinen (2021, p. 5922) explain in their paper, “If structured data is not available, data mining and machine learning techniques can sometimes be used to reconstruct a latent structure hidden in seemingly unstructured data”. However, they also argue that it is not all black and white what is structured, semi-structured and unstructured data. Where in other cases some data can be considered unstructured like a photo, in other cases it can be made into structured data with algorithms creating visual structures (Aaltonen & Penttinen, 2021).

Structured data is often easier to directly utilize using statistical models (Tayefi et al., 2021). However, regardless of form, all digital objects consist of sequences of bits. No matter how simple or complex the digital object is a

set of instructions is needed to it come into being which are computed (Baskerville, 2020). Therefore, in order to work, machines need interaction between technology and human user. When a human user performs a physical action, such as typing something on a keyboard, it creates a digital object, and this started input project is transformed to bitstrings with a help of lower-level machine-readable instructions (Baskerville, 2020).

However, as mentioned by Baskerville (2020) the interaction between technology and human is necessary, nowadays from a research perspective, an interesting observation is that theory could originate from a computer (Agarwal & Dhar, 2014). As more data is collected, computers can be used to create cause-and-effect relationships, as they are able to form new insights. As Agarwal and Dhar (2014) presented as an example, a computer can find an area in a large dataset where people are increasingly getting diabetes. Based on the data it can test different hypotheses regarding whether diabetes can be caused by, for example, diets or other factors that humans would not necessarily have been able to hypothesize (Agarwal & Dhar, 2014). Furthermore, they emphasized that this is powerful, highlighting that computer-based theory creation is already among us.

This positions data as a fundament for more accurate decision-making. However, more important than data is the actual information or hypothesis created based on the data available. As the amount of data increases, finding the relevant information to transform it to essential insights gets harder. Luckily, utilizing artificial intelligence and large language models serves a great potential to this problem, which has raised a lot of expectations.

2.2 Understanding XBRL: Formats, Taxonomies, and Applications

As aforementioned, big data can be combined in various forms. In this thesis XBRL formatted sustainability reports are utilized as a machine-readable data format for the study. It is a commonly used reporting format in sustainability reports and also coming to new sustainability reporting standards such as the CSRD (XBRL, 2025). For that reason, it is important to understand the characteristics of XBRL. The main feature will be presented more closely in this chapter.

As Hoitash et al., (2021) explained, eXtensible Business Reporting Language (XBRL) is a structured data format and an open standard for reporting structured financial information. They also mentioned that starting from 1996 companies were required to file their financial reports electronically by Securities and Exchange Commission (SEC). In 2005 XBRL adaption was first

voluntary and in 2009 companies were required to report their financial statements in it (Hoitash et al., 2021).

Moreover, in XBRL, taxonomy tags are predefined to represent specific concepts as a machine-readable dictionary. These tags, including both numerical values and TextBlock elements, categorize data into taxonomy or extended tags which allows the structured representation of information facilitating the extraction of detailed data (Hoitash et al., 2021). Thus, if the taxonomy does not meet the reporting needs of the company, it may create the need for the company to extend the taxonomy by creating its own extended tags (Hoitash et al., 2021).

According to Ramanan and Warren (2023), XBRL report formats vary depending on different use cases. They state that XBRL reports are machine-readable, and the standard defines four different formats at the moment. They also state that, originally, it was based on XML but over time the standard has developed and has now different formats as an answer of different needs. Any of these formats can be used with business reports even though they have been created to address various types of cases, as they explain. The XBRL reports can be provided in XBRL-XML, Inline XBRL (iXBRL), XBRL-JSON and XBRL-CSV formats (Ramanan & Warren, 2023).

Additionally, there is an XBRL taxonomy giving meaning for the report's facts and the same taxonomy is provided with each XBRL report format (Ramanan & Warren, 2023). In this study only the XBRL-XML and XBRL-JSON formats of the XBRL data are applied, hence only those main features are opened more closely in this chapter. In addition, the study uses a taxonomy file as it examines its benefits.

XBRL-XML is the original eXtensible Markup Language (XML) based format enjoying a strong community support (Ramanan & Warren, 2023). It is a markup language useful to present information of business reports in machine-readable format as defined by Koskentalo (2020). Additionally, Koskentalo explains that the information in XML-format is inside tags which are marked as "<element>value/element>" for information to be easy to handle by computers. Figure 1 shows an example of one of the reports used in the study (Adani Green). In the figure 1, the tags in XBRL-XML format are shown.

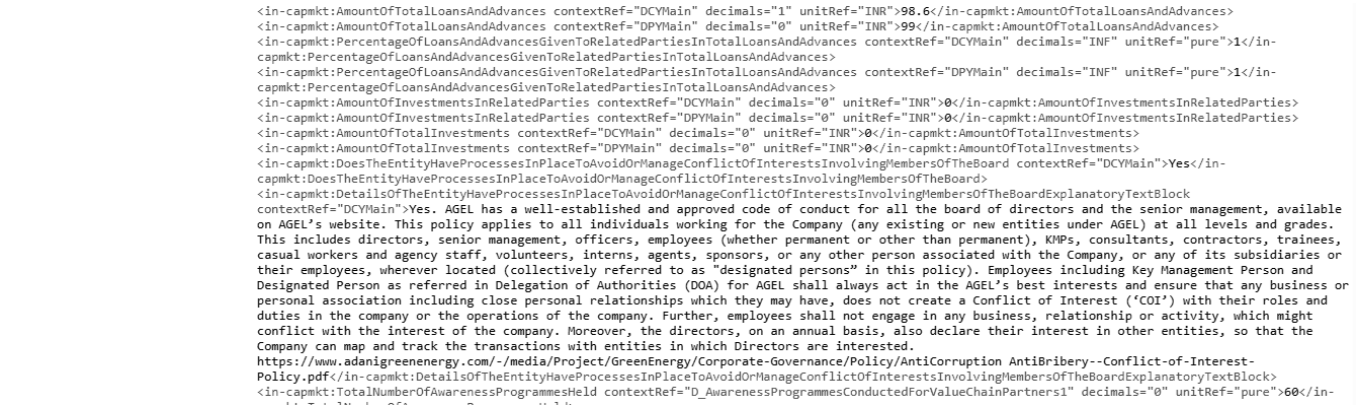


Figure 1. Example of XBRL-XML format. Snapshot from Adani green energy Limites BRSR XBRL-XML file for financial year 2024.

Moreover, according to Ramanan and Warren (2021), XBRL-JSON is programmatically the best format for consuming data from an XBRL report. This is caused because the values are linked with a dimension id which define the meaning of the value as shown in figure 2. XBRL-XML format can be converted into XBRL-JSON with XBRL software packages (Ramanan & Warren, 2021). Additionally, an example shown in figure 3 describes how the two XBRL formats differ in data formats such as ways of presenting same dates.

The extract below shows a representation of a simple monetary fact for *Assets*; in xBRL-JSON.

```
"f1": {
    "value": "1230000",
    "decimals": 0,
    "dimensions": {
        "concept": "eg:Assets",
        "entity": "lei:00EHHQ2ZHDCFXJCPCL49",
        "period": "2020-01-01T00:00:00",
        "unit": "iso4217:EUR"
    }
},
```

Figure 2. Representation of a fact in XBRL-JSON (Ramanan & Warren, 2021).

xBRL-XML		xBRL-JSON
Instant	2019-12-31	2020-01-01T00:00:00
Duration	start: 2019-01-01 end: 2019-12-31	2019-01-01T00:00:00/2020-01-01T00:00:00
Instant	2020-01-01T00:00:00	2020-01-01T00:00:00
Duration	start: 2019-01-01T00:00:00 end: 2020-01-01T00:00:00	2019-01-01T00:00:00/2020-01-01T00:00:00

Figure 3: The difference in period values with XBRL-XML and XBRL-JSON (Ramanan & Warren, 2021)

According to NSE (2025), in addition to XBRL, there is a taxonomy file that gives the meaning for facts in all XBRL reports. They state that the taxonomy and instance document make XBRL data machine-readable. Based on regulatory requirements, elements and relationships between them are created in the taxonomy as explained by NSE (2025). The taxonomy defines XBRL identifiers, in addition to which other relevant information such as time and unit of measure is included in the instance document (NSE, 2025).

Additionally, as stated by XBRL International (n.d.a) the taxonomy is defined as following: "A taxonomy links and defines a number of taxonomy components that provide the meaning for facts in an XBRL report. For example, a taxonomy for an accounting standard would include definitions of concepts such as 'Profit', 'Turnover', and 'Assets'". Usually, taxonomy is kept in a set of files which is upheld by a website (XBRL, n.d.a).

Most taxonomies are created, upkept and developed by policy makers and regulators (XBRL, n.d.b). However, organizations and projects can change and add tags according to their own needs as taxonomies are often built on international standards and baselines (XBRL, n.d.b). XBRL gathers globally

used common taxonomies on their official website, where they are available for the public (XBRL, n.d.b).

2.3 The Role of Generative AI and Large Language Models in Natural Language Processing

As the focus of this study is to examine the leverage of artificial intelligence (AI) in analysing Business Responsibility and Sustainability Reports, it is important to understand the underlying concept of Generative AI and Large Language Models in Natural Language Processing. This chapter gives the basics in these technologies and dives into some operational examples and outcomes.

Modern AIs like Large language models (LLM's) are based on converting textual input into "mathematical" representations and dividing words to smaller parts. The article written by Mikolov et al. (2013), that has been cited almost 50 000 times, explains how neural networks can be used in word representation (word embeddings). This means that text is divided into tokens (parts of text, not necessarily whole words) and created into vectors. They also define that the vectors represent the occurrence of the token in comparison to another. As a simplified example, they explain that the relationship with words "uncle" and "aunt" might be similar to the vector relationship with words "man" and "woman". With these token vector relationships, a model can be taught to "think" by looking at patterns based on vector locations (Mikolov et al., 2013).

For this thesis, the term AI represents mostly large language models, generative AI, and natural language processing (NLP), all relevant topics for this study. Feuerriegel et al. (2024, p.111) defines generative AI as so: "The term generative AI refers to computational techniques that are capable of generating seemingly new, meaningful content such as text, images, or audio from training data". As described by the popular math educator Sanderson (2024), GPT is short for Generative Pre-training Transformer. Moreover, he explains that the words "generative" presents the characteristic of a bot that generates text, "pre-trained" references to the learning stage where a model has been taught with a large amount of data, and "transformer" references to a neural network that is the core element in modern AI (Sanderson, 2024).

Furthermore, according to Feuerriegel et al., (2024) generative AI, such as ChatGPT, is based on generative modeling. They explain that it means it can create new content, such as text, photos, videos and conclusions, on top of predicting answers based on probabilities. The authors also define that unlike discriminative models that only divide answers into categories, for example like spam email, generative AI aims to understand the whole data and

its structure. As mentioned previously, the model is taught with huge datasets. They explain that the training data is given, along with the correct answers from which the model learns. After, the model creates deep neural networks and later, when tasks are given, it uses the learned complex patterns to generate answers to tasks (Feuerriegel et al., 2024).

Nowadays, the combination of algorithms and the advancement of large language models enable increasingly precise analysis of textual content. Generative AI models such as ChatGPT are changing the way we create text and analyze problems (Feuerriegel et al., 2024). Consequently, in both organizations and the research world, artificial intelligence and its utilization have increased in popularity, and thus there are many interesting examples related to them available from a variety of different fields.

The study by Hasan et al. (2019) applied NLP techniques to analyze textual data from Twitter (now X) and classify sentiments as negative or positive. The researchers used a pre-processing framework to clean the text and sentiment classification was performed using a combination of models with a logistic regression classifier. The study found that the proposed specifically developed and improved model achieved an accuracy of 85.25% which demonstrates its efficiency in sentiment analysis. This highlights the ability of language models to recognize nuances in tone as well as aspects such as objectivity, all of which are important factors also in report analysis.

Moreover, LLMs like GPT-4 offer improved services powered by NLP, enabling tasks such as language translation and text generation (Bouquet et al., 2024). With ChatGPT, users can also interact more easily with the fundamental LLM (Bang et al., 2023). Gregory et al. (2021) also argue that AI benefits from positive data network effects, a new network effect phenomenon. As AI tools have a lot of users, such as ChatGPT as researched in this thesis, the algorithms learn from user inputs and therefore improve. However, despite these advances and an increasing understanding in LLM's and the utilization of NLP's, the lack of transparency in neural networks weakens trust in AI (Bouquet et al., 2024).

Even so, this ease of use helps leverage artificial intelligence in the workplace. In the study by Brynjolfsson et al. (2025), researcher investigated how generative artificial intelligence can improve work efficiency. The study looked at the use of an AI tool that served as a conversational assistant for customer service representatives. The result showed that productivity increased by 15% when employees had the opportunity to use an AI assistant. Employees were therefore able to solve more customer problems per hour than without the help of AI. However, an interesting observation according to the study is that the efficiency increase was not evenly distributed among the employees, as

less experienced and lower-skilled customer service employees benefited more from it than experienced ones. This suggests that artificial intelligence systems can help widen the operating methods that more productive employees have (Brynjolfsson et al., 2025).

Kang et al. (2020) in turn, reviewed studies utilizing NLP to explore how textual data can contribute to advancing management theories across various disciplines. They highlight the potential of analyzing past research from different fields as a foundation for further NLP-based studies. Firstly, NLP seeks to transform natural language into a format that computers can process, afterwards it can be explored from two key perspectives which are Natural Language Understanding (NLU) and Natural Language Generation (NLG) (Kang et al., 2020).

According to their findings, within the information systems discipline, consumer behavior and organizational research have attracted the most interest in NLP applications. Research using NLP covers both individual and organizational levels, with methods often applied to extract information for measurement purposes. For instance, studies have used NLP to assess the personalities of customers, data analytics capabilities of companies and innovation potential (Kang et al., 2020).

Additionally, Kang et al. (2020) found that NLP can assist researchers in various fields by determining the sentiments positive, neutral or negative, making it useful for assessing tone, for example in articles. They also explain that text representation and term frequency-inverse document frequency (TF-IDF) remain the most widely used method for generating word vectors. This addresses how diversly NLP can already be utilized in diverse business-related activities, which creates expectations for conducting broader analyses using AI.

Moreover, as the amount of textual data and the potential of AI increases, interest in their effective utilization is growing. For example, the use of NLP on unstructured healthcare data has already raised considerable interest among researchers (Tayefi et al., 2021). Another study by Li et al. (2022) related to examining the use of NLP to harness unstructured data from electronic health records (EHR) showed that it can enhance data utilization from many different perspectives. In the study, it was noted that answering questions and supporting decision systems among other things can be developed using NLP. This can respond to healthcare staff queries about patient health information and thereby speed up decision-making and improve patient care (Li et al., 2022).

EHR data consists of clinical notes, diagnoses and laboratory values among other things, making it comprise both fixed numerical and categorical fields as well as free-form text, with about 80% being unstructured (Li et al., 2022). Therefore, from the perspective of this thesis, EHR data processing provides a good applicable perspective for Business Responsibility and Sustainability reports, as their content possesses similar characteristics consisting of both the written text but also including predefined categorized values and metrics.

To continue, Kang et al. (2020) also pointed out certain challenges that arise in applying NLP to specific disciplines. For example, accurately processing domain-specific terms in accounting presents a notable issue. Words such as "taxes" are assigned a sentiment orientation in lexicon-based sentiment analysis, despite not necessarily conforming to the research context. Additionally, while sentiment analysis is considered useful and is sometimes combined with deep learning techniques, determining whether a text conveys a positive or negative tone is often framed as a classification problem (Kang et al., 2020). Therefore, despite its benefits, it is important to be aware of the potential for errors to effectively identify inaccurate outputs in analyses using NLP.

This study is concluded using ChatGPT, and there are also challenges related to its analysis. Possibility of hallucinations and lack of transparency may distort the result (Salah et al., 2023). Moreover, for the scope of this thesis, it is useful to understand the differences in AI's use of structured and unstructured data. A study done by Smailhodžić and Oehmichen (2025), examines how AI can understand accounting information from annual reports and their research found that context format matters in error rates.

Additionally, as Tayefi et al. (2021) define, a characteristic of unstructured data, such as a text format, lacks a predetermined structure as defined also in section 2.1. Therefore, utilizing natural language processing (NLP) on unstructured data is not entirely straightforward as the text consists of a string of tokens and words must be separated from delimiters before processing, which can lead to incorrect word segmentation (Tayefi et al., 2021).

More closely Tayefi et al. (2021) also specified an example of this challenge whether something like "20.4 mg" should be kept as one token or split into two separate entities "20.4" and "mg" or even further into smaller parts. As a result, both the narrative text and numerical indicators appearing in reports may face challenges in being accurately read by AI models, compared to structured data where the indicator is already in a machine-readable format. This is important to consider when leveraging NLP models.

Furthermore, a study from Zou et al. (2025) presents ESGReveal, a methodology for extracting and analyzing ESG data from reports using large language models. The authors tested how unstructured data can be transformed into structured form using artificial intelligence tools to improve analysis efficiency. As a result, they found that the ESGReveal framework showed the utility of automated methods in processing unstructured ESG reports which improved the accuracy and efficiency of ESG disclosure assessments. More closely, the study found that 76.9% accuracy was achieved in data extraction and 83.7% in ESG disclosure classification with GPT-4. This demonstrates both the interest in and the need for using artificial intelligence to analyze sustainability reporting more effectively as well as the importance of structured data formats in that process.

In summary, the development of AI's has changed how we analyse and produce data and conclusions. It can be a useful tool when analysing vast amounts of data and has proven to be useful for wide variety of different purposes. It is however important to have some sort of critical thinking as their lack of transparency, probabilistic nature and sensitivity to input formats also introduce risks.

2.4 Sustainability reporting

"Sustainability reporting is measuring, disclosing, and being accountable to internal and external stakeholders for organizational performance toward sustainable development." (Agama & Zubairu, 2022, p. 32).

Over time, companies have realized that it is no longer enough for businesses to just report on financial matters; the future lies in comprehensive reporting that takes sustainable development into account in reporting (Brondoni & Plata, 2022). It's about competitiveness and gaining a competitive edge through comprehensive reporting (Brondoni & Plata, 2022) some experts even stating that "sustainability reporting will now be on an equal footing with financial reporting" (European Commission, 2022). Governments and Unions have come up with their own sustainability frameworks to further this trend (Securities and Exchange Board of India, 2021; European commission, 2022). Overall, there has been a strong upward trend in sustainability reporting and related research, especially in the last ten years (Nasreen et al., 2023).

In terms of mitigating climate change, which is one of the biggest challenges globally at the moment, reducing greenhouse gas emissions and increasing the circular economy are important measures (Janik et al., 2020). This makes sustainability reports an optimal choice for examining companies' operations related to these themes.

In this chapter two frameworks, India's Business Responsibility and Sustainability Reporting (BRSR) (chapter 2.4.1) and EU's Corporate Sustainability Reporting Directive (CSRD) (chapter 2.4.2) will be discussed. The focus is on these two as the research analyses BRSR reports using a LLM and draws conclusions on how the results give insights that can be applied to the very new, still not fully implemented CSRD in EU. For that reason, it is crucial to understand both sustainability reporting frameworks.

2.4.1 Indian Sustainability Reporting BRSR

Although India's emissions relative to its population are small, in general, India was the third-largest producer of carbon dioxide emissions in 2023 (Ritchie et al., 2024). India's approach to reducing emissions, tracking them, and consequently reporting them is therefore crucial in the global picture.

According to PwC (2021), in 2009 Ministry of Corporate Affairs issued National Voluntary Guidelines (NVGs). They also explain that quick after, in 2012, the Securities and Exchange Board of India (SEBI) made 100 top listed companies in India report a Business Responsibility Reports (BRR) as a part of their annual report. In 2015 this was extended to include 500 largest listed companies and in 2019 to 1000 (PwC, 2021).

Moreover, In 2018, IICA and UNICEF conducted a joint study (PwC, 2021). The study revealed gaps in the Business Responsibility Reporting (BRR) framework previously developed by SEBI, as the information provided by companies was not entirely clear and accurate (PwC, 2021). To address these issues and enhance transparency in support of sustainability goals, the Indian Institute of Corporate Affairs (IICA) formulated and developed the BRSR in collaboration with UNICEF (PwC, 2021). Consequently, the BRSR was designed to improve non-financial reporting by companies (PwC, 2021).

This lead in May 2021 SEBI to introduce the Business Responsibility and Sustainability Reporting (BRSR), an integrated sustainability reporting framework, which requires the 1,000 largest publicly listed Indian companies to report in accordance with it according to Securities and Exchange Board of India (2019 & 2021). They state that companies were given time to prepare for the new BRSR, implementing it gradually. During the financial years 2021-2022, it was voluntary but encouraged, and from the financial year 2022-2023 onwards, it became mandatory for those 1,000 largest listed companies determined by market capitalization (Securities and Exchange Board of India, 2021).

Securities and Exchange Board of India (2021) defines that the objective of BRSR is to facilitate the comparability of reports across different companies,

sectors and financial years. Additionally, according to them, it aims to increase the use of quantifiable indicators and metrics that are easy to track and compare analytically. They also state that the various reported data and figures illustrate the sustainability of production chains and corporate business practices.

As PwC (2021) states companies are required to report in BRSR the KPIs that are in alignment with the nine principles of the National Guidelines on Responsible Business Conduct (NGRBC). The nine principles are Integrity, Sustainable and Safe, Well-being, Responsive, Human Rights, Environment, Responsible, Inclusive Growth and Provide Value (PwC, 2021).

The BRSR targets are in line with all the three ESG pillars Environmental, Social and Governance as stated by PwC (2021). They explain that in environmental reporting this is seen as reporting emissions, energy consumption, water usage, waste management and use, life cycle assessments, 3Rs (reducing waste, reusing and recycling resources and products) and extended producer responsibility. Furthermore, they explain that in the social aspect, BRSR focuses on employee well-being and training, assessing social impact and employee equality, and other corporate social responsibilities. Lastly, in Governance BRSR focuses on corruption, conflicts, stakeholder engagement and follow through of policies according to PwC (2021). BRSR has also strong links with Sustainable Development Goals SDGs (PwC, 2021).

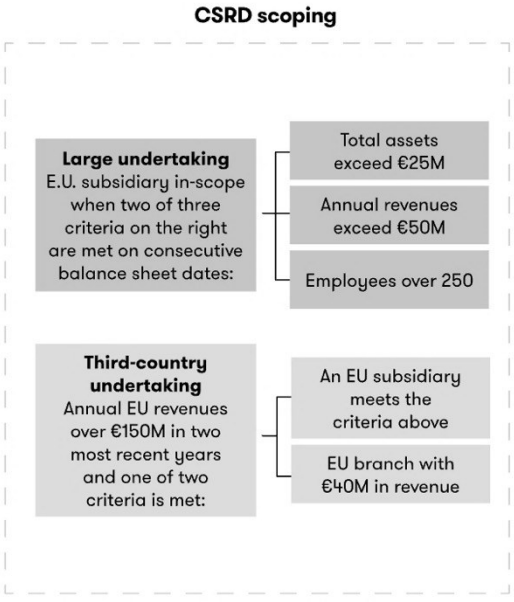
As EY (2023) states, BRSR reports have to be published in an annual report and 1 MCA21 portal through 2 XBRL languages. The BRSR report focuses in both quantitative and qualitative disclosures, including 140 questions (98 mandatory and 42 leadership ones) (EY, 2023). Additionally, as Design My Report (2021) explains there are two formats, one comprehensive one for large companies that already list in Listing Regulations, and the Lite that requires less information for companies new to sustainability reporting

Moreover, according to a study executed by PwC (2024) a bit over half of the companies disclosed their scope 3 emissions (financial year 2023) and 44% did a life cycle assessment on their product services. According to the reports and their results PwC (2024) found that 49% of companies increased renewable energy consumption and reduces their scope 1 (29%) and scope 2 (29%) emissions. 89% reported their leadership indicators and 31% reported net-zero targets (PwC, 2024).

2.4.2 EU Context: CSRD

Sustainability reporting is becoming mandatory for most companies in the EU (European Commission, 2022). On April 2021 the European Union published a proposal on Corporate Sustainability Reporting Directive (CSRD) from which we will be seeing the first obligatory sustainability reports in 2025, reporting from financial year 2024 (European commission, 2022). CSRD aims to develop the European Union towards a more sustainable economy (Baumüller and Grbenic, 2021), increase transparency and align with the Green Deal.

The new CSRD will be implemented in phases depending on the size of the company, whether the company is listed, and whether it operates within the EU (Thornton, 2024). Originally the phases were as presented in figure 4. Midst writing this thesis, on 11 February 2025 a new proposition from the EU commission called the Omnibus package emerged (European Commission, 2025). The proposition will reduce and postpone the scope of companies reporting concerns and the amount of data needed to report and "the Commission has a clear target to deliver an unprecedented simplification effort, by achieving at least 25% reduction in administrative burdens and at least 35% for SMEs before the end of the mandate" (European Commission, 2025).



Phase-in of CSRD reporting

Entity type	Financial year	Reporting in
Entities already subject to the NFRD, which includes entities with listed securities and more than 500 employees	2024	2025
Large undertakings (see first set of criteria on the left), not currently subject to NFRD	2025	2026
Small and medium-sized undertakings (except micro undertakings) listed on an EU-regulated exchange, small and non-complex credit institutions, and captive insurance undertakings	2026	2027
Third-country undertakings with net turnover over €150 million in the EU	2028	2029

Figure 4: Phase in CSRD reporting for different companies (Thornton, 2024)

Companies must report according to the ESRS (European Sustainability Reporting Standard), and the standards in ESRS apply regardless of the industry or sector they operate in (European Commission, n.d.). The ESRS

requires companies to report both their impact on the environment and people as well as the outer environmental and social impacts on the company (European Commission, 2023), which is called double materiality. The European Commission has also made sure that there is a high level of alignment between the Global Reporting Initiative (GRI) taking foundation in the Global framework (European Commission, 2023). This enhances coherence and helps to standardize the reporting format globally.

Additionally, European Commission (2023) explains that the ESRS is divided into crosscuttings (ESRS 1 and 2), environment (ESRS E1-E5), social (ESRS S1-S4) and governance (ESRS G1). Companies analyze themselves which topics are material and relevant to them (European Commission, 2023).

As the CSRD is still changing and new proposals are being made (European Commission, 2025), it is crucial to understand the most effective ways to shape the directive. It is important to understand and learn from previous frameworks such as BRSR and uncover insights that can be brought to the new CSRD. In BRSR XBRL format is already in use whereas in CSRD it is still a proposal. This thesis aims to contribute to that process by examining the BRSR to uncover insights that could inform the development and implementation of CSRD.

3 Research methodology

This chapter describes the methodology for the study. Chapter 3.1 describes the research approach and goes through the research in short. Chapter 3.2 explains how and why the data is selected, converted and narrowed. The data source, different data file formats, and processing of the data are discussed. In chapter 3.3 the systematic way of executing the study is gone through in detail. Chapter 3.4 opens how the results are evaluated, and chapter 3.5 discusses limitations and constraints in the study.

3.1 Research Approach

This study analyzes how OpenAI's ChatGPT-4o Plus and Pro processes some of India's largest sustainability reports in the energy sector across three file formats: PDF, XBRL-XML and XBRL-JSON. The 19 selected Business Responsibility and Sustainability Reports (BRSR) from the National Stock Exchange of India, were used to test the LLM's ability to extract, calculate, visualize, rank and categorize information from the reports. Standardized prompts are repeated three times, and results were then assessed looking at correctness and consistency. The study highlights differences in file format, question types and use of Chat GPT 4o Plus and Pro.

For the analysis, we utilize OpenAI's ChatGPT-4o (both in Plus and Pro version), as it is well-suited for processing and analysing documents. Consequently, the study is conducted within the capacity and analytical capabilities of ChatGPT at the time of research. The study examines the content of selected sustainability reports using the LLM. Based on both successful and unsuccessful results, a conclusion for the most effective approach to analysing sustainability reports are constructed while highlighting the major shortcomings in LLM-based analysis.

The GPT-4o version of ChatGPT (March 2024 release) is used in this study. OpenAI does not always disclose the exact version number to users. ChatGPT 4o was, at the time of conducting the study, the model that could process all format types and was available for all ChatGPT versions. OpenAI describes the GPT-4o model as follows: "GPT-4o is an autoregressive omni model, which accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. It's trained end-to-end across text, vision, and audio, meaning that all inputs and outputs are processed by the same neural network. " (OpenAI, 2024a).

The focus is on analysing numerical data by asking GPT to extract, create, visualize and rank data from a single or multiple sustainability reports as well as categorizing text.

3.2 Data selection

This study is based on Business Responsibility and Sustainability Reports collected by NSE - National Stock Exchange of India Ltd. PDF and XBRL formats. The research data comes from the National Stock Exchange of India Ltd. (NSE), which hosts 1,182 Business Responsibility and Sustainability Reports (BRSR). The reports are from annual year 2023-2024 from India's largest companies. These reports are available in both PDF and XBRL-XML formats. Both data types are incorporated into the study to compare differences in the LLM's ability to analyze reports across these unstructured and structured formats. The XBRL-XML is also converted to XBRL-JSON format, which is detailed in chapter 3.2.1.

This research leverages NSE reports while keeping in mind the upcoming European Corporate Sustainability Reporting Directive (CSRD). India's sustainability reports serve as a valuable data source for testing AI's capability in analyzing corporate responsibility disclosures, making this study relevant for broader global applications in AI-driven sustainability reporting analysis.

Selecting the industry and reports

To ensure that the analysis is as accurate as possible and the categories are the same across reports, the selectable reports are limited to the same industry. The energy sector has a major environmental impact as in 2018 it caused over 83% of GHG emissions in the EU (Janik et al., 2020). Therefore, in this study companies have been limited to those operating in the energy sector.

Companies were initially filtered from all reports by selecting those whose names contained specific keywords related to the energy sector. This was done by importing the company names into Excel and filtering the list based on the presence of the desired words. Next, the filtered list, which initially included over 50 companies, was reviewed using ChatGPT-4o to determine whether they truly belonged to the energy sector. Based on this assessment, 10 companies were excluded, as their activities were found to be unrelated to the energy industry despite their names being suggested otherwise.

The remaining 40+ reports were then reviewed again by asking ChatGPT to identify which companies operate in the energy sector. The results were further verified by cross-checking manually the companies' business activities online. Through this process, a final selection of 19 energy sector companies

operating in India were identified and used as the dataset for this study which are shown in table 1. The aim was to have a diverse group of energy industries sustainability reports, with focusing on choosing companies specialized in different areas. these areas different renewable and non-renewable energy production as well as energy transmission.

A maximum of 20 files can be attached to the project folder in ChatGPT 40. Therefore, in this study, the number of reports is limited to 19 to allow the inclusion of an additional file containing the XBRL taxonomy. Additionally, ChatGPT indicates that large number of files may reduce the quality of the responses.

Company	Energy Industry
KPI Green Energy Limited	Renewable energy (solar)
Torrent Power Limited	Power generation, transmission, distribution and
Power Grid Corporation of India	India's largest Electric Power Transmission Utility
Adani Green Energy Limited	Renewable Energy
Confidence Petroleum India Limited	LPG and CNG Cylinders
Inox Wind Energy Limited	Renewable Energy (Wind)
Suzlon Energy Limited	Renewable Energy (Wind, solar)
Adani Power Limited	Thermal power
Coal India Limited	Coal Mining and Production
NTPC Limited	Generation and distribution of electricity
Indian Oil Corporation (IOC)	Oil refinery and distribution
Tata Power	Renewable solutions and electricity Transmission & Distribution
Petronet LNG Limited	Oil and gas company that builds LNG terminals
Oil and Natural Gas Corporation (ONGC)	Oil & Natural Gas Exploration and Production
Oil India Limited	Oil & Gas Exploration and Production
Reliance Industries Limited (RIL)	Oil to Chemicals (O2C), Gas
Bharat Petroleum Corporation Limited (BPCL)	Oil & Gas Refining
RattanIndia Power Limited	Thermal Power
Savita oil Technologies Limited	Petroleum specialties

Table 1: Companies used in the study and their specified focus on the energy sector.

Decisions of data formats

The study is implemented by using three different formats that have the same information, as well as taxonomy file to explain the facts in XBRL -files. These data formats in the study are PDF, XBRL-XML, and XBRL-JSON.

As mentioned, both machine-readable XBRL data formats and more human readable PDF format are researched in this study. Portable document format (PDF) is a commonly used human readable reporting format (Chao & Fan, 2004). Companies publish annual and sustainability reports to read in PDF format in CSRD and BRSR, both large sustainability reporting frameworks. The PDF files, as well as XBRL-XML files were published in the National Stock Exchange of India (NSE) official database. As one of the research questions in this thesis is to analyse how human readable and machine-readable file formats analysis with ChatGPT differ, the PDF and XBRL are a good fit for the purpose.

XBRL-XML is the original XBRL data form and the file format in which business and responsibility reports are given in NSE. For data collection XBRL-XML is, according to Ramanan and Warren (2023), the default choice for the new closed reporting requirements. This Closed reporting requirement means that the data points are prescribed fully by the collector (Ramanan & Warren, 2023). In this study XBRL-XML was chosen as it was the original machine-readable file format given in NSE web site.

XBRL-JSON on the other hand is a simple representation of XBRL data and preferred choice for publishing (Ramanan & Warren, 2023). Moreover, XBRL-JSON format was used in other related study from Ramanan (2024b) where business reports were analysed with the help of ChatGPT. Additionally, taxonomy file is provided in this study for both XBRL-XML and XBRL-JSON formats as the effect of having descriptions for the facts is part of the study.

Files in each format are downloaded separately into project folders where analyses are conducted. In addition to the XBRL files, there is a taxonomy file that gives the meaning for facts in all XBRL reports. The effect of the taxonomy file in the LLM responses is also tested in this thesis. As is stated by NSE (2025) “XBRL makes the data readable, with the help of two documents - Taxonomy and instance document. Taxonomy defines the elements and their relationships based on the regulatory requirements.”. Therefore, both ChatGPT’s performance when taxonomy is and is not given are tested.

3.2.1 Methods for Data Collection and Processing

Selected reports have been downloaded into the project folders, which are used throughout the research process. The reports are saved separately in dedicated project folders for PDF, XBRL-XML, XBRL-XML + taxonomy, XBRL-JSON, and XBRL-JSON + taxonomy formats. Inside each folder, a new conversation is created for each research question. Memory is turned off between conversations, ensuring that ChatGPT does not store or learn from prior discussions. The same folders and conversations are tested using both the Plus and Pro versions of ChatGPT.

The study focuses on 19 companies, as ChatGPT 40 can efficiently process only this number of files at a time according to OpenAI. The file amount is limited to 20 files. While it is technically possible to provide a larger number of files, such as PDFs, to ChatGPT in a zip file, issues arise in analyzing them effectively. Furthermore, to ensure comparability and to specifically focus on differences between data formats, all project folders have been created using individual files rather than merging all company files into one large file, as was done in previous research from Ramanan (2024b).

Additionally, the aim of this study is to explore how reports can be analyzed and insights generated without extensive data preprocessing or the need for coding skills. Despite the smaller sample size, the research offers guidance on whether similar methods could be scaled to larger datasets in the future, particularly when utilizing more powerful AI models.

The XBRL-XML files have been converted to XBRL-JSON format using Arelle, which is an open-source tool for XBRL data processing. With the help of taxonomy file, XBRL-XML structure could be accurately transformed into XBRL-JSON format in Arelle. This enables the analysis of how ChatGPT processes and interprets information from different structured machine-readable data formats. This supports forming a recommendation on which format is most suitable for various types of AI-assisted analysis.

3.3 Execution of the study

The research is executed with a systematic approach for each file formats. First the 19 different business reports are selected. The study is done with both ChatGPT Plus and Pro 40 versions. In both versions Project files are generated (PDF, XBRL-XML, XBRL-XML + Taxonomy, XBRL-JSON and XBRL-JSON + Taxonomy) in which the files are uploaded separately in a right format for the project.

For each research question, new chat is created where the short context is given and afterwards the research question is asked. This is repeated three times in each project in both GPT versions. After each trial the answers are checked and documented for analysis. Lastly the results from each file format for each question is compared and conclusions drawn. This process is shown in a research process flowchart (figure 5). More detail for each phase is described next.

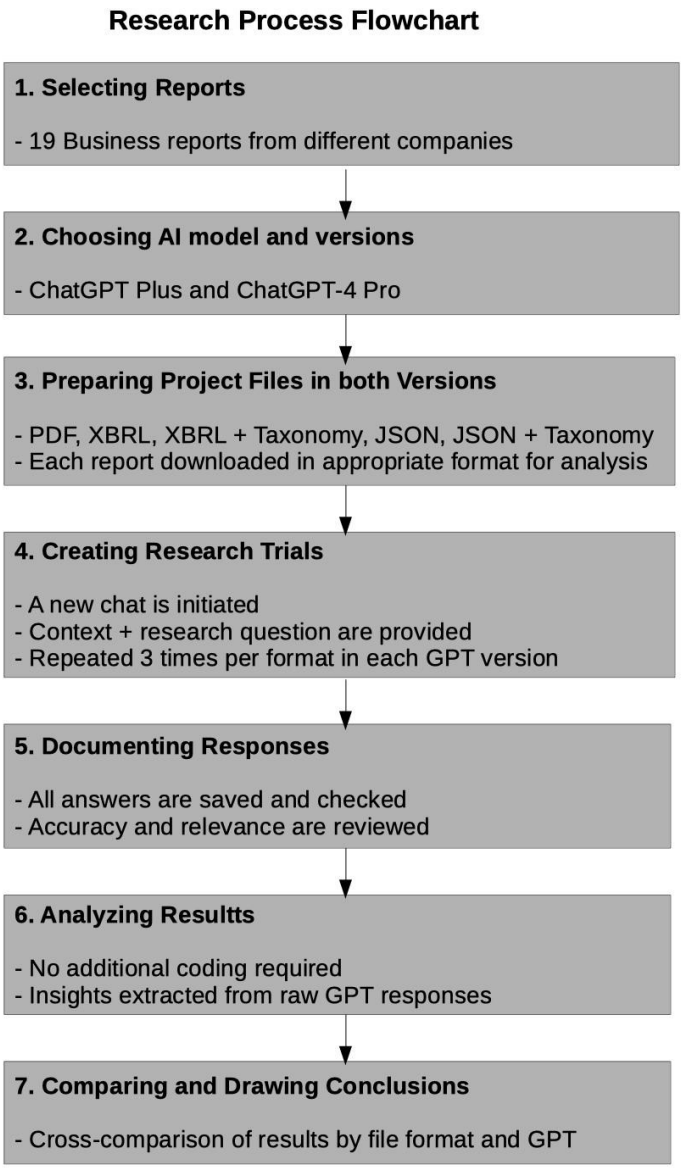


Figure 5: Research process flow.

After each new chat created there is a short context given in each chat before asking the question as shown on table 2. Each question is repeated three times in each project in both ChatCPT Plus and Pro. The questions are shown on table 3.

File format	Chat Context
PDF Files	There are 19 business sustainability and responsibility reports from different companies as PDF files in this project. Answer based on the project files.
XBR-XMLL Files	There are 19 business sustainability and responsibility reports from different companies as XBRL files in this project. Answer based on the project files.
XBRL-XML + Taxonomy	There are 19 business sustainability and responsibility reports from different companies as XBRL files and one taxonomy xlsx file for the XBRL files in this project. Answer based on the project files.
XBRL-JSON Files	There are 19 business responsibility and sustainability reports from different companies as JSON files in this project. Answer based on the project files.
XBRL-JSON + Taxonomy	There are 19 business sustainability and responsibility reports from different companies as JSON files and one taxonomy xlsx file for the JSON files in this project. Answer based on the project files.

Table 2: The context that is given in each new chat. Each file format has its own context that is identical except the file format name.

Each file format (with and without taxonomy) has its own project file, where all chats and tasks for the specific file type are executed. Project 1 has 19 Business Responsibility and Sustainability Reports as PDF files. Project 2 has reports as XBRL-XML files. Project 3 has reports as XBRL-XML files, accompanied by one taxonomy XLSX file providing schema details for the XBRL files. Project 4 has reports as XBRL-JSON files. Project 5 has reports as XBRL-JSON files, accompanied by one taxonomy XLSX file providing schema details for the XBRL-JSON files.

As previously mentioned, the memory has been turned off to make sure ChatGPT does not learn from previous conversations. Context is given, so that answer is based on given files (sustainability reports) in project folder in ChatGPT. However, the context is kept short and simple, so that it does not guide ChatGPT too much.

The capability of ChatGPT is tested by giving it five separate tasks. These five questions are shown in table 3. The aim is to test ChatGPT's ability to extract data (Question 1), calculate new values from existing values (Question 2), find and divide the correct data and visualize it (Question 3), rank companies based on specific values (Question 4) and ability to analyze and categorize text (Question 5). For each project, the same three questions are asked three separate times, each time in a new chat. This repetition allows to assess the

consistency, accuracy and robustness of the LLM's answers, still making it possibly to manually check answers when needed.

Each question is asked tree times in this study and this amount of repetition is selected in order to examine ChatGPT's performance capability, as it has been observed that it is not yet consistent. However, as each question is asked three times in each five project files, as well as five questions, and in two ChatGPT versions, the total answers analysed is 150. Therefore, the trials are limited to three as all answers are relatively long and time consuming to check from all 19 reports. In order to ensure reliable results, answers must be carefully verified for each individual tiral.

Question #	Question
1	Can you give the companies' total energy consumed from non-renewable energy sources for all the 19 project file companies separately?
2	Calculate how much each company has increased or reduced (%) their waste generation from FY 2022-23 to FY 2023-24.
3	Extract water withdrawal by source for Adani Power. Calculate ratio for each water source usage and generate a bar chart with data labels. Present the chart and a table.
4	Rank all companies based on the Water intensity per rupee of turnover in descending order.
5	Analyze question '8. Does the Company have any project related to reducing Green House Gas emission? If yes, then provide details.' in Coal India Limited report. Identify which initiatives have quantifiable outcomes, and which have non-quantifiable claims. List and categorize each initiative.

Table 3: Each question and the specific word by word phrasing of the task given to ChatGPT.

After each set of queries, the answers ChatGPT gives for each question and each (of the three) tries are recorded. The extracted values are compared to the actual data in reports. Finally, the number of answers generated are documented. Wrong answers are percentual portion of given answers. The comparison to ChatGPT answers and values in the report is done partially manually and partially with Python. The Python code extracts correct values from the machine-readable files.

This systematic evaluation enables a quantitative assessment of the accuracy rate and rate of wrong answers. It also enables comparison across questions, filetypes and ChatGPT Plus and Pro. To do this the averages and variances are calculated to see patterns and evaluate the performance through these different attributes (question type, file format, ChatGPT version). It evaluates AI's reliability and consistency throughout.

3.4 Constraints / limitations

Despite the consistent implementation of the study, some constraints are recognized related to research methods. One major constraint is the file upload limitation in ChatGPT project. The limitation is 20 files at the time this study is concluded. For that reason, in this study, it is limited to analyse 19 company sustainability reports in order to add an additional taxonomy file. This, however, makes sure consistency is put throughout, even though it still restricts the sample size.

The given prompts may also affect the outcome received from ChatGPT. The given short context and prompts might not translate to real-world prompting but rather examine what ChatGPT is capable of with short, straightforward prompts, that provide little information or background to the task. The effect of different prompting on the generation of the answers is not examined.

Additionally, we cannot be certain whether OpenAI will make any changes to ChatGPT's functionality, algorithms, efficiency, or any other factor that could affect its performance at the time of our study. OpenAI has a website that shows the release notes of ChatGPT. There are no major changes or upgrades to be noticed during the period this study was carried out. However, as ChatGPT is a closed-source system, it is not possible to be fully certain that its behavior will remain exactly the same throughout the duration of the study.

Moreover, an essential note is that this study is made looking at only ChatGPT 4o. It would be beneficial to compare other LLM's (Copilot, DeepSeek etc.) and GPT versions (GPT 3.5, 1o etc.) to gain a better understanding of LLMs' capabilities to analyse company reports. Despite the name OpenAI, ChatGPT is a closed source model and does not openly provide information on algorithms, training data or the architectures details. For that reason, it is hard to evaluate what other conditions affected answers (like time of day, traffic on site etc.).

However, as seen from the figure 6 of model evaluations (OpenAI, 2024b), the ChatGPT 4o shows best performance over text evaluation accuracy of compared models. Thus, is important to note that the comparison is from OpenAI's own pages and therefore the transparency cannot be fully confirmed.

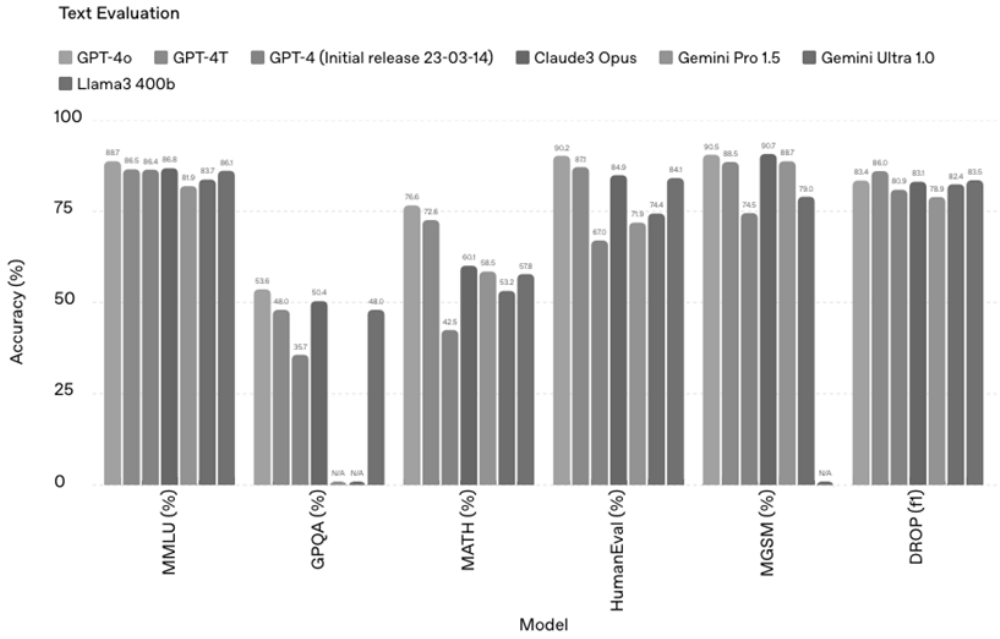


Figure 6. Comparison of Accuracy (%) of text evaluation between different models (OpenAI, 2024b)

4 Results

This section presents the results from the empirical part of the study. The research focused on evaluating how well ChatGPT-4o Plus and Pro perform when analysing BRSR reports from India's largest companies in PDF, XBRL-XML, and XBRL-JSON formats in different tasks. The focus is on how accurately ChatGPT-4o is able to extract, create and interpret information when given multiple reports.

The chapter is divided into three parts. 4.1 covers the first four questions that examine how GPT can extract data, create new metrics from existing ones, visualize data into graphs and rank companies based on a specific attribute. The 4.2 analyses how accurately GPT-4o interprets text, particularly performance in categorizing text's claims to quantifiable and non-quantifiable actions. In the 4.3 all results are summarized.

Overall, this chapter aims to provide a clear comparison of GPT's capabilities and limitations in sustainability reporting analysis, offering insights for companies, researchers and practitioners who rely on LLM tools to analyze sustainability reports and data.

4.1 Chat GPT analysing numerical data

Question 1 Data extraction

The first question tests how GPT performs when it is asked to extract a specific value from 19 separate reports and gather it together. The specific question given to ChatGPT after the context text was: "Can you give the companies total energy consumed from non-renewable energy sources for all the 19 project file companies separately."

Success rate (correct answers) for question 1, extracting 19 companies energy consumption data						
		Try 1 success rate	Try 2 success rate	Try 3 success rate	Average success rate (correct answers)	Variance correct answers*
PDF	**Plus**	68.42%	78.95%	78.95%	75.44%	1.33
PDF	**Pro**	57.89%	0.00%	84.21%	47.37%	67.00
XBRL-XML	**Plus**	100.00%	0.00%	0.00%	33.33%	120.33
XBRL-XML	**Pro**	36.84%	31.58%	31.58%	33.33%	0.33
XBRL-XML + TAX	**Plus**	100.00%	100.00	100.00%	100.00%	0.00
XBRL-XML + TAX	**Pro**	100.00%	100.00	100.00%	100.00%	0.00
XBRL-JSON	**Plus**	15.79%	10.53%	15.79%	14.04%	0.33
XBRL-JSON	**Pro**	5.26%	5.26%	5.26%	5.26%	0.00
XBRL-JSON +TAX	**Plus**	10.53%	10.53%	10.53%	10.53%	0.00
XBRL-JSON +TAX	**Pro**	10.53%	5.26%	10.53%	8.77%	0.33

*Variance is calculated from correct number of answers in absolutes not percentages

Table 4: The table shows the number of correct responses in percentages for each trial (Try 1, Try 2, Try 3), along with the average and variance of correct answers for each combination of file format (PDF, XBRL-XML, XBRL-JSON) and ChatGPT model version (ChatGPT Plus 4o, ChatGPT Pro 4o). For XBRL-XML and XBRL-JSON, the models were tested both with only the raw files and with an accompanying taxonomy file (TAX) that provides the structural metadata and explains the tags used in the sustainability reports.

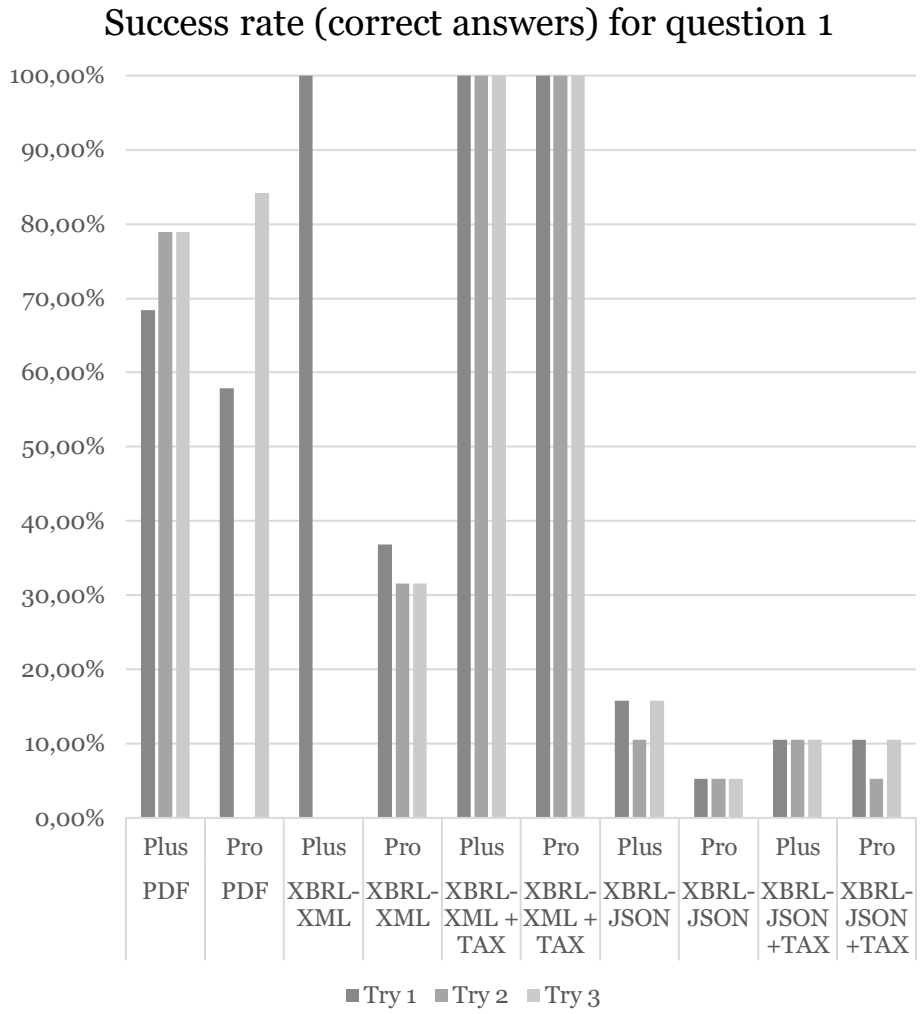


Figure 7: The Graph shows the number of correct responses in percentages for three separate tries for each combination of file format and taxonomy (PDF, XBRL-XML, XBRL-JSON) with and without taxonomy file (TAX) and model version (Plus, Pro).

As is shown in table 4 and figure 7, XBRL-XML with taxonomy performs best out of all file formats. It gives all 19 correct answers (100%) for all 19 companies values from their reports in both GPT Plus and Pro. The second-best performance comes from PDF with an average of 14.33 (75.4%) with GPT Plus and 9 (47.4%) with GPT Pro. All the other formats (XBRL-XML without taxonomy, XBRL-JSON and XBRL-JSON with taxonomy) achieve far lower rates with an average of less than 33% correct answers.

The largest variance between tries occurs at PDF in Pro and XBRL-XML without taxonomy in Plus having values 67,00 and 120,33 respectively as seen in figure 7. The variance is calculated with absolute values not percentages. This shows that ChatGPT performs very differently between trials. Additionally, even though the number of correct answers is lower with some file formats, as they give less answers, their variance became smaller as values

are less spread. This is due to the number of correct answers compared to the number of answers given being higher even though there are answers missing. This is important to notice as only looking at the variance may not give the comprehensives illustration of the performance.

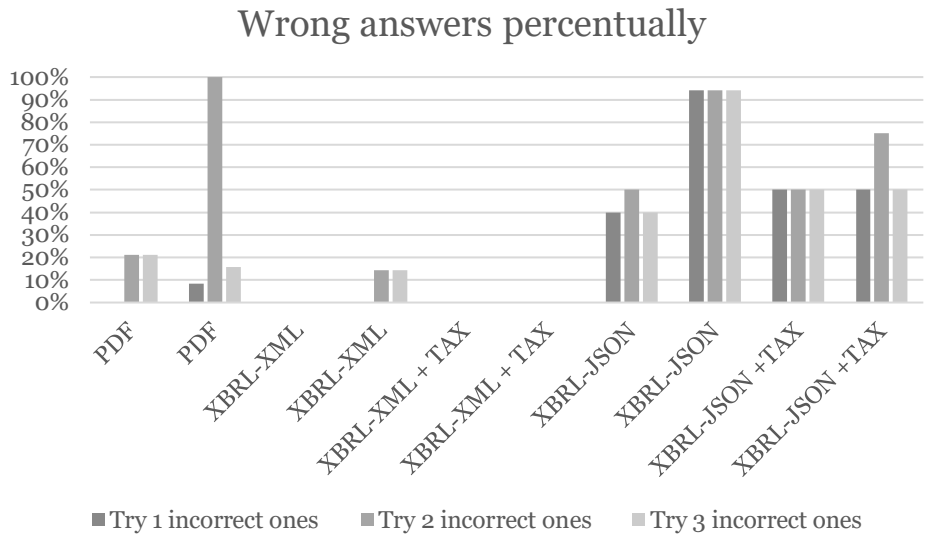


Figure 8: The figure shows the number of incorrect (or hallucinated) responses as a percentage of given answers for three separate tries for each combination of file format and taxonomy (PDF, XBRL-XML, XBRL-JSON), with and without a taxonomy file (TAX), and with two model versions (Plus, Pro). For example, if the model gives answers for 5 out of 19 companies, and 3 of those are incorrect, the error rate shown in the chart is 60% (3/5).

It is also important to understand not only the number of correct answers but also how many incorrect or hallucinated answers GPT gives for the task given. The number of incorrect answers percentual in figure 8 have been calculated with how many of all given answers are incorrect. In question 1, as can be seen in figure 8, XBRL-JSON and XBRL-JSON with taxonomy give incorrect answers far more. XBRL-JSON without taxonomy gives 16 (16/17 = 94.1%) incorrect answers while only 1 correct answer. XBRL with taxonomy has zero incorrect answers as output as it gave only correct answers.

There are some observations on what caused the incorrect answers that were seen for question 1. GPT can add a decimal point in the wrong place due to punctuation in original data, especially in PDF. In some answers GPT takes the value from wrong part of the report, for example when asked to give total energy consumed from non-renewable energy sources it gives total fuel consumed from non-renewable energy sources. In these situations, it is always a value with a name very close to the one asked and happened mostly in XBRL-XML and XBRL-JSON.

As a conclusion, when asking Chat GPT to find a value from a file, XBRL-XML + TAX performs the best. Even though the XBRL-XML alone consists of the same files as XBRL-XML + TAX, only the taxonomy file missing, they perform very differently. Additionally, to compare, XBRL-JSON files with or without taxonomy gave only a couple answers.

Question 2: value calculation

The second question tests how GPT performs in case of creating a new value from existing values. The specific question given to ChatGPT after the context text was: "Calculate how much each company has increased or reduced (%) their waste generation from FY 2022-23 to FY 2023-24." As is shown in table 5 different formats and trials perform very differently, which is seen also with high values of variance.

Success rate (correct answers) for question 2, calculating change in waste generation for 19 company reports						
		Try 1 Success rate	**Try 2 Success rate**	**Try 3 Success rate**	**Average success rate (correct answers)**	**Variance correct answers***
PDF	**Plus**	36.8%	47.4%	36.8%	40.4%	1.33
PDF	**Pro**	42.1%	42.1%	42.1%	42.1%	0.00
XBRL-XML	**Plus**	21.1%	21.1%	0.0%	14.0%	5.33
XBRL-XML	**Pro**	0.0%	42.1%	0.0%	14.0%	21.33
XBRL-XML + TAX	**Plus**	100.0%	100.0%	100.0%	100.0%	0.00
XBRL-XML + TAX	**Pro**	100.0%	0.0%	21.1%	40.4%	100.33
XBRL-JSON	**Plus**	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-JSON	**Pro**	100.0%	0.0%	0.0%	33.3%	120.33
XBRL-JSON +TAX	**Plus**	0.0%	0.0%	100.0%	33.3%	120.33
XBRL-JSON +TAX	**Pro**	5.3%	0.0%	0.0%	1.8%	0.33

*Variance is calculated from correct number of answers in absolutes not percentages

Table 5: The table shows the number of correct responses in percentages for each trial (Try 1, Try 2, Try 3), along with the average and variance of correct answers for each combination of file format (PDF, XBRL-XML, XBRL-JSON) and ChatGPT model version (ChatGPT 4o Plus, ChatGPT 4o Pro). For XBRL-XML and XBRL-JSON, the models were tested both with only the raw files and with an accompanying taxonomy file (TAX) that provides the structural metadata and explains the tags used in the sustainability reports.

Out of all format types XBRL-XML + taxonomy performs the best in this task as shown in figure 9 and table 5. ChatGPT Plus gives all 19 correct answers (100%) for all 19 companies values from reports in all trials. GPT Pro also gives 100% correct output on the first try but has an average of 7.67 (40.4%) correct answers as the other trials performed less well. The second-best results are given by PDF with an average of 8.00 (42.1%) with GPT Pro and 7.67 (40.4%) with GPT Plus (figure 9, table 5).

All the other formats (XBRL-XML without taxonomy, XBRL-JSON and XBRL-JSON with taxonomy) achieve a lower average of less than 33% correct answers. In PDF sometimes the final values are off by 0,01%-0,03%, in the scale insignificant but still an error, probably due to rounding in intermediate steps. In these situations, answers have, however, been evaluated to be correct.

XBRL-JSON with GPT Pro and XBRL-JSON + taxonomy with GPT Plus both give 100% correct answer with one try. However, other trials give 0% correct answers in both formats, creating a variance of 120.33. Additionally, XBRL-XML + taxonomy has a variance of 100.33 in GPT Pro. All these three formats (XBRL-XML + taxonomy, XBRL-JSON and XBRL-JSON + taxonomy) are therefore able to perform 100% in best case scenario when creating a new value but as their variance is high it indicates a large spread between the possible outcomes. XBRL-XML in GPT Pro has a variance of 21.33 as the best trial gives 42.1% of correct answers and two other tries give 0% of correct answers. Other formats have variances less than or equal to 5.33.

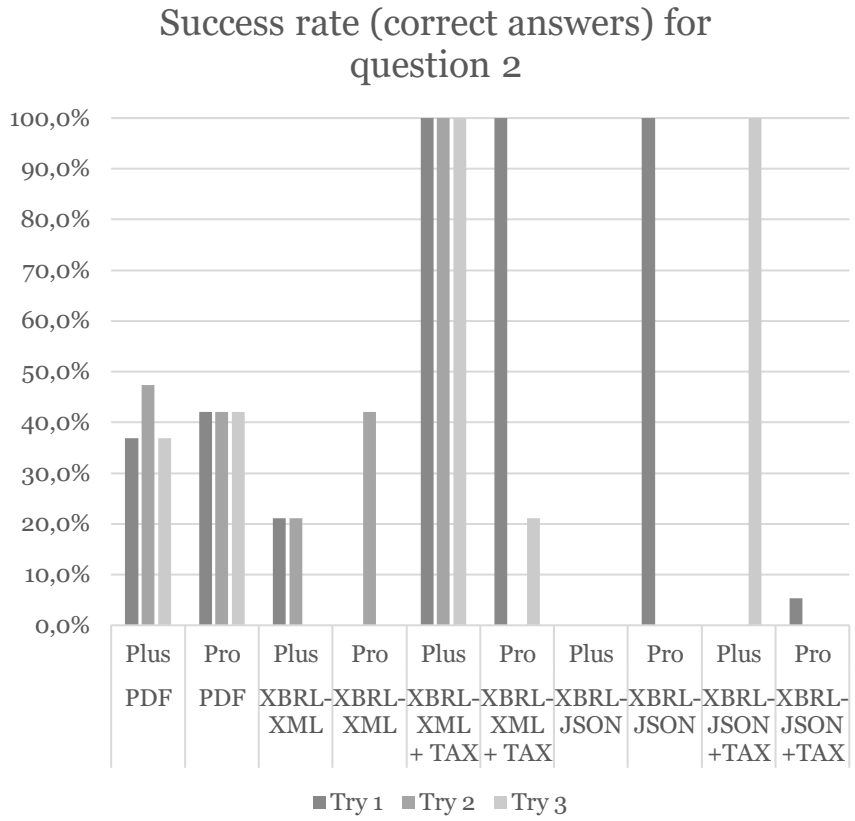


Figure 9: The figure shows the number of correct responses in percentages for three separate tries for each combination of file format and taxonomy (PDF, XBRL-XML, XBRL-JSON) with and without taxonomy file (TAX) and model version (Plus, Pro).

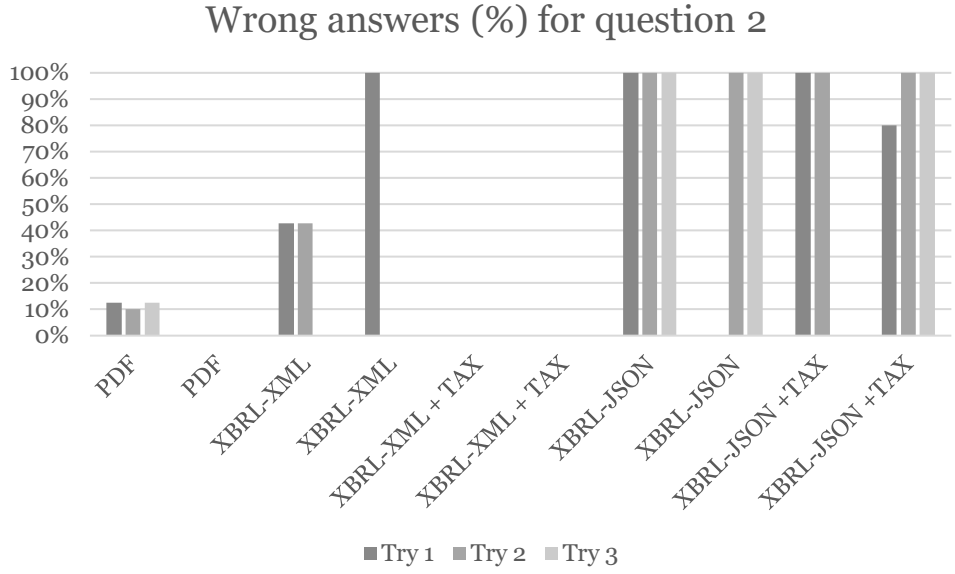


Figure 10: The figure shows the number of incorrect (or hallucinated) responses as a percentage of given answers for three separate tries for each combination of file

format and taxonomy (PDF, XBRL-XML, XBRL-JSON), with and without a taxonomy file (TAX), and with two model versions (Plus, Pro). For example, if the model gives answers for 5 out of 19 companies, and 3 of those are incorrect, the error rate shown in the chart is 60% (3/5).

As shown in figure 10. GPT gives also some incorrect or hallucinated answers for the task given. XBRL-JSON without and with taxonomy give incorrect answers far more. More specifically, XBRL-JSON without taxonomy gives mostly incorrect answers as despite the first attempt with GPT Pro, all the answers it gave were wrong. XBRL-JSON + taxonomy has the same kind of results. XBRL-XML with GPT Pro gave 19/19 (100%) hallucinated answers in first try and in third trial only an empty table with company names but no values were given. Additionally, PDF with GPT Plus has one incorrect answer in each try.

Some possible reasons for incorrect answers in question 2 were related to taking the values from wrong parts of the reports or from different years. Sometimes a similar but not exactly correct metric was used, such as total waste disposed instead of total waste generated or total energy consumed versus total fuel consumed and so forth.

Question 3 Data visualization:

The third question tests how GPT performs when asked to create visualizations, specifically bar charts from a single report's values. The specific question given to ChatGPT after the context text was: “Extract water withdrawal by source for Adani Power. Calculate ratio for each water source usage and generate a bar chart with data labels. Present the chart and a table.”

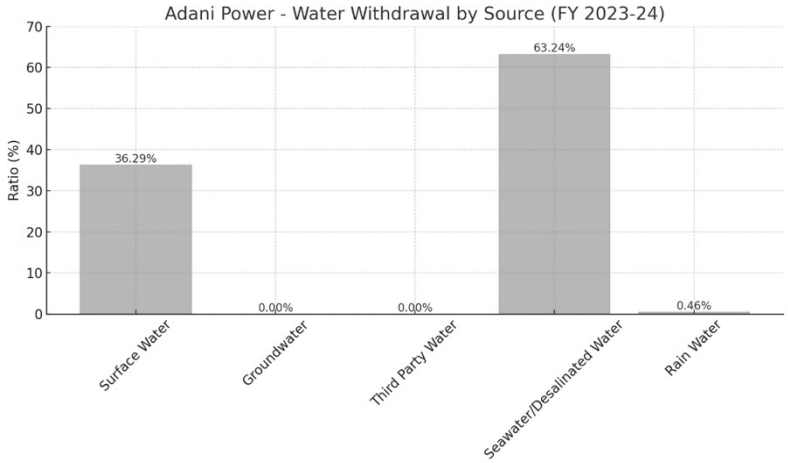


Figure 11: A typical data visualization from Chat GPT for question 3.

Success rate (correct answers) for question 3, creating bar chart from water sources						
		Try 1 Success rate	Try 2 Success rate	Try 3 Success rate	Average success rate (correct answers)	Variance correct answers*
PDF	**Plus**	100.0%	100.0%	100.0%	100.0%	0.00
PDF	**Pro**	100.0%	100.0%	100.0%	100.0%	0.00
XBRL-XML	**Plus**	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-XML	**Pro**	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-XML + TAX	**Plus**	0.0%	0.0%	100.0%	33.3%	8.33
XBRL-XML + TAX	**Pro**	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-JSON	**Plus**	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-JSON	**Pro**	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-JSON +TAX	**Plus**	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-JSON +TAX	**Pro**	0.0%	0.0%	0.0%	0.0%	0.00

*Variance is calculated from correct number of answers in absolutes not percentages

Table 6: The table shows the number of correct responses in percentages for each trial (Try 1, Try 2, Try 3), along with the average and variance of correct answers for each combination of file format (PDF, XBRL-XML, XBRL-JSON) and ChatGPT model version (ChatGPT Plus 4o, ChatGPT Pro 4o). For XBRL-XML and XBRL-JSON, the models were tested both with only the raw files and with an accompanying taxonomy file (TAX) that provides the structural metadata and explains the tags used in the sustainability reports.

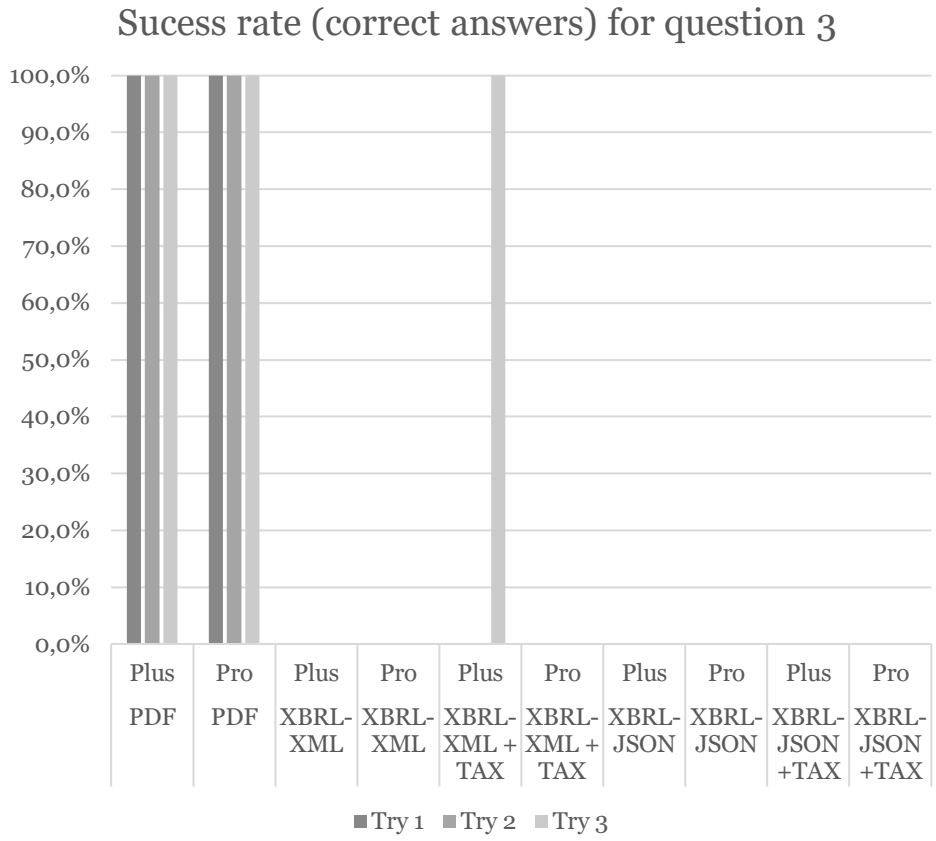


Figure 12: The figure shows the number of correct responses in percentages for three separate tries for each combination of file format and taxonomy (PDF, XBRL-XML, XBRL-JSON) with and without taxonomy file (TAX) and model version (Plus, Pro).

As seen in table 6 and figure 12, PDF format performs by far the best in this third question. It succeeds in getting all correct ratios, well visualized in all three separate tries (100% correct). For all the structured reports, XBRL-XML and XBRL-JSON with and without taxonomy, the performance is poor and there occurs only one time where XBRL-XML with taxonomy gets correct answers. XBRL-XML without taxonomy, XBRL-JSON and XBRL-JSON with taxonomy all get 0% correct answers.

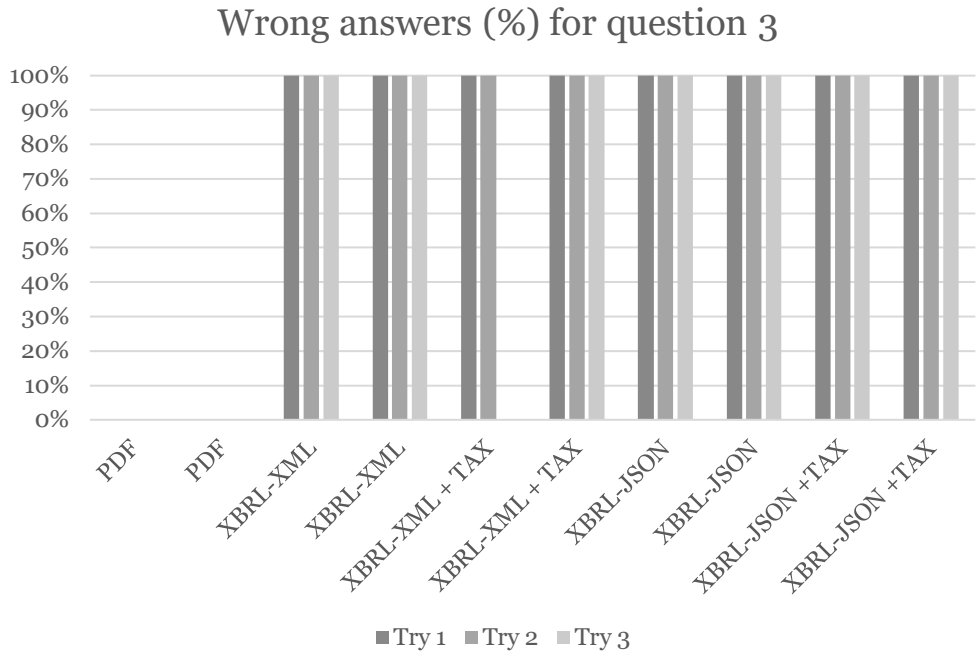


Figure 13: The figure shows the number of incorrect (or hallucinated) responses as a percentage of given answers for three separate tries for each combination of file format and taxonomy (PDF, XBRL-XML, XBRL-JSON), with and without a taxonomy file (TAX), and with two model versions (Plus, Pro). For example, if the model gives answers for 5 out of 19 companies, and 3 of those are incorrect, the error rate shown in the chart is 60% (3/5).

Though this success rate is almost 0%, it is important to understand that these incorrect answers are all due to confusing the general water source values with the water source values in areas under water stress. In other words, GPT returns the data that is labelled as water sources in the reports but in areas with water distress not the general info on water sources. As can be seen clearly in the PDF report there are two separate charts, one with general water source values and later in the file the same chart but for areas with water distress.

Therefore, there are both hallucinated answers as well as values found from wrong places. However, the visualization task itself is correctly executed based on the found or hallucinated values. Therefore, the result from the tasks shows more about the GPT's performance of generating answers based on asked values rather than creating a successful visualisation.

Something worth notable also is that the variance in the answers is smaller in question 3 than in 1, 2, and 4. Not only is the variance in the correct answers 0 in PDF, XBRL-XML, XBRL-JSON and XBRL-JSON with taxonomy, but the variance in wrong answers and the way GPT presents answers is a lot

smaller. Almost all correct and wrong answers were identical, which shows the consistency of ChatGPT's performance between trials in this task.

Question 4 Ranking companies:

The fourth question tests GPT performance to rank values. GPT is asked to rank all companies based on the water intensity per rupee of turnover in descending order. GPT needs to find the correct value from all files and change the values to the same unit for comparison. The specific question given to ChatGPT after the context text was: "Rank all companies based on the Water intensity per rupee of turnover in descending order."

Success rate (correct answers) for question 4, ranking 19 companies based on water intensity						
		Try 1 Success rate	Try 2 Success rate	Try 3 Success rate	Average success rate (correct answers)	Variance correct answers***
PDF	Plus	0.0%	0.0%	0.0%	0.0%	0.00
PDF	Pro	0.0%	0.0%	0.0%	0.0%	0.00
XBRL-XML	Plus	57.9%	36.8%	100.0%	64.9%	37.33
XBRL-XML	Pro	100.0%**	100.0%	100.0%**	100.0%	0.00
XBRL-XML + TAX	Plus	100.0%	100.0%	100.0%	100.0%	0.00
XBRL-XML + TAX	Pro	0.0%	100.0%	100.0%	66.7%	120.33
XBRL-JSON	Plus	100.0%*	0.0%	21.1%	40.4%	100.33
XBRL-JSON	Pro	100.0%	100.0%	0.0%	66.7%	120.33
XBRL-JSON +TAX	Plus	100.0%	100.0%	100.0%	100.0%	0.00
XBRL-JSON +TAX	Pro	0.0%	0.0%	0.0%	0.0%	0.00

** Does ranking correctly but values are incorrect (previous years values)*

** * Gives correct answers but table title has wrong unit (L/INR when values are KL/INR)*

*** Variance is calculated from correct number of answers in absolutes not percentages

Table 7: The table shows the number of correct responses in percentages for each trial (Try 1, Try 2, Try 3), along with the average and variance of correct answers for each combination of file format (PDF, XBRL-XML, XBRL-JSON) and ChatGPT model version (ChatGPT Plus 40, ChatGPT Pro 40). For XBRL-XML and XBRL-

JSON, the models were tested both with only the raw files and with an accompanying taxonomy file (TAX) that provides the structural metadata and explains the tags used in the sustainability reports.

As is shown in table 7 and figure 14, XBRL-XML with taxonomy performs best out of all format types. It gives all 19 correct answers (100%) for all 19 companies values from their reports in both GPT Plus and Pro excluding one try with Pro when zero values are given. This creates averages of 100% and 66.7% accuracy of correct answers respectively. Also, XBRL-XML without taxonomy performs well when ranking the values. GPT Pro gives all 19 correct answers (100%) for all 19 companies but as a side note the table title has the wrong unit (L/INR when values are KL/INR). GPT Plus gives also 100% correct answers on the third try but performs worse with only 58% and 37% correct answers on the first and second try, having 64.9% as an average outcome of correct answers. With PDF GPT gives 0% correct answers within both GPT Plus and Pro.

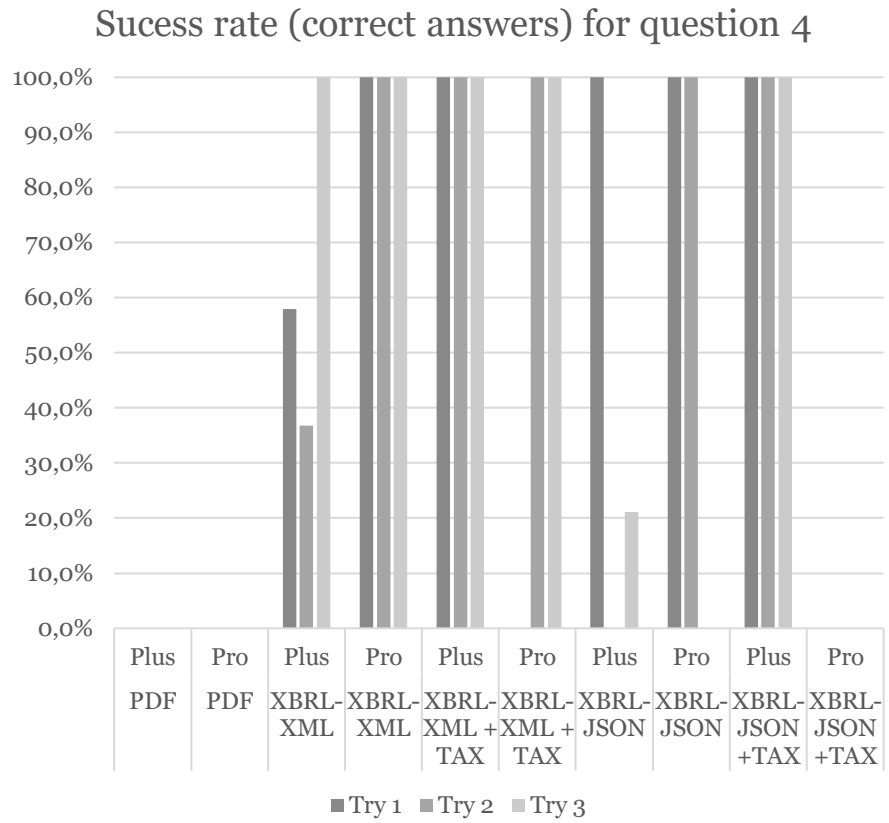


Figure 14: The Graph shows the number of correct responses in percentages for three separate tries for each combination of file format and taxonomy (PDF, XBRL-XML, XBRL-JSON) with and without taxonomy file (TAX) and model version (Plus, Pro).

Additionally, in the case of XBRL-JSON and XBRL-JSON + taxonomy GPT can rank companies in correct order, but values are incorrect as they are from previous year. However, in this task the answers were marked as correct as the ability to rank values was being analysed.

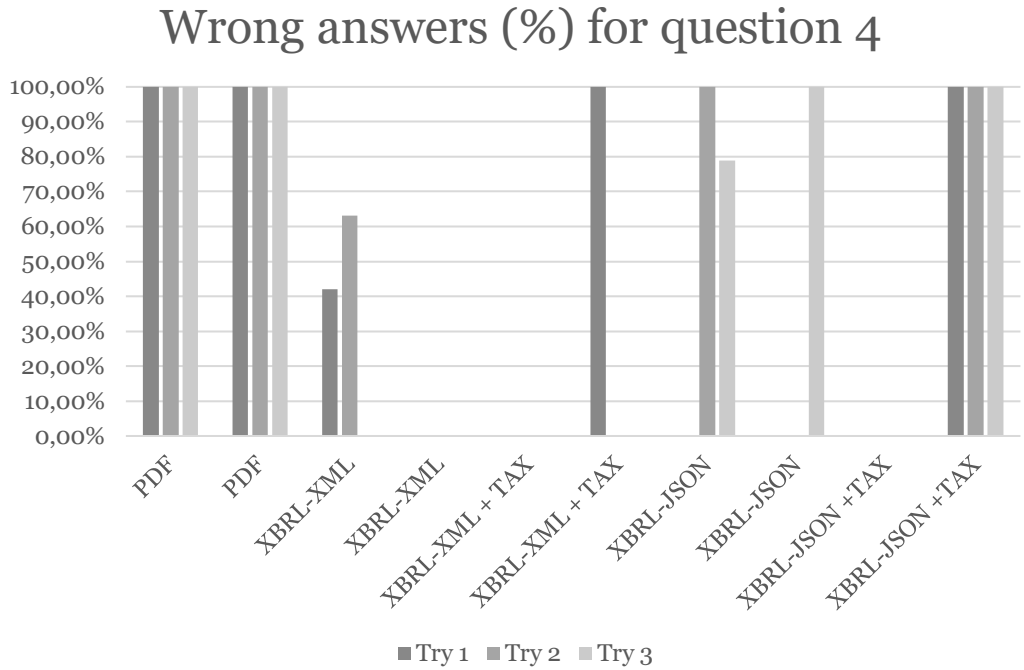


Figure 15: The figure shows the number of incorrect (or hallucinated) responses as a percentage of given answers for three separate tries for each combination of file format and taxonomy (PDF, XBRL-XML, XBRL-JSON), with and without a taxonomy file (TAX), and with two model versions (Plus, Pro). For example, if the model gives answers for 5 out of 19 companies, and 3 of those are incorrect, the error rate shown in the chart is 60% (3/5).

The largest variance between tries occurs at XBRL-XML + taxonomy Pro and XBRL-JSON Pro having values 120,33 as seen in figure 16. In both cases this is due as the GPT gives 100% correct answer 2/3 time of trials but 1/3 it gives 0%. This shows that GPT is once again performing very differently in different trials and the outcome depend on which of its best- or worst-case scenarios come true. Additionally, XBRL-JSON in GPT Plus has variance of 100,33 and XBRL-XML in GPT Plus has a variance of 37.33. As the other formats give only 0% or 100% correct answers the variance for them is 0.

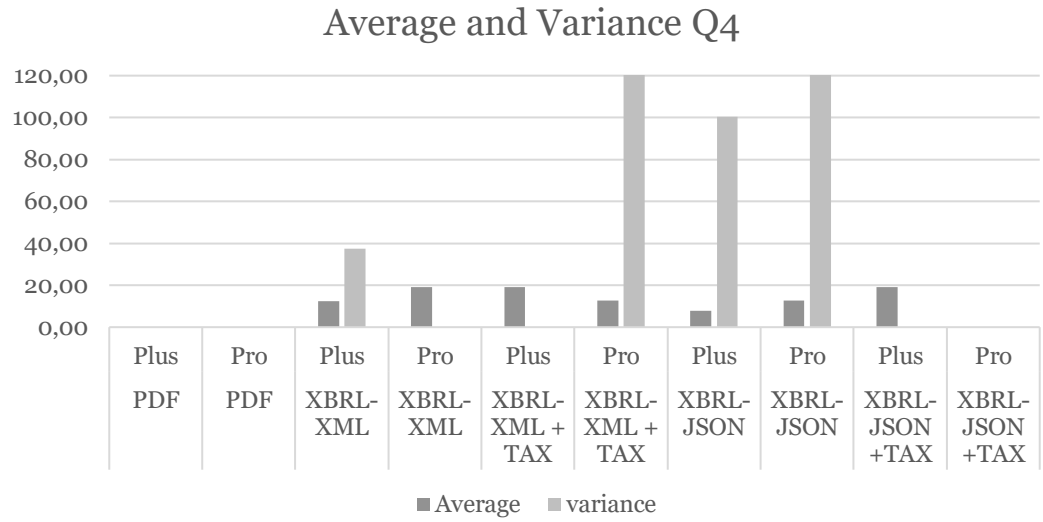


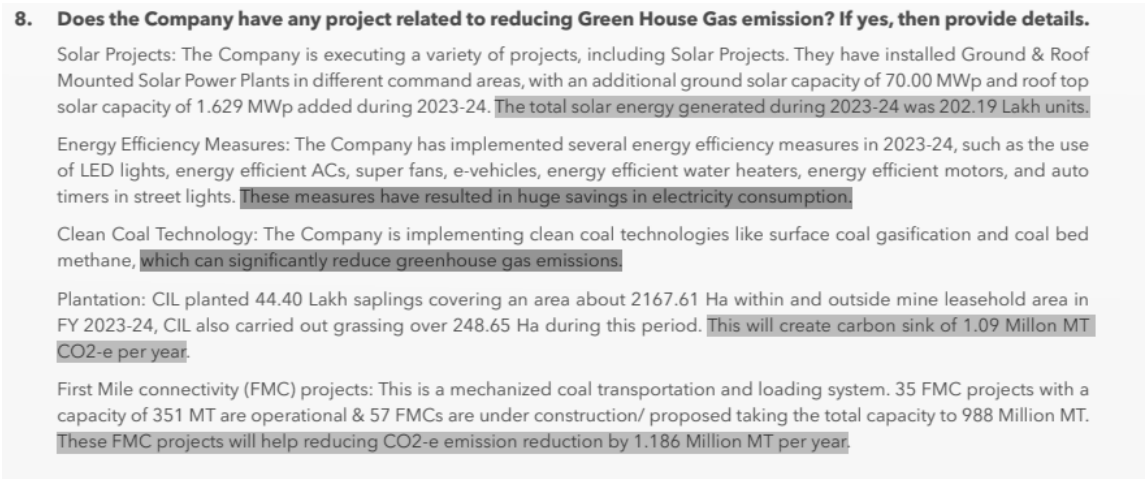
Figure 16: Variation of absolute values and average of success rate for each try (try 1, try 2, try 3) for all file types and GPT versions.

There is also some incorrect or hallucinated answers GPT gives for the given task as shown in figure 15. Some errors are due to values found from wrong parts of the files. In XBRL-JSON Chat GPT finds information for previous year water intensity not the current year (previous year is later in the XBRL-JSON file) in the first two tries. In third try it gives correct water intensity values but can not still rank companies correctly in descending order according to the correct numbers.

Additionally, In PDF the units are given in all different types (kl/INR, KL/Cror etc) and GPT does not correctly transform units. Like mentioned, sometimes it can transform units correctly, but it still ranks them incorrectly. Additionally, in XBRL Plus one try GPT recognizes correct values but still does ranking wrong between two values, claiming incorrectly e.g. 0.0004112759 > 0.0005201019.

4.2 Analysing text in reports

The fifth question tests how GPT performs in text analysis. GPT is asked to analyse text (shown in figure 17) in a specific part of the report and categorize the mentioned initiatives and divide them into one that have quantifiable effects and ones that have unquantifiable effects. The specific question given to ChatGPT after the context text was: “Analyze question '8. Does the Company have any project related to reducing Green House Gas emission? If yes, then provide details.' in Coal India Limited report. Identify which initiatives have quantifiable outcomes, and which have non-quantifiable claims. List and categorize each initiative.”



8. Does the Company have any project related to reducing Green House Gas emission? If yes, then provide details.

Solar Projects: The Company is executing a variety of projects, including Solar Projects. They have installed Ground & Roof Mounted Solar Power Plants in different command areas, with an additional ground solar capacity of 70.00 MWp and roof top solar capacity of 1.629 MWp added during 2023-24. The total solar energy generated during 2023-24 was 202.19 Lakh units.

Energy Efficiency Measures: The Company has implemented several energy efficiency measures in 2023-24, such as the use of LED lights, energy efficient ACs, super fans, e-vehicles, energy efficient water heaters, energy efficient motors, and auto timers in street lights. These measures have resulted in huge savings in electricity consumption.

Clean Coal Technology: The Company is implementing clean coal technologies like surface coal gasification and coal bed methane, which can significantly reduce greenhouse gas emissions.

Plantation: CIL planted 44.40 Lakh saplings covering an area about 2167.61 Ha within and outside mine leasehold area in FY 2023-24, CIL also carried out grassing over 248.65 Ha during this period. This will create carbon sink of 1.09 Millon MT CO2-e per year.

First Mile connectivity (FMC) projects: This is a mechanized coal transportation and loading system. 35 FMC projects with a capacity of 351 MT are operational & 57 FMCs are under construction/ proposed taking the total capacity to 988 Million MT. These FMC projects will help reducing CO2-e emission reduction by 1.186 Million MT per year.

Figure 17: Picture of Coal India Limited PDF Sustainability report which is analysed in this task. Quantifiable (a lighter highlight over the text) and non-quantifiable (a darker highlight over the text) measures for each initiative have been highlighted.

			Try 1 Success rate	Try 2 Sucess rate	Try 3 Success rate	Average success rate	Variance
PDF	Plus	Recognizes initiatives	100 %	100 %	100 %	100 %	0,0
		Correctly categorized	100 %	100 %	100 %	100 %	0,0
		No errors in claims	100 %	100 %	100 %	100 %	0,0
PDF	Pro	Recognizes initiatives	100 %	100 %	100 %	100 %	0,0
		Correctly categorized	100 %	100 %	100 %	100 %	0,0
		No errors in claims	100 %	100 %	100 %	100 %	0,0
XBRL-XML	Plus	Recognizes initiatives	60 %	100 %	100 %	87 %	1,3
		Correctly categorized	20 %	80 %	80 %	60 %	3,0
		No errors in claims	0 %	60 %	60 %	40 %	3,0
XBRL-XML	Pro	Recognizes initiatives	100 %	60 %	40 %	67 %	2,3
		Correctly categorized	80 %	20 %	20 %	40 %	3,0
		No errors in claims	60 %	0 %	0 %	20 %	3,0
XBRL-XML + TAX	Plus	Recognizes initiatives	60 %	80 %	80 %	73 %	0,3
		Correctly categorized	60 %	60 %	60 %	60 %	0,0
		No errors in claims	20 %	40 %	40 %	33 %	0,3
XBRL-XML + TAX	Pro	Recognizes initiatives	60 %	60 %	60 %	60 %	0,0
		Correctly categorized	20 %	20 %	20 %	20 %	0,0
		No errors in claims	0 %	0 %	0 %	0 %	0,0
XBRL-JSON	Plus	Recognizes initiatives	100 %	100 %	100 %	100 %	0,0
		Correctly categorized	80 %	80 %	80 %	80 %	0,0
		No errors in claims	60 %	60 %	60 %	60 %	0,0
XBRL-JSON	Pro	Recognizes initiatives	20 %	100 %	20 %	47 %	5,3
		Correctly categorized	20 %	80 %	20 %	40 %	3,0
		No errors in claims	0 %	40 %	0 %	13 %	1,3
XBRL-JSON +TAX	Plus	Recognizes initiatives	100 %	100 %	20 %	73 %	5,3
		Correctly categorized	80 %	80 %	0 %	53 %	5,3
		No errors in claims	60 %	60 %	0 %	40 %	3,0
XBRL-JSON +TAX	Pro	Recognizes initiatives	100 %	0 %	100 %	67 %	8,3
		Correctly categorized	80 %	0 %	80 %	53 %	5,3
		No errors in claims	60 %	0 %	60 %	40 %	3,0

Scale
100 %
80 %
60 %
40 %
20 %
0 %

Table 8: The table shows the number of correct responses in percentages for each trial (Try 1, Try 2, Try 3), along with the average and variance of correct answers for each combination of file format (PDF, XBRL-XML, XBRL-JSON) and ChatGPT model version (ChatGPT Plus 40, ChatGPT Pro 40). For XBRL-XML and XBRL-JSON, the models were tested both with only the raw files and with an accompanying taxonomy file (TAX) that provides the structural metadata and explains the tags used in the sustainability reports.

As can be seen in table 8, in text analysis, when analysing one report, PDF performs best. PDF can correctly recognize and categorize the initiatives without making any mistakes having a 100% success rate in all categories. In structured machine-readable reports (XBRL-JSON and XBRL-XML) the text analysis does not perform as well.

There are more differences in performance not between filetypes but rather within the three tries. Sometimes GPT gives correct answers and then when asked to repeat the same task with same filetype again it performs poorly.

The only file type that gives consistently similarly great or poor answers are PDF and XBRL-XML with taxonomy.

The notable difference with PDF and XBRL-XML with taxonomy is that in all other file types (XBRL-XML, XBRL-JSON and XBRL-JSON + tax), GPT hallucinates its own initiatives. This means that it invents initiatives that are not mentioned in the text it is supposed to base its answer on.

It was also observed that GPT sometimes take initiatives from outside the asked part and takes answers from wrong parts of the report. When it was supposed to analyse only text under title "8. Does the Company have any project related to reducing Green House Gas emission? If yes, then provide details." It takes answers from under title "If the entity provided below taken any specific initiatives or used innovative technology or solutions to improve resource efficiency or reduce impact due to emissions / effluent discharge / waste generated, please provide details of the same as well as outcome of such initiatives, as per the following format".

However, accuracy is affected by the fact that in this study, the features of a successful answer have been very precisely defined. Even if part of the AI's generated output is correct, if it also contains incorrect facts that have been found from a different part of the report than what was asked, the item is classified as incorrect in the evaluation. Thus, in some cases ChatGPT finds more correct answers than the results show, but they also include incorrect ones and therefore when evaluating correctness and accuracy, they are recorded as incorrect.

4.3 Summary of findings

Several observations are made regarding the different file formats, question types, as well as ChatGPT models. This this chapter presents the main results and provides explanation. Additionally, the most significant shortcomings and an overall analysis are highlighted.

		Average					
		Q1	Q2	Q3	Q4	Average of averages	Averages for file format
PDF	**Plus**	75%	40%	100%	0%	**54%**	**51%**
	Pro	47%	42%	100%	0%	**47%**	
XBRL-XML	**Plus**	33%	14%	0%	65%	**28%**	**32%**
	Pro	33%	14%	0%	100%	**37%**	
XBRL-XML + TAX	**Plus**	100%	100%	33%	100%	**83%**	**68%**
	Pro	100%	40%	0%	67%	**52%**	
XBRL-JSON	**Plus**	14%	0%	0%	40%	**14%**	**20%**
	Pro	5%	33%	0%	67%	**26%**	
XBRL-JSON +TAX	**Plus**	11%	33%	0%	100%	**37%**	**19%**
	Pro	9%	2%	0%	0%	**3%**	

Table 9: Success rates based on averages of filetypes, questions and GPT versions.

As can be seen in table 9. XBRL-XML with taxonomy performs by far best out of all file formats having an average accuracy of 68% when considering questions 1-4. XBRL-XML performs much better when taxonomy is given than when it is not included. Without taxonomy the average success rate for Q1-4 is 28% (Plus) and 37% (Pro), and with taxonomy it is 83% (Plus) and 52% (Pro) (table 9). Even though the files themselves are identical, adding the taxonomy file to the project opening the tags as names, improve the performance significantly from 28% to 83% in Plus (table 9). In XBRL-JSON however taxonomy does not increase success rate rather giving 1% units less accuracy (table 9). Overall, XBRL-JSON performs worst out of the other two file types (PDF and XBRL-XML).

PDF performs best in data visualization and text categorization with 100% success rate (table 9). Other file types have only 0-33% success rate in data visualization question. However, this is due to wrong data, not problems in the actual visualization. More specifically, ChatGPT can visualize the data it finds correctly, in this task however it finds the wrong data, and therefore the answer is not marked as correct. PDF performs worse out of all file types in Q4 Ranking the data, with 0% success rate (table 9).

GPT Version	Q1	Q2	Q3	Q4	Q5
Plus	47%	38%	27%	61%	71%
Pro	39%	26%	20%	47%	51%
Average	43%	32%	23%	54%	61%

Table 10: Performance per question for questions one to five.

As shown in table 10, the performance of Q4 is the best out of questions 1-4 when comparing the average % of correct answers from all trials and file types (excluded Q5, as analysed separately). Results show that question 4 which compared and then ranked different values in order between given data files, was the best executed task from AI. It gives an average of 54% of correct answers. Additionally, data extraction (Q1) gives the second-best performance 43% of average from all file formats. This shows that the LLM performs well with tasks related to data extraction, more specifically finding the asked values from the datasets and using the data to deduct analysis.

Overall, question 5 outperforms all other question types in terms of percentage of accurate outcomes (table 10), but the results are not straightly comparable as the nature of the question and analysis is significantly different. In question Q5, the focus was on understanding narrative text and categorizing it accordingly as the other four questions are related to analysing numerical values. In question 5, an important observation is that GPT sometimes takes initiatives from the wrong parts of the report which affect the correctness of the answer generated.

Continuing the analysis by focusing more closely on how performance varies across file formats. In analysing data for Questions 1-4, the file formats (PDF, XBRL-XML, XBRL-XML + taxonomy file, XBRL-JSON, XBRL-JSON + taxonomy file) performed as shown in table 11. The overall performance ranking from best to worst is: XBRL-XML + taxonomy file, PDF, XBRL-XML, XBRL-JSON, XBRL-JSON + taxonomy file.
For question five, however, the ranking differs. Ranking from best to worst is: PDF, XBRL-JSON, XBRL-JSON + taxonomy file, XBRL-XML and XBRL-XML + taxonomy file. Notably, each file format occupies a different position when comparing the rankings from Q1-Q4 to that of Q5. Interestingly, XBRL-XML with taxonomy files goes from the best-performing format in Q1-Q4 to the worst in Q5.

	Q1-Q4 average	Q5
PDF	51%	100%
XBRL-XML	32%	52%
XBRL-XML + TAX	68%	41%
XBRL-JSON	20%	57%
XBRL-JSON +TAX	19%	54%

Table 11: Performance with the average of Plus and Pro in questions one to four and then question five.

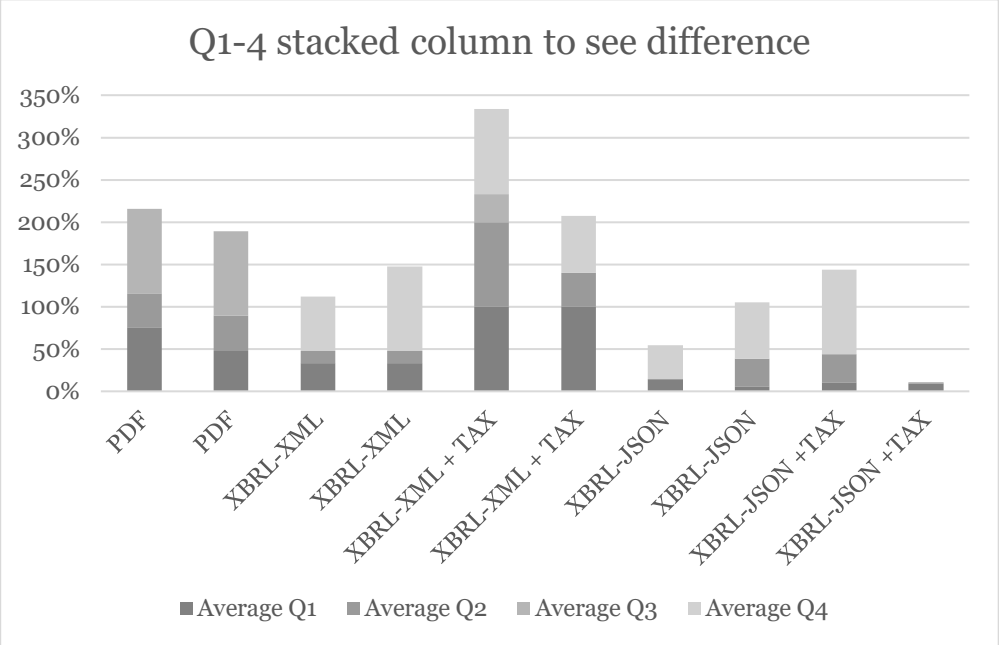


Figure 18: Average success rate in stacked column chart on Q1: data extraction, Q2: data creation, Q3: Data visualization, Q4: Data Ranking.

Figure 18 shows that overall, when questions 1-4 are all put together XBRL-XML with taxonomy with ChatGPT Plus version outperforms all the other success rates. It performs 54% better than the second-best format, PDF, with GPT Plus. However, it is important to note that when the taxonomy is provided with XBRL-XML files the average success rate for Q1-Q4 with taxonomy is multiple times higher than without it.

Moreover, the main objective in this study is to analyse the performance of the ChatGPT in general when analysing the Business Responsibility and Sustainability Reports with AI, not so much about the differences between the GPT versions (Plus & Pro). However, a few interesting observations were made. These two main findings observed when analysing differences in Plus and Pro are listed below.

1. Chat GPT **plus performs better on average in each question** when all file format averages average is looked.

2. When looking at file format performance on average in all questions
    a. **Plus performed better in:** PDF, XBRL-XML with taxonomy and XBRL-JSON with taxonomy
    b. **Pro performed better in**: XBRL-XML without taxonomy, and XBRL-JSON without taxonomy

What is interesting is that Chat GPT Plus outperforms Pro on average in each question even though the subscription of Pro version is 10 times more than Plus version monthly. ChatGPT Plus version has 20$ + tax monthly subscription fee as the cost of Chat GPT Pro is 200$ + tax monthly. Despite the notable price difference, Plus performs better on average in each question (Q1-Q5) when all file format averages are looked as shown in table 10. The percents are calculated as an average of all trials from different project with different file formats for both, Plus and Pro version, separately.

Interesting observation is that the average accuracy percent of all questions percentage differences between Plus and Pro is notable. In each question type regardless of the file format used (PDF, XBRL-XML, XBRL-JSON) the average accuracy of correct answers is every time greater with Plus (table 10). Additionally, both Plus and Pro versions have the same result when comparing the order in which question has performed the best. In both cases, the highest average accuracy percent is with Q5, followed by Q4, Q1, Q2, and Q3 respectively. However, it is only when taking average of all results. When examining separately each format, Pro outperforms Plus version in some tasks.

PDF	PDF	XBRL-XML	XBRL-XML	XBRL-XML + TAX	XBRL-XML + TAX	XBRL-JSON	XBRL-JSON	XBRL-JSON +TAX	XBRL-JSON +TAX
Plus	**Pro**	**Plus**	**Pro**	**Plus**	**Pro**	**Plus**	**Pro**	**Plus**	**Pro**
63%	58%	35%	38%	78%	47%	27%	28%	40%	13%

Table 12: question 1-5 all average of the average success rate per versions.

As said, even though Plus perform every time better when looking at averages in question types it is not always the case when looking specifically at file formats. In both XBRL formats, XML and JSON, Pro performs better when the taxonomy is given (table 12). However, if company has only XBRL-JSON formatted reports and no taxonomy provided, an unlikely situation to be in, Pro version would have better performance based on the results. It is also important to note that in cases when Plus performs better it is 5% units, 31% units and 27% units better whereas in Pro the differences in accuracy in percentage units are 1% units and 3% units as can be seen from table 12. Therefore, ChatGPT Pro does not seem to offer significant additional value for the tasks examined in this study despite it has more expensive price.

		Variance			
		Q1	Q2	Q3	Q4
PDF	**Plus**	1.33	1.33	0.00	0.00
PDF	**Pro**	67.00	0.00	0.00	0.00
XBRL-XML	**Plus**	120.33	5.33	0.00	37.33
XBRL-XML	**Pro**	0.33	21.33	0.00	0.00
XBRL-XML + TAX	**Plus**	0.00	0.00	8.33	0.00
XBRL-XML + TAX	**Pro**	0.00	100.33	0.00	120.33
XBRL-JSON	**Plus**	0.33	0.00	0.00	100.33
XBRL-JSON	**Pro**	0.00	120.33	0.00	120.33
XBRL-JSON +TAX	**Plus**	0.00	120.33	0.00	0.00
XBRL-JSON +TAX	**Pro**	0.33	0.33	0.00	0.00

Table 13: Q1: data extraction, Q2: data creation, Q3: Data visualization, Q4: Data Ranking.

Moreover, as table 13 shows the variance across questions, file formats and GPT versions does not have any clear pattern. The largest variance is 120,33 while, in many cases, it is as low as 0. Variance, however, does not indicate the quality of the answers alone, since variance of 0 can mean that all answers on all tries are incorrect.

Even when querying with a brief context, the LLM can perform very differently in different tries. It can get a fully correct answer on try 1 but when the exact same question is repeated it can get it wrong. This shows that depending on the realized scenario the outcome may vary significantly. More closely, the best-case scenario 100% and worst-case scenario 0% of correct answers may both be reach in only two consecutive attempts as shown in many cases in the study. This shows a lack of consistency in LLM's outputs.

ChatGPT capacity with number of reports it can handle falls, being usually able to only return 2-15 companies' data reliably and consistently. This occurred even though many of the questions explicitly asked to return data for all 19 companies. However, AI can be helpful in analysing a single value or a certain point from a large number of reports which may otherwise require more advanced coding skills.

Another significant source for incorrect answers from the LLM is due to values found from wrong parts of the files. This was most common in files that conform to the structured XBRL type where information is saved as tags defined by taxonomy. With PDF files, fewer of these kinds of errors occur. However, PDF hallucinates more, meaning it comes up with answers that do not

really exist in the report, or the answer is not based on the asked part of the report.

Additionally, here are some individual observations that were also noticed when conducting the study. At the time of this study was done, ChatGPT Plus can not read too large tables from PDF file. The table examined was about 1550 words and 6 pages long, and ChatGPT do not have enough answering capacity to recreate the chart and all its data. Therefore, it comes up with its own imaginary data (hallucinating).

Moreover, GPT-4o understands numbers well, but special characters can sometimes cause issues. When dealing with tables, the accuracy of extraction depends on how precisely they are requested and how clearly, they are structured. GPT-4o cannot always analyze an entire table correctly within a broad analysis task, even if the file format is appropriate. If an analysis prompt is too complex, GPT may extract and analyze parts of the table but fail to reconstruct it fully. However, when given a highly specific command, such as "Recreate the table from [company, specific part from the report]", it correctly extracts the table word for word. This demonstrates the importance of precise and direct prompts when requesting specific sections from a report.

When a request is too broad or unclear, GPT may hallucinate missing data, making up content instead of extracting it accurately. This is especially problematic when analyzing large reports where the model's capacity is exceeded. However, when asked to regenerate a specific table, ChatGPT can complete the task correctly. Therefore, it is crucial to be precise in formulating prompts. If the request is vague or allows for broader analysis, GPT may hallucinate or omit parts of the data. This can be difficult to detect without verifying the output manually.

Generally, across all questions, great volatility is observed. This shows the AI's ability to already function in analyses, but it is not yet systematic. ChatGPT is able to categorize and "understand" also narrative text, but the success of answer generations has very high volatility. Due to this, the reliability of the answers lowers. Additionally, because of the high quality of ChatGPT's text generation ability, the answers may seem very accurate, which makes it difficult to observe errors and hallucinations without closer inspection.

Additionally, the different nature of tasks implemented in this study affects the performance between file formats. For example, Q1, Q2 and Q4 are related to all the reports in the project files as Q3 and Q5 are only related to one report. Therefore, it is notable that when an analysis is done from one report from a specific section in it, the unstructured data format (PDF), clearly works best, providing 100% accuracy of correct answers every time in this study. These observations are important because analyses are need from

both, across a large number of reports for comparison purposes as well as more in-depth analyses of individual reports.

In conclusion, when comparing all the tasks and file formats, XBRL-XML format with taxonomy delivered the best performance. However, the high variance in many comparisons between different trials indicates inconsistency in ChatGPT's responses. In addition, the lack of transparency and prone to mistakes reduces its reliability but shows great potential for improving efficiency and expanding analytical capabilities. Efficiency can therefore be achieved within the current capabilities of ChatGPT, although the tendency to errors must be carefully taken into account.

5 Discussion

The aim of this discussion is to analyse results and reflect them on the previous literature. The focus in chapter 5.1 is on theoretical implications as how performance varied across different tasks and file types is examined, and how these findings align with previous research. Chapter 5.2 explores the managerial implications of the study and gives particular attention to ChatGPT's ability to analyse sustainability reports. It gives potential applications and provides practical examples. Additionally, recommendations for the CSRD, future sustainability reporting frameworks, companies and investors are provided. Lastly, chapter 5.3 outlines the limitations of the study and future research recommendations.

5.1 Theoretical implications

Data is everywhere and is being leveraged in a wide variety of ways for many different purposes. Information of business activities is important for organizations as arguments based on data have a valuable effect on organizational decision-making (Bhimani, 2015). Having these large data sets can certainly be beneficial for companies (Müller et al., 2018), however as Grover et al. (2018) notes it is important to know and utilize state-of-the-art tools to get the benefits from the data. This is where this thesis brings some new research into AI tools, specifically ChatGPT 4o, and its ability to use sustainability data.

The results of the study are consistent with earlier research available on the topic. To compare, in related study from Bang et al. (2023), they tested performance of ChatGPT in 10 different reasoning categories, in logical reasoning, commonsense reasoning and non-textual reasoning. The researchers found out that ChatGPT had 63.41% accuracy on average. The accuracy percentages in our results are a bit lower in most tested tasks due to the slightly different nature of our study, showing still the same order of magnitude in terms of accuracy percentage. Though there is a growing amount of research on NLP, there is a lack of the type of research we conducted in this study. As a result, there is little previous comparison to the more detailed and sophisticated analysis performed using artificial intelligence.

Additionally, the more advanced the given task is, the more there is room for errors and misleading analysis. Compared to classifying sentiments, specific context is more prone to misclassification (Kang et al., 2020). This impact of more sophisticated and different types of tasks is visible when comparing results from the correctness percent accuracy in our study to accuracy of

63.41% from Bang et al. (2023). However, regardless of the given task, the basic functioning of AI remains the same.

AI can be seen as a tool to utilize data more efficiently. However, challenges involved in creating new insights from it as applying data analytics does not automatically lead to better decisions (Müller et al., 2016). LLM generated answers can thus be seen as an aid to this challenge to find relevant data. When analyzing business and responsibility reports, AI can be useful to find values from several reports when asking with narrative "human readable" language without need of knowing specific tags or coding. Additionally, machine-readable file formats like XBRL-XML and XBRL-JSON may be hard for human users to understand in their current form.

However, the answers generated by LLM cannot be trusted blindly which limits the gained benefit for the user. Hallucinations and lack of transparency may distort the result (Salah et al., 2023). In addition, to correct answers, ChatGPT can hallucinate values that seem like they could be correct, as observed from the results. This can be problematic in cases where data is not known from inside out and one falsely trusts the hallucinations.

Another notable observation from the study is that sometimes GPT uses incorrect data but makes correct conclusions. For example, finding wrong data but then visualizing the wrong data successfully. Sometimes it is able to find correct data but does not make correct conclusions as it finds correct values but still ranks them in wrong order. Additionally, GPT sometimes may take initiatives from outside the asked part of the report. These actions undermine the credibility of AI analysis as without comparing the response to the source data, reliability of the output may not be certain.

This may be due to the fact that especially the narrative text appearing in reports may face challenges in being accurately read by AI models, compared to structured data where the indicator is already in a machine-readable format as explained by Tayefi et al. (2021). They also mention that this challenge is related to splitting words and marks correctly into separate entities. This is observed from the fact that the variation in accuracy between different trials is generally very high. Sometimes GPT succeeds in finding all the correct projects and classifying them correctly as "quantifiable" and "unquantifiable," and sometimes the accuracy is 0% as noticed in categorizing task in our study.

The data format also has a major effect on the outcome. As a conclusion in this study, the machine-readable XBRL-XML format, with taxonomy provided, performs better than XBRL-JSON or PDF format when analyzing multiple Business and Responsibility reports with ChatGPT. This is because the

taxonomy in XBRL provides semantic meaning to the reported facts, enabling more accurate interpretation (Ramanan & Warren, 2021). Also, a paper by Beelitz (2017) announced that XBRL-XML was the winner of comparison in their study, where XBRL-XML and XBRL-JSON were compared for the purpose of intra-linkage of financial information. Therefore, the author encourages to improve XBRL-XML form for integrating information which also supports the finding from this study.

One factor affecting the significant difference in performance between two different XBRL data types (XML vs. JSON) can be identified relating to the different structure between them. This is also visible in figure 3 presented in section 2.3 about XML–JSON differences, where XBRL's official site, Ramanan and Warren (2021), explains how, for example, dates are presented differently. The differences in instants and durations in XBRL-XML and XBRL-JSON may have affected how GPT analyzes the years. Therefore, in the case of ranking values with XBRL-JSON, some values found by GPT were from previous years when values from the current year were requested.

Additionally, since the PDF project file is almost three times larger, it supports the observation of why PDF works much better when analyzing one report at a time rather than all reports in the project file simultaneously. Related to file sizes, another interesting observation is that XBRL-JSON is the smallest in size, and therefore it does not explain its poor performance compared to XBRL-XML.

The number of reports to be analyzed can therefore be seen as one contributing factor in determining which file format should be chosen for the analysis. Machine-readable XBRL-XML or XBRL-JSON formats, which were used in this study, struggle with retrieving information from the requested section. For example, they may consider other parts of the data file or use numbers from the wrong year. However, when it comes to questions which emphasize finding a specific value present in the report from several reports and then using found information to complete the requested task, structured, machine-readable formats work better. In such cases, the large amount of input likely affects capacity. Therefore, the structured XBRL-XML format generally performs the best when all different tasks are considered.

One last observation related to the different file formats that may affect ChatGPT's ability to generate answers, is that Chat-GPT 4 is primarily trained with webpages, books and other human-readable documents from e.g. third parties (OpenAI, 2023). As PDF is a common reporting format (Chao & Fan, 2004), it could indicate that also the training data, that uses web pages, might have more pre-training with the model on these PDF files than XBRL ones.

The aim in this study is analysing AI's ability to generate outputs from different format, but additionally, some interesting observations were made between performance of Plus and Pro versions. Those differences in performance are likely due to other factors than different models as both utilize the same underlying model. Therefore, benefits of ChatGPT Pro are more related to accessibility and limitations as it uses same technology as Plus. According to OpenAI (2025) ChatGPT Pro includes all features of Plus, with additional benefits. Also, Pro has unlimited access to all reasoning models and a more powerful version of o1 (OpenAI, 2025). For example, Pro has faster responses, priority access and deep research (OpenAI, 2025). This goes against the results in this study and supports the claim that differences in Plus and Pro are likely due to other limitations.

Prioritized traffic, no peak hour limits and minimized disruptions even during highest demand times, are factors especially relevant in our study context. Other characteristics better in Pro version are early access to new features, models and to be among the first to try new capabilities as they launch as well as extended access to deep research (OpenAI, 2025). All testing was conducted using the GPT-4o model and focused strictly on analysing provided material. Therefore, such benefits like early access to new models, their features and extended research capabilities did not play a significant role in this study.

Additionally, LLMs have a pre-defined context window size for maximum amount of input tokens. When conducting long conversations or summarizing long documents, its limit is often outpaced (Chen et al., 2023). However, differences in size of context windows does not matter in this study as all the prompts are kept very short.

However, as Open AI algorithm is not an open source, it is hard to analyse the reasons behind the different performance of GPT Pro and GPT Plus. The differences in accuracy of outputs in this study may have more likely been a cause of other limitations than the different features of GPT versions. More of the limitations are described in chapter 5.3.

5.2 Managerial implications

The study provides some important insights for current and future sustainability frameworks, organizations, investors and sustainability data information systems. It gives implications on how the EU's Corporate Sustainability Reporting Directive (CSRD) can develop its framework. By testing how different file formats and question types perform in a LLM, we can get insights on future possibilities and how to best prepare for them. Additionally, as Dwivedi et al. (2021) mention, the current research focuses more on impacts than performance of AI, whereas our research brings clear results on

AI's abilities and how it can be utilized, as well as its weaknesses. All crucial for organizations and policy makers.

The number of reports ChatGPT is able to analyze at the same time is still limited. AI helps find and analyze values from business reports, but the opportunity for a more comprehensive analysis with a larger sample size still needs further development. However, the ability to rank and find values from several reports is already possible and therefore seen as enhancing valuable insight into decision making.

Based on the findings XBRL-XML works best in overall analysis for sustainability reporting especially in analysing numerical data. It does however need the taxonomy file to accompany the XBRL-XML file. It thrived when multiple companies data is asked all at once. The results show that including the taxonomy was very important as in XBRL-XML the success rate jumped from 28–37% without taxonomy to 52–83% with it. Therefore, taxonomy development and enforcement should be prioritized in CSRD as well in order to get better results.

In contrary PDF format excels when a single reports information is asked. For that reason, it is recommended for CSRD to mandate the use of XBRL-XML and PDF both meaning that PDF should not be fully dismissed. It outperformed XBRL formats in certain tasks likely due to GPT's strong pretraining on human-readable content (OpenAI, 2023). Human readable formats in reports are also needed for publishing for people to read and analyse. This would help comparability, large analysis from separate reports and comparison throughout yearly development. PDF is also human-readable, which makes it important for accessibility and direct human analysis.

Overall, as mentioned, structured and unstructured data perform differently in different tasks. However, as Aaltonen and Penttinen (2021) point out, the boundary between them is often unclear and ambivalent. It is also important to note that many companies collect and store the majority of their data in unstructured formats (Chen et al., 2012), which influences what kind of analytical approaches are feasible and how data processing methods should be chosen.

Another notable finding is the significant performance gap between XBRL-XML and XBRL-JSON. XBRL-JSON performance was comparably poor and for that reason specifically XBRL-XML is recommended. According to this study, the use of XBRL-JSON is not recommended in sustainability report analysis as its success rate is low. This aligns with earlier claims about the efficiency and better performance of XBRL-XML over alternatives such as XBRL-JSON (Beelitz, 2017).

Moreover, companies, investors, governments and unions could benefit from having structured data, as the results of this study suggest it can be analysed more effectively with AI. This would require resources and increased focus on building structured data and reporting practices within organizations. As the trend is towards automation it would likely pay off in the future and can help get more accurate analysis when AI is utilized.

Furthermore, as shown in the study by Zou et al. (2025), unstructured ESG data was converted into a structured format using AI to enhance the accuracy and efficiency of disclosure assessments when reviewing a large number of reports at once. This further supports the need for analyse responsibility data in a structured format that can be utilized across large data set. This is also consistent with the results of our study and indicates the need for a machine-readable data format, in the context of this work in particular XBRL-XML (with a taxonomy file) for efficient analysis. Therefore, it is also seen as a necessary recommendation for the CSRD.

Moreover, as data driven decision making is on the rise, also sustainability data is moving towards more measurable impacts and reporting. Using structured reporting can give more reliable accurate results as shown in this study. When analysing text, PDF seems to have better accuracy and performance in the other hand. This indicates that maintaining some risk analyses and textual data in unstructured forms could be beneficial.

This shows the need for both concentrating the AI development as well as skills of human analysts. Developing AI systems for better functions, using the most suitable file formats and optimizing answers are important factors affecting the results. It is important to train expert human analysts who can critically interpret and validate AI-generated outputs, as AI is also prone to making errors.

Errors are important aspect to be aware of when creating insights based on AI generated data analysis, as there is high volatility in the answers generated. The LLM's output cannot be blindly trusted even if the correct result is previously given. As variance between trials is high, it indicates a large spread between the possible outcomes. ChatGPT is able to perform with 100% correctness in the best-case scenario, but in the next try, it may have 0% accuracy. This can be considered one of the most important observations in our study, as how well ChatGPT succeeds in the given task can be, so to speak, "a matter of luck."

To continue, it is important to observe also minor errors in order to analyse AI generated outputs correctly. For example, if the numbers appear correct

but the decimal point is in the wrong place, it can change the value by millions, leading to a completely incorrect insight. Additionally, for example when asked to give total energy consumed from non-renewable energy sources and the AI finds total fuel consumed from non-renewable energy sources, the error may be hard to notice. ChatGPT and LLMs in general can thus make many errors that are not direct hallucinations. The data is searched from the correct source and is not made up, but it is either written incorrectly or retrieved from the wrong section.

Additionally, another minor note which does not affect the creation of insights from the data but is worth noting is the presence of rounding errors. This is evident sometimes in PDFs, when the final values are off by 0.01%–0.03%, probably due to rounding in intermediate steps. Similarly, a small error that does not affect the outcome of the analysis but can unknowingly mislead is that units are transformed incorrectly.

All these insights offer practical implications regarding the recommended file formats to use and the caution organizations and investors should take when using AI as a tool for sustainability analysis. In addition, practical examples were also generated based on findings. It was discovered that a new value can be easily calculated based on the data provided in the report. This value or change in value can be further linked to the company's operations (e.g. the increase in renewable energy use in relation to total energy use) and comparisons can be studied across several companies by ranking them in order of importance. In addition, the narrative disclosures can be analysed more accurately, for example by classifying concrete quantifiable operations from vague descriptions which can also be useful in identifying greenwashing.

As a conclusion, from a practical perspective, both shortcoming and value creation potential are recognized when utilizing generative AI to analyse organizational reports. Work that would take more time if done manually can be demonstrably enhanced. This reveals also a strong potential for utilization by both organizations and other stakeholders during the planning, execution as well as analysis of reporting.

5.3 Limitations and future research

Due to the scope of this study, it has been necessity to make choices and simplifications that constitute some limitations. For example, the average accuracy percentage could have been more accurate if there had been more than three trials in each task. The constraint to three times was set, as some parts of checking had to be done manually being time consuming. Additionally, tendency to inconsistency and spread between answers is seen already in three tries.

Additionally, some answers were hard to analyse whether they were correct or incorrect due to the complexity of the tasks. For example, there were some small rounding errors and typos seen in multiple answers. This may have led to over or underestimation of accuracy of the answers when the manual checking was done. Recording the results in a uniform format may also have suffered from simplifications which can affect their analysis.

There are also limitations to be aware of with answers generated by ChatGPT. ChatGPT reports the shortcoming: "Responses may be worse due to the number of files used in this project." This may partly explain the decline in response quality and completeness observed in some cases. Additionally, the size of the project files varied in this study which could have affected ChatGPT's ability to generate analyses. Especially, PDF files were multiple times larger in size than XBRL-XML or XBRL-JSON. The file sizes were as follows: XBRL-XML: 15,6 MB, PDF: 44,8 MB, XBRL-JSON: 14,9 MB and the Taxonomy files: 0.5 MB.

Another minor attention related to the file formats in this research is the inconsistency related to names used in the ChatGPT projects. Context and questions were given for XBRL-XML with filename mentioned as "XBRL" and for XBRL-JSON files as "JSON". This might have affected the LLM's answers even though both contexts are correct in essence.

It is also important to note that memory in ChatGPT was turned off, as learning between answers was not wanted. The research simulates only a no memory one shot scenario where ChatGPT did not have a learning curve and was not corrected. Answers might be worse for this reason, but the target was to make study as objective as possible and look at ChatGPT capabilities without teaching the model. Additionally, at the beginning of the research process, ChatGPT Plus version was tested more than PRO with the same files used in this study. However, the study had been executed in empty folders in both versions, and the memory was turned off as described earlier. Therefore, the effect of using Plus version more cannot be excluded, but in this context, the impact is not believed to be significant.

Lastly, there are several directions which would be interesting to dive further into for future research in using LLM's to analyze sustainability reports. Generative AI is developing rapidly, which creates expectations for more use cases and better accuracy already in the near future. One main direction would be researching how LLM's can be taught and bettered in this specific task of analyzing data and especially sustainability reports. As this study only looks at results from a non-teaching perspective, this could bring new insights to complement this study. Moreover, it is good to note that all reports are from India under the BRSR reporting framework and from the energy sector when looking at results. For future research, other sectors and

sustainability frameworks could be analysed as well as other tasks like policy alignment, risk and opportunity analysis and greenwashing characteristics in reports.

Furthermore, looking at other LLM's like Copilot, DeepSeek and other models would be interesting. Comparing different AIs could show different outcomes between different kinds of tasks. Additionally examining other ChatGPT versions such as ChatGPT 3.5 or 10 would be interesting for comparing the results observed in this research. This could help find the most suitable versions and LLM's for different tasks with specific features.

6 Conclusions

This study examines how ChatGPT 4o can analyse sustainability reports in various file formats (PDF, XBRL-XML and XBRL-JSON), and how it performs in different tasks (data extraction, creating values, visualization, ranking and data categorization). The study identifies the most efficient file formats and task types for data analysis and the biggest shortages that affect analysis and provides insights and recommendations to CSRD policy makers, companies, and investors.

6.1 Summary of Key Findings

The best success rate for analysing multiple reports and data was when using XBRL-XML format with the taxonomy file provided. When there is need to compile, create and rank data, using XBRL-XML format is the pre-eminent choice. The best success rate for analysing a single report was achieved when using PDF. When wanting to analyse text or visualizing data, PDF format gives better accuracy. Use of XBRL-JSON is not recommended in this context.

Overall, the LLM worked best in ranking data, analysing text and extracting values from reports. Calculating new values from existing ones and creating visualizations showed lower accuracy. Surprisingly Chat GPT Plus version (20$ + taxes) outperformed Chat GPT Pro (200$ + taxes) in all questions and most file formats. The explanation to this was left unclear.

6.2 Implications

For CSRD policy makers, companies and other entities it would be recommended to enhance and promote the use of machine-readable reporting formats, specifically XBRL-XML and its relevant taxonomies. However, human readable formats like the PDF remain important in text analysis and analysing single reports. Therefore, providing reports that are easily human-readable are still needed.

In general, AI can make report analysis more efficient. There is potential and an increase in efficiency, which is needed, as reporting itself is useless if the information it provides is not analysed. As sustainability themes are important, reported actions can be more accurately analysed, compared, and developed when the information is better available with the help of artificial intelligence. In the future, the potential of artificial intelligence and its capacity will develop into even more diverse and reliable analyses and from an even larger sample, at which point the benefit will increase even further.

However, rather than being ambivalent, the results demonstrate that the success of AI applications relies on both technical features and the context of use. It is recommended to take careful consideration and critical thinking with answers generated by artificial intelligence. The study found, among other things, that ChatGPT sometimes gives very low accuracy, makes mistakes, hallucinates and takes numbers and data from the wrong places or year in reports.

Additionally, GPT can only process a limited amount of information (reports) at a time, which is why it cannot yet be utilized efficiently to a large number of reports. The study showed that the same prompt can give a success rate of 0% or 100% in same conditions, showing the huge volatility and inconsistency in GPT results. Human readable reports such as PDF are important for the manual checking that needs to be done. Despite these limitations, the use of LLM to analyse reports is already seen as beneficial.

7. References

Aaltonen, A., & Penttinen, E. (2021). What makes data possible? A sociotechnical view on structured data innovations. *In Annual Hawaii International Conference on System Sciences* (pp. 5922-5931).

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the association for information systems*, *17*(2), 3.

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, *16*(1), 3-9.

Agama, E. J., & Zubairu, U. M. (2022). Sustainability reporting: A systematic review. *Economics, Management and Sustainability*, 7(2), 32-46. https://doi.org/10.14254/jems.2022.7-2.3:contentReference[oaicite:0]{index=0}.

Agarwal R., Dhar V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research 25*(3):443-448.
https://doi.org/10.1287/isre.2014.0546

Ahmad, V., Goyal, L., Arora, M., Kumar, R., Chythanya, K. R., & Chaudhary, S. (2023, September). The Impact of AI on Sustainability Reporting in Accounting. In 2023 6th *International Conference on Contemporary Computing and Informatics (IC3I) (Vol. 6,* pp. 643-648*)*.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q., Xu, Y. & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv*:2302.04023.

Baskerville, R. L., Myers, M. D., & Yoo, Y. (2020). *Digital first: the ontological reversal and new challenges for information systems research. MISQ 44* (2): 509–523.

Baumüller, J., & Grbenic, S. O. (2021). Moving from non-financial to sustainability reporting: Analyzing the EU Commission's proposal for a Corporate Sustainability Reporting Directive (CSRD*). Facta Universitatis, Series: Economics and Organization,* (1), 369-381.

Beelitz, C. (2017). The dilemma of XBRL-XML versus XBRL-JSON regarding linkage of financial information. *In CEUR Workshop Proceedings (Vol. 1890,* pp. 1-11).

Bhimani, A. (2015). Exploring big data's strategic consequences. *Journal of Information Technology, 30*(1), 66-69.

Bouquet, P., Molinari, A., & Sandri, S. (2024, November). Challenges in Working with Unstructured Data in the LLM Era: NER Processes Using Graph Neural Networks. In *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-6). IEEE.

Brondoni, M. S., Plata, E., (2022). Ouverture de 'Global Competition and Sustainability Management'. *Symphonya. Emerging Issues in Management (symphonya.unicusano.it),* (2), 1-5.

Brynjolfsson E., Li D., Raymond L., (2025). Generative AI at Work. *The Quarterly Journal of Economics (Vol. 140*(2), May 2025, pp. 889–942). https://doi.org/10.1093/qje/qjae044

Chao, H. and Fan, J., (2004). Layout and content extraction for PDF documents. In Document Analysis Systems VI: *6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004. Proceedings 6* (pp. 213-224). Springer Berlin Heidelberg.

Chen, S., Wong, S., Chen, L. and Tian, Y. (2023). Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:*2306.15595. Available at: https://arxiv.org/abs/2306.15595

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165-1188.

Cusumano, M.A., Gawer, A. and Yoffie, D.B., (2020). The Future of Platforms. *MIT Sloan Management Review.* [online] Available at: https://www.oreilly.com/library/view/the-future-of/53863MIT61304/chapter001.html [Accessed 13 May 2025].

Design My Report. (2021). 2022: BRR to BRSR – The shift towards sustainability. Available at: https://designmyreport.com/blog/2022-BRR-to-BRSR.php (Accessed: 3 May 2025).

De Villiers, C., Dimes, R. and 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), pp.96-118*.

di Tullio, P., La Torre, M., Rea, M.A., Guthrie, J., & Dumay, J. (2023). Beyond the planetary boundaries: exploring the role of accounting in addressing ecological overshoot. *Sustainability Accounting, Management and Policy Journal, 14*(2), pp. 96-118. DOI: 10.1108/SAMPJ-02-2023-0097.

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P.V., Janssen, M., Jones, P., Kar, A.K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L.C., Misra, S., Mogaji, E., Sharma, S.K., Singh, J.B., Raghavan, V., Raman, R., Rana, N.P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P. & Williams, M.D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International journal of information management*, *57*, 101994.

European Commission. (2022). Sustainable finance: EU adopts new rules on corporate sustainability reporting. *European Commission*. https://ec.europa.eu/commission/presscorner/detail/en/mex_22_3966 [Accessed 10 Feb. 2025].

European Commission. (2023). Questions and answers on the European sustainability reporting standards (ESRS). Available at: https://ec.europa.eu/commission/presscorner/detail/en/qanda_23_4043 [Accessed: 10 Feb. 2025].

European Commission. (2025). Q&A on simplification omnibus I and II. Available at: https://ec.europa.eu/commission/presscorner/detail/en/qanda_25_615 (Accessed: 13 May 2025).

European Commission. (n.d.). Corporate sustainability reporting. Available at: https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en (Accessed: 13 May 2025).

EY. (2023). BRSR reporting and the evolving ESG landscape in India. Available at: https://www.ey.com/en_in/insights/climate-change-sustainability-services/brsr-reporting-and-the-evolving-esg-landscape-in-india (Accessed: 3 May 2025).

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*., 66(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Gharehchopogh, F. S., Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. *2011 5th International Conference on Application of Information and Communication Technologies.* [Online]. IEEE. pp. 497–4.

Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021). The role of artificial intelligence and data network effects for creating user value. *Academy of management review, 46*(3), 534-551.

Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. *Journal of management information systems*, 35(2), 388-423.

Hasan, M.R., Maliha, M. and Arifuzzaman, M., (2019). Sentiment analysis with NLP on Twitter data. *In 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2)* (pp. 1-4). IEEE.

Hillebrand, L., Pielka, M., Leonhard, D., Deußer, T., Dilmaghani, T., Kliem, B., Loitz, R., Morad, M., Temath, C., Bell, T., Stenzel, R., & Sifa, R. (2023). sustain.AI: a Recommender System to Analyze Sustainability Reports. *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023),* June 19–23, 2023, Braga, Portugal. ACM.

Hoitash, R., Hoitash, U., & Morris, L. (2021). eXtensible business reporting language (XBRL): A review and implications for future research. *Auditing: A Journal of Practice & Theory*, *40*(2), 107-132.

Janik, A., Ryszko, A., & Szafraniec, M. (2020). Greenhouse Gases and Circular Economy Issues in Sustainability Reports from the Energy Sector in the European Union. *Energies, 13*(22), 5993. https://doi.org/10.3390/en13225993

Kang, Y., Cai, Z., Tan, C.W., Huang, Q. and Liu, H., (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics, 7*(2), pp.139-172.

Kolk, A. (2005). 'Sustainability reporting'. *VBA Journal, 21*(3), pp. 34–42.

Koskentalo, E. (2020). Mitä on XBRL?. *XBRL Suomi*. Available at: https://fi.XBRL.org/faq/mita-on-XBRL/ (Accessed: 5 May 2025).

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W.P., Nuzumlalı, M.Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R.A., Krumholz, H.M. & Radev, D. (2022). Neural natural language processing for unstructured data in electronic health records: a review. *Computer Science Review*, *46*, 100511.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:*1301.3781. Available at: https://arxiv.org/abs/1301.3781 (Accessed: 13 May 2025).

Müller, O., Junglas, I., Brocke, J. V., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289-302.

Müller, O., Fay, M., & Vom Brocke, J. (2018). The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of management information systems, 35*(2), 488-509.

Nasreen, T., Baker, R., & Rezania, D. (2023). Sustainability reporting–a systematic review of various dimensions, theoretical and methodological underpinnings. *Journal of Financial Reporting and Accounting*. DOI 10.1108/JFRA-01-2022-0029

NSE. (2025). XBRL Information. National Stock Exchange of India. Available at: https://www.nseindia.com/companies-listing/XBRL-information (Accessed: 1 May 2025).

OpenAI. (2023). GPT-4 Technical Report. Available at: https://cdn.openai.com/papers/gpt-4.PDF

OpenAI. (2024a). GPT-4o system card. [online] Available at: https://openai.com/index/gpt-4o-system-card/ (Accessed 18 Mar. 2025).

OpenAI. (2024b). Hello GPT-4o. Available at: https://openai.com/index/hello-gpt-4o/ (Accessed: 16 May 2025).

OpenAI. (2025). What is ChatGPT Pro? Available at: https://help.openai.com/en/articles/9793128-what-is-chatgpt-pro (Accessed: 16 May 2025).

PwC. (2020). Why XBRL is the future of ESG reporting. PwC. Available at: https://www.pwc.com/us/en/tech-effect/ai-analytics/why-XBRL-is-future-of-esg-reporting.html

PwC India. (2021). Business Responsibility and Sustainability Report And attempt tp mainstream ESG. Available at: https://www.pwc.in/assets/PDFs/consulting/esg/business-responsibility-and-sustainability-report.PDF .

PwC India. (2024). 51% of India's top 100 listed companies disclosed their Scope 3 data for FY23 despite it being a voluntary disclosure in the BRSR: PwC India Report. Available at: https://www.pwc.in/press-releases/2024/51-of-indias-top-100-listed-companies-disclosed-their-scope-3-data-for-fy23-despite-it-being-a-voluntary-disclosure-in-the-brsr-pwc-india-report.html (Accessed: 13 May 2025).

Rainer, R.K., Prince, B., Sanchez-Rodriguez, C., Splettstoesser-Hogeterp, I. and Ebrahimi, S., (2020). *Introduction to information systems*. Hoboken, NJ: John Wiley & Sons.

Ramanan R., Warren. (2021). XBRL JSON tutorial. Available at: https://www.XBRL.org/guidance/XBRL-JSON-tutorial/ (Accessed: 16 March 2025).

Ramanan, R. (2024a). 'Narrative disclosure analysis with GPT-4', XBRL International. Available at: https://www.XBRL.org/narrative-disclosure-analysis-with-gpt-4/ (Accessed: 29 January 2024).

Ramanan, R. (2024b). Why structured data and definitions vastly outperform unstructured PDFs in LLM analysis. *XBRL International*. Available at: https://www.XBRL.org/why-structured-data-and-definitions-vastly-outperform-unstructured-PDFs-in-llm-analysis/ (Accessed: 5 February 2025).

Ramanan, R. & Warren, P. (2023). XBRL report formats – which one to choose. Available at: https://www.XBRL.org/guidance/XBRL-report-formats/ (Accessed: 27 March 2025).

Ritchie, H., Roser, M. and Rosado, P. (2024). CO_2 and Greenhouse Gas Emissions. *Our World in Data*. Available at: https://ourworldindata.org/co2-emissions [Accessed 1 May 2025].

Salah, M., Al Halbusi, H., Abdelfattah, F. (2023). May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research. *Computers in Human Behavior: Artificial Humans, 1*(2), p.100006. Available at: https://doi.org/10.1016/j.chbah.2023.100006

Sanderson. (2024). Transformers (how LLMs work) explained visually | DL5. Available at: https://www.youtube.com/watch?v=wjZofJX0v4M (Accessed: 13 May 2025).

Securities and Exchange Board of India. (2019). Extension of applicability of Business Responsibility Reporting (BRRs) to top 1000 listed entities from present requirement to 500 listed entities, based on market capitalization. *SEBI*. Available at: https://www.sebi.gov.in/sebi_data/meetingfiles/dec-2019/1576469077048_1.PDF [Accessed 26 Feb. 2025].

Securities and Exchange Board of India. (2021). SEBI issues circular on Business Responsibility and Sustainability Reporting by listed entities. *SEBI*. Available at: https://www.sebi.gov.in/media/press-releases/may-2021/sebi-issues-circular-on-business-responsibility-and-sustainability-reporting-by-listed-entities-_50097.html?utm [Accessed 26 Feb. 2025].ramanan

Smailhodžić, E., & Oehmichen, J., (2025). AI determinants of success and failure: The case of financial statements. *Munich: LMU Munich School of Management*.

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in information systems. *Information Systems Research*, *34*(2), 399-408.

Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A. and Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics, 13*(6), p.e1549.

Thornton, G. (2023). CSRD reporting: What you need to know. [online] Available at: https://www.grantthornton.com/insights/articles/esg/2023/csrd-reporting-what-you-need-to-know [Accessed 10 Feb. 2025].

Wong, A., Plasek, J. M., Montecalvo, S. P., & Zhou, L. (2018). Natural language processing and its implications for the future of medication safety:

a narrative review of recent advances and challenges. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *38*(8), 822-841.

XBRL. (n.d.a). XBRL glossary. Available at: https://www.XBRL.org/guidance/XBRL-glossary/#taxonomy-defined-dimension (Accessed: 18 March 2025).

XBRL. (n.d.b). Taxonomies. XBRL. Available at: https://www.XBRL.org/the-standard/what/key-concepts-in-XBRL/taxonomies/ (Accessed: 2 June 2025).

XBRL. (2025). EU sustainability reporting shake-up arrives — without impacting digital reporting. Available at: https://www.XBRL.org/news/eu-sustainability-reporting-shake-up-arrives-without-impacting-digital-reporting/# (Accessed: 5 May 2025).

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang S., Tong, H., Xiao, L. & Zhou, W. (2025). ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *Journal of Cleaner Production*, 489, 144572.