Framework Development and Data-Driven Approach for Automating Corporate ESG Governance Evaluation Using GPT Models

Yangjiang Yu

Master of Science
School of Informatics
University of Edinburgh
2024

Abstract

In this thesis, we investigate the use of state-of-the-art Natural Language Processing (NLP) techniques, in particular GPT models to assess corporate governance as a part of ESG performance metrics. Most existing ESG rating systems are manually based and have been criticized for non or counterintuitive ratings. In this paper, we investigate whether GPT models can be used to automatically review corporate annual reports based on the governance-related paragraphs and improve objectivity in ESG scoring. A large dataset that was constructed by aggregating metrics across several ESG rating agencies, governance may be useful for analyzing companies through the use of GPT-based models. The study assesses the performance of various batch processing techniques and prompt designs, thereby seeking to balance model accuracy against efficiency. Findings highlight the potential of using GPT models for a higher quality ESG evaluation; however, barriers to process length and alignment with existing ESG ratings remain. This work should be further expanded to cover more data sets, include other sources of information, and deepen it using the power of AI-based evaluation techniques that can help in making ESG assessments more transparent and reliable.

Research Ethics Approval

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Yangjiang Yu*)

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Tiejun Ma, whose invaluable guidance and support have been pivotal to the success of this project. Through our regular meetings, Professor Ma ensured that I remained on track with the overall direction of the project, preventing me from veering off course. His insightful feedback and constant emphasis on maintaining rigor and professionalism in my research have greatly contributed to the quality of this dissertation.

I would also like to extend my heartfelt thanks to Mengyu Wang, a research assistant under Professor Ma's guidance. Mengyu's technical expertise was an immense help when I encountered challenges during the development process. His ability to provide clear and practical suggestions often illuminated the path forward when I felt stuck. Additionally, during moments of stress, Mengyu's encouragement was a source of great comfort, and his timely responses to my inquiries made all the difference in keeping the project moving forward. His support has been invaluable throughout this journey.

Table of Contents

1 Introduction 1

1.1 Background 1

1.2 Motivation 3

1.3 Objective 3

2 Literature Review 5

2.1 Existing ESG Evaluation Methods 5

2.2 Research on large language models in processing financial texts . . . 8

2.2.1 Handling Long Documents in Financial Tasks 8

2.2.2 Structured vs. Unstructured Text 8

2.2.3 Current Limitations and Future Directions 9

2.3 The way machine learning methods currently being used in the field of ESG 10

2.4 Summary of Literature Review 10

3 Methodology 12

3.1 Data 12

3.2 GPT Framework 15

3.2.1 Extracting Governance Information from Annual Reports . . . 16

3.2.2 Assess extracted information by Key Metrics 17

3.2.3 Summary and Rating 18

3.3 Evaluation Method 20

4 Results and Discussion 22

4.1 Extracting Governance Information 22

4.2 Assess extracted information by Key Metrics 23

4.3 Get Summary and Rating 26

4.4 Summary 30

5 Conclusion and Future Work 31

Bibliography 33

A Data Sources of each Key metrics in Governance from MSCI 35

Chapter 1

Introduction

1.1 Background

ESG, which stands for Environmental, Social, and Governance, is a framework used to evaluate the sustainability and societal impact of investments in companies or businesses. ESG criteria are a set of standards for a company's operations that socially conscious investors use to screen potential investments. Environmental criteria consider how a company performs as a steward of nature, social criteria examine how it manages relationships with employees, suppliers, customers, and the communities where it operates, and governance deals with a company's leadership, executive pay, audits, internal controls, and shareholder rights. The significance of ESG has grown in the financial sector, driven by an increasing awareness of the impact of corporate activities on the environment and society, alongside a demand for greater corporate responsibility [6]. This shift reflects a broader trend towards integrating non-financial factors into investment analysis, recognizing that long-term financial performance is increasingly influenced by sustainability and social factors [15].

Within just a few years, the market value of ESG-focused investments has risen to amazing proportions, testifying to the enhanced value that markets are beginning to place in sustainability. Already in 2021, ESG-focused portfolios were nearly $40 trillion in size in assets under management; by 2025, this is forecast to reach $53 trillion—about one-third of all global assets under management. This growth is driven by investors who increasingly require that their investments should not only yield financial returns but also help achieve positive societal and environmental goals[6]. COVID-19 has accelerated this further, as the imperative to secure a 'sustainable recovery' underlined the importance of integrating ESG factors into investment decisions. Consequently, ESG has moved from the periphery to mainstream; what started as a niche-like investing style now represents sustainable funds that continue fetching substantial inflows in professional asset management, showcasing successor taste and strong growth among investors in sustainability-themed investments[2].

ESG ratings are designed to grade the quality of a company's environmental, social, and governance practices, providing investors with insight into non-financial risks and opportunities that may bear on a company's performance. However, methods for determining these ratings vary and become complex, often leading to inconsistencies between rating agencies. Current ESG rating methodologies generally fall into two main categories: internal disclosure frameworks and external evaluation frameworks. The internal disclosure frameworks, to a large degree, are grounded on the self-reported information from companies, such as through sustainability reports, regulatory filing requirements, and other forms of voluntary disclosure. These frameworks, such as the Global Reporting Initiative and Sustainability Accounting Standards Board, aim to standardize the reporting of ESG activities that companies undertake, hence it is easy to compare and assess the ESG performance across firms[6]. The external assessment frameworks consist of third-party assessments of a company's ESG performance. Rating agencies like MSCI, Sustainalytics, and ISS ESG source data from a wide range of independent public and private sources to deduce risks and opportunities of any company pertaining to ESG. With the different approaches—privately owned algorithms and weighting schemes—ESG score providers use to come up with ESG scores that reflect a firm's relative performance vis-à-vis its peers in the same industry or relative to global benchmarks[15].

Inconsistencies in ratings for the same company, because of a lack of standardization across frameworks, along with many differences of scope about what is measured, as well as different methods applied for weighting and aggregating ESG factors, cleanup investment decisions. Furthermore, the large number of factors considered under ESG, along with the difficulty in quantifying these very factors, adds complexity to building reliable and comparable ESG ratings. Despite all odds, ESG ratings present a critical tool for investors, albeit requiring an interpreter and attention to underlying methodologies[3, 14].

Advanced NLP technologies, especially those using GPT models, have developed potential for analyzing financial texts in corporate disclosure settings, more specifically annual reports, within the past years. While direct applications of GPT in estimating ESG performance are still in the early days and reportedly relatively scant, the ability to process and understand huge qualitative data sets which the model possesses opens wide possibilities in this regard. GPT models are designed to extract and synthesize complex information from unstructured texts with ease; hence, they will be quite ideal in parsing nuanced language found in corporate reports. This understanding and text generation capability allows researchers and analysts the potential to open new perspectives in assessing corporate performance on environmental, social, and governance principles and possibly automate what has so far been a manual and time-consuming process[10, 9].

1.2 Motivation

The current approaches to ESG scoring have a number of important limitations that make most of them not very effective and less accessible. For instance, traditional ESG assessments are usually based on manual processes such as analyst-driven evaluations and questionnaires. Such approaches take up enough time and, at the same time, are not transparent, with criteria and processes that result in those scores usually black-boxed to end-users, particularly small investors. Most investors have access only to the final ESG score and lack deep insight in the way these scores were ultimately derived[4, 7]. This lack of transparency and accessibility often results in misinformed investment decisions and general distrust in the precision of ratings.

Against this backdrop, there is an urgent need to examine the next generation of NLP technologies, epitomized by GPT, for the delivery of much greater efficiencies and objectivities in ESG scoring. Specifically, GPT models are potentially able to automate a large volume of unstructured textual data from corporate disclosures to drive more consistency and objectivity in assessing corporate ESG performance. It could be the case that more transparency and granularity from GPT capabilities in the evaluation process will help institutional and small investors alike to derive deeper insights into ESG considerations beyond just a score [18, 12].

1.3 Objective

Knowing that the literature lacks any easily accessible comprehensive ESG datasets, the first objective of this research became one of collecting a robust ESG dataset. This should have ESG scores and associated assessments for different companies against various rating agencies. Through the acquisition of data from multiple sources, including web scraping of ESG rating sites, it is the ambition to come up with a dataset detailing all the information available on every company's ESG performance against the different evaluation criteria.

It also aims to explore the possibility of using GPT models to measure and evaluate corporate ESG performance. This study is proposed to delve into exactly what GPT models are capable of by processing ESG data, in order to identify the benefits and challenges of such advanced technology in natural language processing within domains pertinent to ESG evaluation. Ultimately, this research will provide important findings and signs for future research, thus opening the way to develop increasingly effective tools to assess ESG performance. Investors are therefore going to achieve a greater understanding of a company's sustainability practices by having more help in making effective and informed investment decisions.

Chapter 2

Literature Review

2.1 Existing ESG Evaluation Methods

The existing approaches to ESG reporting can be grounded in various formats, which have been developed in order to allow consistency in reporting and the disclosure of non-financial information from the dimensions of environmental, social, and governance corporate performance. These frameworks include the Global Reporting Initiative, the Sustainability Accounting Standards Board, and the Task Force on Climate-related Financial Disclosures. All of the frameworks attempt to provide better understanding in corporate sustainability efforts through clear, comparable data for investors. Nevertheless, a great number of the initiatives are voluntary, lacking coherence in terms of reporting, clearly affecting ESG assessment effectiveness. Secondly, on grounds of complexity, it is hard to come up with broadly accepted metrics for certain ESG factors, especially social and governance factors [6].

One of the prominent problems in the ESG rating space today is that rating agencies tend to diverge, including MSCI, Sustainalytics, and ISS ESG. All these providers have their methodology for estimating companies based on their respective ESG performance. These differently very approaches often have obviously different results on the ESG ratings for the same company. These differences originate from the different sets of criteria applied by each agency, how they are weighted, and the sources of data used. Some rating agencies, for instance, are more focused on environmental risks, while others may be biased toward companies' governance practices. Since the rating methodologies used by these agencies are not standardized, it becomes hard for investors to compare ESG scores across agencies and industries [15]. This has therefore led to an increased call for more coherent and harmonized ESG assessment practices[3].

Worth noting, given this complexity and lack of standardization in ESG evaluations, it becomes imperative to learn how different rating agencies construct their ratings. Each rating agency uses some methodology for constructing its ratings; accordingly, these lead to different results even when rating the same company. It becomes important for an investor who relies on ESG scores for making investment decisions. The paper provides an extensive comparison of ESG rating methodologies by some of the key agencies: MSCI, FTSE Russell, Sustainalytics, and S&P Global.

1. MSCI ESG Ratings

MSCI's ESG Ratings Methodology scores companies on their exposure to material risks and respective risk management. The ratings are relative to industry, covering key environmental and social issues besides governance structures. Companies are rated by MSCI across three main pillars: Environmental, Social, and Governance. Under these three pillars come several key issues like climate change, human capital, and corporate governance. These are the key issues scored, and the final ESG rating is derived based on a weighted average of the scores, adjusted relative to industry peers. The MSCI then assigns ratings based on performance with respect to relative ESG risk management on the seven-point scale from AAA to CCC [19].

2. FTSE Russell ESG Ratings

FTSE Russell uses a granular scoring system that includes rating companies against over 300 indicators organized into 14 themes and three main pillars: Environmental, Social, and Governance. A considerable proportion of the data utilized in the computation of the FTSE Russell ESG ratings exists in the public domain. The methodology focuses on the materiality of ESG issues with regard to industry and region. Moreover, FTSE Russell embeds the United Nations Sustainable Development Goals as a part of its scoring system. That definitely makes it all-inclusive for measuring corporate ESG performance. The ratings find application in constructing ESG-focused indices, and companies' scores impact their weight in these indices [18].

3. Sustainalytics ESG Risk Ratings

Sustainalytics assesses the unmanaged ESG risk of companies, focusing on material ESG issues. The ESG Risk Ratings measure a company's level of exposure to ESG risks and how well it manages those risks. Sustainalytics provides a single score that reflects the overall ESG risk of the company, designed to be comparative across various industries. This focused approach to risk gives investors a very clear view of where a company stands in terms of potential financial risks related to ESG issues[12].

4. S&P Global ESG Scores

Industry-specific questionnaires are conducted by S&P Global by directly approaching companies to get the desired ESG information. In case of non-responding companies, SP Global completes the assessment using publicly available information. Companies are rated against weighted criteria on themes related to environment, social, and governance aspects that matter for each particular industry. These ratings get updated at periodic intervals so that continuous ESG-related controversies and new emerging information get encapsulated and timely relevant data gets passed on to investors [11].

As a way of determining the sources of data that would be needed for the later stages of this research, a desktop review targeting all the ESG rating methodologies publicly available was carried out. Out of these rating agencies, only MSCI gave an elaborate description of how it carries out its evaluation and the exact Key Metrics applied. As shown below, this involved scrutinizing MSCI's methodology and accounting for and counting all data sources necessary to compute each Key Metric[16]. Table 2.1 presents the five most frequently required sources of data to compute the Key Metrics in each ESG pillar.

Table 2.1: Data Sources for MSCI ESG Key Metrics

Pillar	**Data Source**	**Count**	**Frequency**
Environment	Company's Annual Report	85	0.3219697
Environment	MSCI ESG Research	77	0.2916667
Environment	Refinitiv	35	0.1325758
Environment	INFORM Risk Index	5	0.0189394
Environment	Inter-Agency Standing Committee and European Commission	5	0.0189394
Governance	Company's Annual Report	76	0.4724096
Governance	Proxy Statement	48	0.2981367
Governance	Company's Corporate Charter and Bylaws	18	0.1118012
Governance	Corporate Governance Documents	6	0.0372670
Governance	Regulatory Filings	4	0.0248447
Social	Company's Annual Report	111	0.3313428
Social	MSCI ESG Research	58	0.1731343
Social	Refinitiv	27	0.0805970
Social	World Bank	9	0.0268657
Social	Bureau of Labor Statistics	6	0.0179105

This analysis has revealed that the single largest source of data within MSCI's ESG rating process comes from an organization's annual report. More importantly, every Key Metric except a few under the Governance thief pillar can be assessed based on information contained in the annual reports. Not only does this finding underscore the critical role of an annual report in an ESG assessment, especially in evaluating governance-related factors, but it also establishes their use as a prime data source for ESG performance analysis in this research.

2.2 Research on large language models in processing financial texts

2.2.1 Handling Long Documents in Financial Tasks

Although large language models, such as GPT-3 and GPT-4, showed their potential for handling natural language processing tasks quite a long ago, their application in finance is quite unique and meaningful, with additional challenges mainly when one has to deal with long and complex financial documents. One of the most evident problems is that such financial reports usually run to hundreds of pages in length, which creates problems for models limited in tokens. Agrawal et al[1]. presented another goal-directed extractive summarization method, specifically oriented to financial documents and more exactly to Management Discussion and Analysis sections from 10-K filings. The results show that the ability of LLMs in Summarizing long documents after task-specific fine-tuning, but the models still creation problems in preserving coherence across longer text.

On the other hand, BloombergGPT[17] emphasizes that to handle the intricacies of financial text, there is a need for domain-specific training. This model was tailored for undertaking financial tasks and hence had an incorporation of large-scale corpora of financial data so as to enhance the comprehension and generation capabilities. However, the capacity of BloombergGPT to manage long documents remains constrained by token limits, and therefore, though domain-specific models are promising, handling long texts complex in nature remains far from being handled within this domain.

2.2.2 Structured vs. Unstructured Text

Another challenge for LLMs would be structured and unstructured financial data processing. Structured data usually refers to tables, financial metrics, and charts in a financial report, while narrative texts form part of the unstructured data. A model called PIXIU[21] combining this structure with unstructured data ro provides all-rounded information on finance. PIXIU provides integration of both textual and tabular data, thus demonstrating the strengthening of LLMs to accommodate varied input types. But structured data still remains a stretch to incorporate into LLMs because forms are diverse and require an understanding by the model of what data to extract from each.

The data-centric financial LLMs[5] further refine the way structured and unstructured data may be processed together. This approach focuses on preprocessing financial texts and integrating domain-specific knowledge prior to improving the model's performance. Alone, the structured data of the financial tables is combined with the unstructured narrative sections to create more detailed and realistic financial analysis studies. Results from these models have shown structured data could be handled as well as the unstructured one, but still with a lot of room for improvement in terms of consistency and accuracy.

2.2.3 Current Limitations and Future Directions

Though much improvement has been noted in the application of LLMs to financial tasks, several limitations remain. CFGPT[13] cites that although LLMs do well in text generation, their reasoning and analysis capabilities within the financial domain are not quite mature yet. The model performs poorly on complex reasoning that requires a deep understanding of basic finance-based principles, especially for domain subtleties. This underlines an even greater need for fine-tuning and consequent training on relevant financial datasets so as to enhance the interpretive abilities of the model.

Furthermore, FinGPT[8] investigates instructional tuning over financial tasks, thus showing that an LLM is adaptable to specific financial benchmarks. Again, progress is hampered by the lack of big, high-quality labeled datasets in the domain of finance. More robust datasets and corresponding benchmarks, such as the one proposed by FinGPT, are critical to drive further improvements in LLM performance within finance.

Though LLMs have shown some potential in financial document processing, how well these models work in handling long documents or integrating structured data remains a subject of active research. In future research, there should be more focused development of domain-specific datasets and leveraging ways to enhance reasoning ability and analytical capability for the financial domain. Hence, further efforts in instruction tuning and domain-specific pretraining are highly critical to going beyond such limitations and unlocking the full potential of LLMs in finance.

2.3 The way machine learning methods currently being used in the field of ESG

In recent times, the marriage between Machine Learning methodologies and analysis of Environmental, Social, and Governance (ESG) factors has been deemed very paramount, especially to most successful companies and investors who hope to have an effective and scalable approach. The most important aspect on which the ML-ESG work of 3DS Outscale underlines is to be able to bring forward multilingual data sets so that identification of types of ESG impact can be improved across the board. This series of shared tasks was co-organized with several institutions and underlined the importance of integrating machine learning with domain knowledge in conducting appropriate analysis of ESG data [20].

These projects are using the powerful functionality of modern language models—like GPT and BERT—to process ESG-related news and reports in support of the automatic classification of risks and opportunities. Much of the existing research, however, is relatively superficial and limited to classification tasks featuring short texts. For instance, the typical application has been to categorize ESG-oriented news articles according to predefined classes. Although such models have been effective in detecting trends within ESG data, the models are still constrained with respect to their capacity for processing larger and more complex documents and undertaking deeper analysis. This shows the need for further research into the models' capacity to process full-scale ESG reports and complex financial data[22].

2.4 Summary of Literature Review

The place that machine learning and large language models hold within ESG analysis is so full of opportunities as it is rife with challenges. Traditional approaches to ESG assessment are typically underpinned by well-established frameworks such as the Global Reporting Initiative, the Sustainability Accounting Standards Board, and the Task Force on Climate-related Financial Disclosures. Across the rating agencies, like MSCI, FTSE Russell, Sustainalytics, and S&.P Global, inconsistencies in methodologies have resulted in different ESG scores for the same firms, and this is clearly making it very difficult for investors to make decisions. This clearly underlines the necessity of more standardization and transparency in ESG evaluation practices.

In the financial domain, LLMs have demonstrated enormous potential for processing financial texts, namely GPT-3 and GPT-4. However, they do not perform very well on long and complex documents such as ESG reports. In fact, the lower token capacity and poor coordination of inferences across extended articles limit the potential of a model to be applied to tasks that entail the processing of a large volume of data related to ESG criteria. Integrating structured and unstructured data is a hard task, especially if it is going to be done for financial analysis.

While most models, such as PIXIU, have been put forth to try and solve these problems, further refinement is necessary for its correct and complete analysis.

The current research on the integration of ESG with ML has been mainly limited to short-text classification tasks, which indeed does not leave much room for conducting detailed analysis of the comprehensive ESG reports. Furthermore, the G from ESG is greatly overshadowed by other constituents, such as Environment and Social, while in fact, it is impossible to have good governance or good corporate governance, leadership, and accountability without it. The low representation of governance within the ESG analysis underscores the demand for more specialized research into this. My research should address these challenges by the following: First, improvement in analysis would be possible with the expansion and standardization of ESG datasets, especially in factors relating to governance. Second, LLM-based analysis can maintain the rigor brought by conventional established methods for ESG evaluation, provided that it can adapt them effectively. Finally, increasing the capacity of LLM to hold long ESG documents and shifting attention to governance analysis will go a long way in balancing and rounding out the approach toward ESG evaluation. These are directions that can be pursued, greatly improving the transparency, consistency, and depth of AI-driven work related to ESG assessment in any future setting.

Chapter 3

Methodology

This section describe how this study extracted and processed information pertaining to governance from corporate annual reports and inputted it into the GPT-based models. This section will be further divided into several subsections covering data collection, framework design, and evaluation methods. The first subsection is on the dataset, which describes the sources and structure of the ESG data gathered from three agencies. Next, the GPT framework used for governance information extraction, assessment against Key Metrics, summary, and scoring will be outlined. Finally, an explanation on the evaluation methods adopted in checking for accuracy and effectiveness of the model's performance will follow. This methodology will ensure a systematic and transparent approach while assessing corporate governance performance within the ESG context.

3.1 Data

The dataset collects information on the ESG performance of companies in the FTSE-Allshare and Russell-3000 indices from three platforms: MSCI, S&P Global, and Sustainalytics.

The dataset dimensions are **2906 x 204**, with each row representing a company for which at least one ESG score is available from the three platforms. The number of companies from the FTSE-Allshare index with at least one, two, and all three ESG scores are **351**, **160**, and **62** respectively. For the Russell-3000 index, the corresponding numbers are **2604**, **1270**, and **448**.

MSCI, S&P Global Ratings, and Sustainalytics are overall ESG assessment frameworks offering a multidimensional view of a company's ESG performance. MSCI measures a company's exposure to ESG risks and how it manages these risks com-

Attribute_Name	Description	Variable Type
Ticker	Ticker of the Company (no unique because different market may have the same Ticker)	str
Industry	Industry of the companies	str
Group	FTSE or RUSSELL, indicate the company stocks belong to which Index	str

Table 3.1: General Company Data

pared to industry peers to identify the areas where the company leads or lags. S&P Global Ratings give granular ESG performance across a range of pillars with industry comparisons. While Sustainalytics focuses more on rating companies based on both their ESG risk exposure and management. These data sources together therefore reflect different vantage points on corporate ESG performance by combining assessment of performance, industry benchmarking metrics, and risk evaluation.

Attribute_Name	Description	Variable Type
MSCI_Score	ESG Rating of the Company evaluated by MSCI (aaa to ccc)	str
History_Score	Historical ESG scores of the company over time	list of dictDate:score
LAGGARD	Key Issues where the company lags behind the industry	list of str
AVERAGE	Areas where the company has an average track record	list of str
LEADER	Key Issues where the company leads its industry	list of str
Banned_Controversial_Weapons	Involvement in activities related to banned controversial weapons	str
Gambling	Involvement in activities related to gambling	str
Tobacco	Involvement in activities related to tobacco	str
Alcohol	Involvement in activities related to alcohol	str
Decarbonization_Target	Information on whether the company has a decarbonization target	str
Decarbonization_Target_ITR	Involvement in the Implied Temperature Rise calculation	str
Decarbonization_Target_Year	Target year of the decarbonization target	str(int)
Decarbonization_Target_Comprehensiveness	% of company footprint covered by the target	str(num)
Decarbonization_Target_Ambition	Projected reduction per year to meet the stated target	str(num)
Controversies_{KeyIssuesName}	Involvement in controversies related to specific ESG issues	str

Table 3.2: Data from MSCI

Figure 3.1 illustrates the correlation analysis of four continuous score data points, which are highly relevant to this study, across a sample of 2,906 companies. The results are shown in the figure.

1. **Moderate Correlation Between Existing ESG Scores:** The correlation coefficients between the ESG scores from the three different agencies (S&P Global, MSCI, and Sustainalytics) are moderate, with absolute values below 0.4. This suggests that while there is some level of agreement between the agencies, the scores are not strongly correlated. Therefore, if the scores generated by my GPT-based model exhibit correlations with any of the existing scores that approach or exceed 0.4, this could be considered a sign of validity. This threshold is based on the observed correlations between industry-standard ESG scores, making it a reasonable benchmark for evaluating

Attribute_Name	Description	Variable Type
SP_Score	Overall ESG Rating of the company by S&P Global	str(num)
Score_CSA	Actual score based on CSA Required Public Disclosure	str(num)
Score_Modeled	Score based on modeling approaches	str(num)
Score_Env	Score of the Environmental Pillar	str(num)
Env_Score_Industry_Mean	Industry mean score of the Environmental Pillar	str(num)
Env_Score_Industry_Max	Industry maximum score of the Environmental Pillar	str(num)
Score_Social	Score of the Social Pillar	str(num)
Social_Score_Industry_Mean	Industry mean score of the Social Pillar	str(num)
Social_Score_Industry_Max	Industry maximum score of the Social Pillar	str(num)
Score_Gov	Score of the Governance Pillar	str(num)
Gov_Score_Industry_Mean	Industry mean score of the Governance Pillar	str(num)
Gov_Score_Industry_Max	Industry maximum score of the Governance Pillar	str(num)
RPDRDARP%	CSA Required Public Disclosure Rate Data Availability	str(num%)
RPDRDARP	CSA Required Public Disclosure Rate Data Availability	str
RPDCSAMIS	Required Public Disclosure CSA Maximum Industry Score	str(num)
PSRPDR	Potential Score based on Required Public Disclosure Rate	str(num)
ASRPD	Actual Score based on Required Public Disclosure	str(num)
ADRDARP%	Additional Disclosure Rate Data Availability	str(num%)
ADRDARP	Additional Disclosure Rate Data Availability	str
ADCSAMIS	Additional Disclosure CSA Maximum Industry Score	str(num)
PSADR	Potential Score based on Additional Disclosure Rate	str(num)
ASAD	Actual Score based on Additional Disclosure	str(num)
NQMA	Number of questions based on modeling approaches	str(num/num)
Score_{KeyIssueName}	Score of the company in the most relevant Key Issue	str(num)
IndustryMaxScore_{KeyIssueName}	Maximum Score of the company's industry in the most relevant Key Issue	str(num)
IndustryMeanScore_{KeyIssueName}	Average Score of the company's industry in the most relevant Key Issue	str(num)

Table 3.3: Data from S&P Global

Attribute_Name	Description	Variable Type
Name	Company Name at sustainalytics.com	str
Market	Exchange code on which the company is listed	str
SUS_Score	ESG Risk Rating of the company evaluated by Sustainalytics	str(num)
Risk	ESG Risk Level of the company evaluated by Sustainalytics	str
URL	URL showing the results of Sustainalytics ESG's research on the company	str(url)

Table 3.4: Data from Sustainalytics

the effectiveness of the model-generated scores.

2. **Strong Correlation Between Governance and Overall ESG Scores:** Results from the analysis yield a strong correlation of 0.94 between the S&P Global Governance score and its overall ESG score. This strong relationship underlines that for a company, governance forms a very substantial part of its overall ESG performance. While the parameters of evaluation differ across various rating agencies, which further adopt different working methodologies, it still does appear that governance is one of the major driving factors in ESG performance. It is justified that the focus of the research

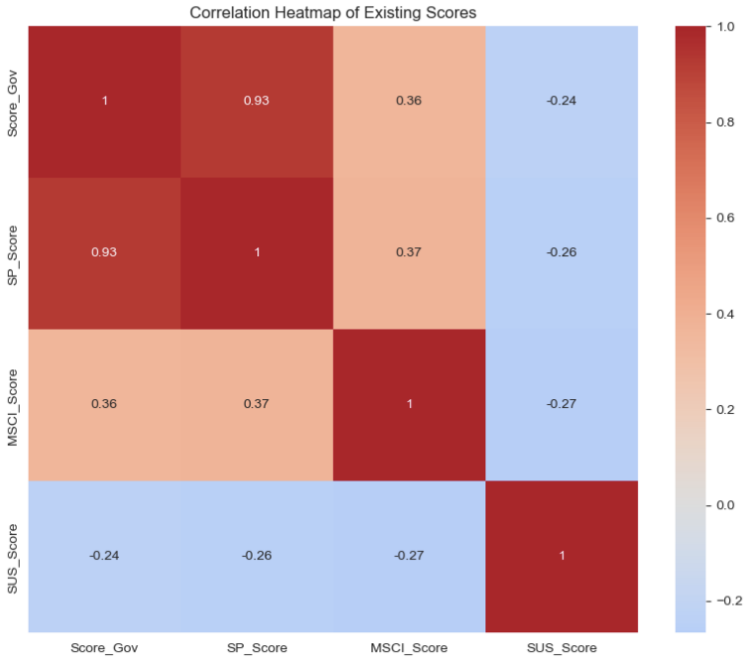


Figure 3.1: Correlation Heatmap of Existing ESG Scores. Score_Gov and SP_Score are both from SP Global, with Score_Gov representing the Governance score and SP_Score representing the overall ESG score.

is on governance, for the simple reason that metrics pertaining to this dimension are maximum in overall ESG ratings, thus trying to understand a very valuable insight into the companies' sustainability practices.

3. **Negative Correlation of Sustainalytics Scores:** Notably, the Sustainalytics ESG score (SUS_Score) is negatively correlated with the scores from the other two agencies. This observation aligns with Sustainalytics' rating system, where a higher score indicates greater ESG risk exposure, implying weaker ESG performance.

By accessing AnnualReport.com, the latest annual reports of all listed companies on 5 exchanges (LSE, NYS, NAS, ASX, TSE) were collected. As discussed in Section 2.1, through a detailed examination of MSCI's methodology, the names, definitions, measurement methods, and value ranges of a total of 327 Key Metrics across 33 Key Issues under the three ESG pillars were identified.

3.2 GPT Framework

With the dataset now collected, the thing need to do next is to utilize this data on a GPT-based framework that would focus on governance information from annual

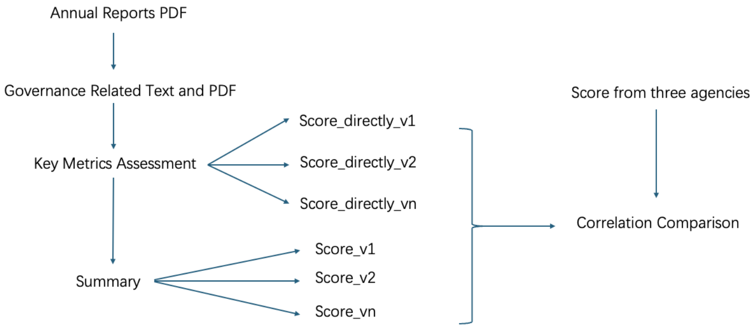


Figure 3.2: Flow chart of GPT Framework

reports. Because of the length of the annual reports, the framework will extract only the relevant portions related to governance rather than processing the whole document. This extracted information shall then get analyzed from different aspects using MSCI's Key Metrics for rating the governance performance. These Key Metrics are summarized at the end of the analysis to form a comprehensive conclusion, which is then used to assign a governance score for the company. The approach is for an easier and more accurate governance assessment.

Unless otherwise noted, the GPT-4o-mini model is used, which offers affordable pricing at \$0.15 per 1 million input tokens and \$0.60 per 1 million output tokens.

3.2.1 Extracting Governance Information from Annual Reports

The extraction of information related to governance from annual reports is an important step in measuring the company's performance on ESG, more particularly on the Governance pillar. Given the comprehensiveness of the annual report, it will be inefficient and useless to process the document as a whole. That is why this framework will apply step-by-step processes in identifying and extracting only relevant parts that deal with issues concerning governance. The solution is designed to efficiently target governance-related content through the combination of text preprocessing techniques and GPT-based models in the extraction process.

1. **PDF Conversion and Text Preprocessing:** The annual reports, originally provided in PDF format, are converted into text files. This conversion is essential for enabling text-based analysis, as models and regular expressions cannot directly operate on PDFs. Once converted, the text is split by page breaks (") to separate the document into individual pages, allowing for more flexible and focused processing.

2. **Mapping Table of Contents to PDF Pages:** Regular expressions are utilized to locate the Table of Contents (ToC) page, which is then extracted. Due to potential discrepancies between the page numbers listed in the ToC and the actual PDF pages, the model extracts sample pages to map the ToC sections to the correct PDF pages. This step ensures that the correct governance-related sections are extracted later in the process.

3. **Generating Section Range Dictionary:** A GPT-based model processes the ToC and generates a dictionary that maps section titles to corresponding PDF page ranges. This dictionary is crucial for identifying the sections most likely to contain governance-related information.

4. **Extracting Governance-Related Sections:** The GPT model, utilizing predefined governance-related keywords and concepts, analyzes the section range dictionary to filter out irrelevant content. The model then returns a refined index list that highlights the sections most pertinent to governance.

5. **Slicing Text and PDF Content:** Finally, the index list is used to slice the original PDF and converted text, extracting only the governance-related sections. This targeted extraction minimizes unnecessary data processing and ensures that the subsequent analysis focuses exclusively on governance-related content.

This extraction process makes the analysis of governance more efficient and accurate by having to process only parts of the relevant texts. By dealing with only that content relevant to governance drawn from annual reports, the framework reduces, therefore, computational overload and increases precision in assessments relevant to governance.

3.2.2 Assess extracted information by Key Metrics

Assessment using extracted text After extraction, the next process in the framework is assessment. The assessment should be performed using the Key Metrics of MSCI. The performance objective of this stage in the framework is to obtain a meaningful, expert-informed assessment at multiple dimensions of governance. A major point of focus in this paper is the assessment of efficiency of processing of several Key Metrics at the same time. bringing out a balance between the related trade-offs in terms of accuracy and time taken in processing.

GPT models harness the assessment process to evaluate governance-related content against defined Key Metrics. Each Key Metric is evaluated using well-constructed prompts that have been defined to elicit structured responses from GPT. To answer the question in each Key Metric, in sum, GPT must output four critical answers cumulatively: (1) a characterization of the bench-mark for that metric, (2) the estimate of the score the governance text should receive to align with the benchmark, (3) a flag system indicating the appropriate information is missing or questionable, and (4) a final score creating the estimate of the score. If there is insufficient detail in governance content for a specific Key Metric, GPT scores it based on the ”Typical Scoring Contribution” as noted in MSCI methodology.

The GPT model is configured with a low temperature setting (0.2) so that it might be more thoughtful about accuracy and consistency than it is about inventing novel information. This moves away from random output generation from the GPT and ensures it is accurate and dependable in its judgments, rather imaginative and varied.

An important area of research within the current framework is that of batch processing. As processing various Key Metrics at each processing step might be faster, the evaluation process is compromised in terms of focus and precision in the model. On the other hand, if there is an evaluation at each step for a specific Key Metric, the process might be more time-consuming but the quality is better. In this framework, this trade-off issue—between efficiency and accuracy—constitutes the core research topic, and different batch sizes are discussed in terms of finding the right balance. The evaluation results, comprising scores, assessments, or any flags raised in the course of the analysis, are output in a structured format. Post-evaluation, it will clean up all related resources such as threads, assistants, and the like, to free up resources efficiently. This approach enables comprehensive and systematic evaluation of the performance of governance. For that, optimization of this batch processing of Key Metrics will further improve both efficiency and accuracy.

3.2.3 Summary and Rating

In the **Summary** phase, what is desired is to synthesize the governance performance assessments of each company into a summary based on one voice. Drawing from outputs from the previous phase, where individual Key Metrics have been assessed, this step will seek to pull together all governance-related information that pertains to different dimensions into an integrated narrative for each company.

1. Multidimensional Information Consolidation: The model consolidates the results from an assessment of the governance-related Key Metrics across various dimensions to form a holistic summary. Here, besides integrating individual metrics, this consolidates the relationships binding them to set up a broader picture of the company's governance performance. A good summary will correctly reflect this combined performance across these dimensions of governance and clearly express this to the reader. This generated text is an important input for the rating phase that follows, where a further analysis of governance performance shall be made.

2. Prompt Design in Summary Phase: The prompt here is designed to keep as much information from the preceding Key Metrics assessments as possible. Here, a relatively simple structure for a prompt is employed to guide the model in synthesizing already evaluated data so that key details embedded therein are preserved in the summary. This means that it is more of the aspect of integrating assessment results rather than garnering new inferences, enabling the model to produce a summary reflecting comprehensive governance performance for the company.

In the **Rating** phase, although the governance analysis is performed for each company individually, the rating process involves comparing the performance of multiple companies within a unified context. The ratings are not merely a compilation of summaries but rather a comparative analysis of each company's governance performance in relation to others, ensuring that the ratings are consistent and comparable.

1. Multi-Company Rating Environment: In this phase, all summaries of the companies are fed in at one go into a single rating framework. This will enable the model to pick up subtle differences between companies and guarantee more reliable and comparative rating results across the dataset.

2. Importance of Prompt Engineering: Since the rating phase involves more subjective judgments compared to the objective assessments in earlier stages, prompt design plays a crucial role in determining the outcome. Various designs of the prompts could make a huge difference in how ratings should actually be done by the model; therefore, much experimentation and optimization would be necessary to ensure that ratings are correct and consistent. On its part, as related to issues of governance, this phase has an inherently high order of subjectivity. Prompt design refinement is, therefore, held to be of central importance for qualitatively improving consistency in rating outcomes.

In summary, the Rating phase focuses on ensuring comparability across companies through a unified rating process, while optimizing prompt engineering to enhance the accuracy of the governance performance ratings.

3.3 Evaluation Method

In this research, several model performance parameters are to be tested against the ground truth for its process to be effective and accurate. This is designed to validate the model for its automated extraction, analysis, and scoring of governance-related information from annual reports without manual intervention in the process.

First, the model's competence in the right extraction of governance-relevant information from annual reports. The intention is to make sure that the model focuses on those parts of the reports that carry the required governance content for further analysis. This is an important stage because unless relevant information is extracted appropriately, the latter analysis and scoring will be performed with incomplete or irrelevant data. This extraction will be evaluated by a manual review of the extracted text to confirm that the correct key governance sections are identified and captured, thus validating an ability of the model to parse the documents.

Second, model testing against the MSCI governance-related Key Metrics should be an important test of analytical accuracy. The companies' scores for each of the individual subcomponents are not available; however, overall ESG scores can be obtained. According to MSCI's methodology, one would expect a high correlation with the overall ESG score from the Governance score. I will estimate the performance of the model using the calculated Governance score based on the Key Metrics assessed by the model and compare these estimated calculated Governance scores reported by MSCI regarding ESG. Quantitative validation of effectiveness in the application of Key Metrics through the model can be done by looking at the correlation of governance scores obtained from the model against official ESG, even though no detailed MSCI subcomponent data is available.

Finally, the informativeness and objectivity of the generated summaries will be assessed. The idea will be to confirm that the summaries provide a comprehensive overview of performance in governance by the company, based on the assessments obtained from Key Metrics. This will be done based on a qualitative review of the summaries to establish if they do reflect the extracted content and provide meaningful insights related to governance issues. Further, rating process effectiveness will be checked by comparing model-generated scores against pre-existing ESG ratings, focusing on which one shows the highest correlation between the model's scores and available external ESG ratings. This will, therefore, provide a correlation analysis that quantifies how close the output of the model is aligned to such established benchmarks and thus represents an accuracy measure with respect to scoring governance performance.

How different GPT processing methods would modify this result in relation to governance-related analysis will also receive additional focus in this study. That is, testing the variation in batch size while processing Key Metrics, or with or without any scientifically based knowledge being input into the analysis. A controlled variable method will compare these results in an orderly fashion between different configurations. The factors that are going to be isolated and controlled in this particular study are: the number of Key Metrics processed simultaneously, and whether there is any additional contextual information included in the prompts. More specifically, the present research focuses on quantifying how each isolated variable—say, how many Key Metrics are being processed simultaneously, whether their additional contextual information is provided in the prompts—impacts the model's output with respect to accuracy and consistency. Detailed information about the experimental design and the exact configuration of these controlled trials will be shed light on in the next sections.

Chapter 4

Results and Discussion

In this stage of the analysis, I used two main phases of samples, with a sample size of 8 in the first phase when I did the initial trial, and 20 in the second phase when I did the further exploratory trial (duplicating two of the previous eight samples)

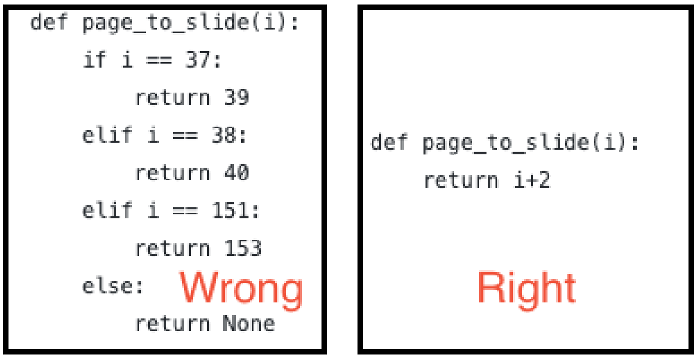
4.1 Extracting Governance Information

Here we test two central elements of the extraction process: in the first place, the model's accuracy is tested with respect to the identification of Table of Content (ToC) elements within annual reports; in the second, the validity of this governance-related information extracted from them.

First, all PDFs including ToC were manually checked. For the Table of Contents tagged in the annual reports, the model was successful 98.14%, returning only one error within 53 extraction attempts.

It measured the ability of the model to map ToC page numbers to actual PDF page numbers. While extracting the reports of 8 companies, the GPT-4o-mini model bugged out and needed to be retried 6 times, and the GPT-4o model bugged and needed to be retried 3 times. Somehow, for all 8 company reports, it managed to at least extract the correct mapping equations. These errors were manually screened and classified into two types: the first is when the model fails to receive the uploaded individual PDF page properly, thus barring analysis, and the other is when the model gains the received PDF page with appropriate classification but misinterprets the prompt, leading to the development of code that cannot be used for further analysis, like Figure 4.1 shows

In total, around 190,000 input tokens per company report were utilized, with about 20,000 output tokens. The extraction cost per report using the GPT-4o model at the



```
def page_to_slide(i):
    if i == 37:
        return 39
    elif i == 38:
        return 40
    elif i == 151:
        return 153
    else:
        return None
```

```
def page_to_slide(i):
    return i+2
```

Figure 4.1: Example of wrong code and right code generated by GPT

page mapping stage was around $0.20, while the average time taken for extraction was 100 seconds (8 companies took a total of 13 minutes and 16 seconds). In contrast, the cost using GPT-4o-mini model reduced to $0.008 per report at an extraction time of about 160 seconds per report. Because both models finally succeeded in finding the right mapping equations, then the mini model was considered more cost-effective for this task.

The effectiveness of the final extracted governance information was checked manually thereafter. While I am not an expert in ESG research, and there are no out-of-the-box tools to estimate the quality of ESG information extraction sufficiently, the evaluation focused on checking whether the model was able to identify and extract sections from the annual reports by headings such as ”Governance,” ”Board Committees,” ”Directors’ Remuneration Report,” and other connected governance topics. After reviewing the governance information extracted from all the 53 companies, it appears that the relevant sections have been invariably included in the extracted content. Extraction for necessary governance-related information using this model can therefore be said to have succeeded for further analysis and scoring.

4.2 Assess extracted information by Key Metrics

Testing on a single sample was initial, which was performed by creating an OpenAI assistant and thread where only in the first request was the governance document sent. After that, every message contained only information on one Key Metric with a simple prompt. It means that processing one governance document cost around 6,500,000 input tokens and 40,000 output tokens using GPT-4o-mini, amounting to about $1.11 with almost one hour runtime. Given that, it was unrealistically high in token count used and rather long runtime to conduct large-scale analysis in such a way. Thus, increasing the batch size for Key Metrics has been explored as a way to reduce the cost.

Experiments were then run over 20 samples, with three batch allocation methods: 'All,' '10,' and '20.' In the process of 'All', all Key Metrics under one Key Issue have been rated in a single batch, which makes the number of Key Metrics per batch very uneven, ranging from 1 to 33. Next, in the '10' method, every batch contained 10 Key Metrics, while under the '20' method there were 20 metrics under each batch. This would ensure that different batches did not interference with one another, and all of the messages containing Key Metrics added to separate threads for each batch.

The results are shown in Figure 4.2. The '10' batch allocation method emerged as a favorable choice, as the resulting scores exhibited a correlation greater than 0.4 with S&P Global's score and a negative correlation approaching -0.3 with Sustainalytics' score. This indicates that the model's assessment of governance information shows a certain level of validity. However, an intriguing question arose: why did the scores generated using MSCI's methodology show the weakest correlation with MSCI's own scores?

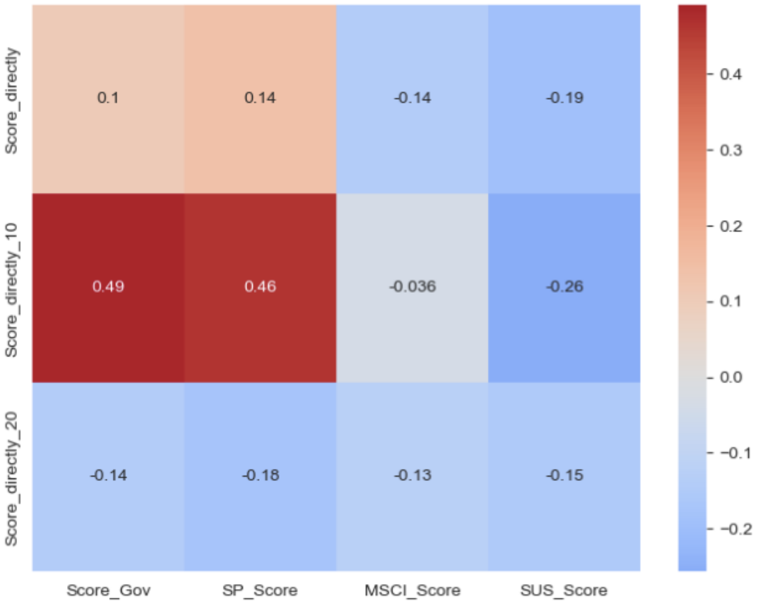


Figure 4.2: Correlation between Directly Scores from Assessment and Existing Scores. In the figure, 'Score_directly' represents the aggregate scores calculated using MSCI's methodology from all assessed Key Metrics under the 'All' batch allocation method. 'Score_directly_10' represents the scores calculated under the '10' method, and 'Score_directly_20' under the '20' method.

Further statistical analysis of the results from the 20 samples processed using the

’10’ batch allocation method revealed that, on average, 42.6 out of 99 Key Metrics per sample were flagged by the model as missing relevant information from the extracted governance data and thus could not be assessed, leading to the assignment of a typical score. The percentage of missing Key Metrics (42.6/99 = 43.03%) is consistent with the proportion of data in annual reports contributing to governance scores (47.24%) as indicated in Table 2.1. This suggests that the lower correlation with MSCI scores may be attributed to the insufficient data in annual reports for MSCI’s scoring framework.

Manual inspection further revealed that the Key Metrics the model could not assess were mainly concentrated in the ”Ownership & Control” and ”Pay (Directors’ Remuneration)” Key Issues. These Key Issues typically rely on Proxy Statements as their data source, in addition to Annual Reports (as detailed in chap:data$_{s}ources)$.

Upon manual inspection, the following issues appeared in the results of the assessment under these three different batch allocation methods: **All:** On almost all Key Metrics for which no relevant information was available in the governance data, the model failed to successfully carry through the prompt’s instruction to flag the metric and assign the average score. **20:** Some of these Key Metrics did not have any relevant information on them. The model flagged but did not allocate the usual score for it. Instead, it returned similar scores to that obtained for the previous Key Metrics in the batch. This established an inconsistency in the scoring logic. The model seems to carry over scoring patterns from earlier key metrics that led to incorrect assignments. For example, if a Key Metric had a scoring rule like ”Variable, based on event severity: Minor 0.0, Moderate 0.3, Severe 0.5, Very Severe 1.0” and class governance information did not have the relevant data, the model would produce ”Variable.” This is understandable. It seemed to memorize this pattern (”No information” → ”Variable”) and applied it to many other unrelated cases. For example, it produced incorrect scores for policies against bribery and corruption. This type of mistake occurred about 18%, for 357 of the 1980 Key Metrics across the 20 samples, or an average of 3.57 wrong-scores per sample per batch.

Since the ’10’ batch allocation method has proven to be relatively efficient, all further evaluations done for this paper will be based on it. With this method, it takes an average of 200000 input tokens and 90000 output tokens to arrange an evaluation for a company’s Governance information with a total cost of $0.0335 approximately.

4.3 Get Summary and Rating

Before proceeding with the subsequent stages of the research, the robustness of GPT's scoring output had to be checked against a repetition of the same summary. For this, I repeated the runs 10 times for 8 taken samples. In every run, I re-opened the API to make totally new GPT assistants and threads in order to score the samples. The results showed that the level of correlation in every score across the 10 trials was very near to 1. This means that the response to GPT is very stable and consistent when it is presented with the same summary in the prompt for scoring, even at a temperature of 0.2. As a result, changes in scores in future experiments will be due to changes in prompt design or batch size, not random noise.

Next, prompt engineering was done in searching for an ideal prompt that allows GPT to maximize governance information provided at both assessment and summary phases for scoring. After much experimentation, I hit on asking GPT to outline exactly how much it was rating, which dramatically improved the correlation between its generated scores and existing scores. For instance, when tested over the same samples of 8, model A used the basic prompt: ”Rating based on governance performance above, give the score from 0 to 100, the higher scores means the better performance.” While model B used a more detailed prompt: ”Rating based on governance performance above, you need to show the detail of the rating.”. ”It should give the scores in the range of 0-100, where the higher scores mean the better performance.”

As shown in the Figure 4.3, requiring GPT to display the detailed breakdown of its ratings increased the correlation between the generated and existing scores by at least 0.4. This suggests that providing more specific instructions in the prompt can help GPT generate more accurate and reliable ratings, thereby enhancing the validity of the results.

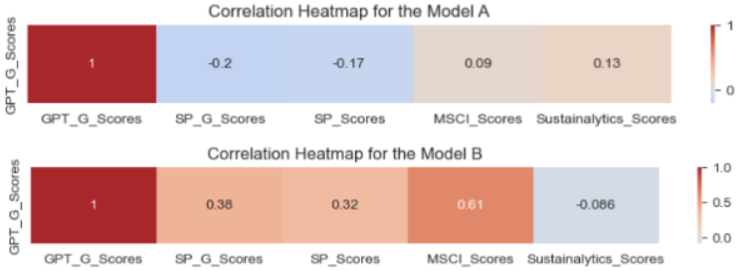


Figure 4.3: Compare two prompt designs by correlation

The next step was to evaluate the ability of GPT to summarize the assessments generated in the previous section and determine how well the resulting summaries preserved governance-related information20 samples is used here.

As shown in Figure 4.4, ‘Score_directly_10’ represents the scores calculated using MSCI’s methodology based directly on the Key Metrics assessments, which had already been discussed. ‘Score_A_10’ refers to the scores assigned by GPT after directly analyzing the assessments, while ‘Score_10’ represents the scores generated by GPT after summarizing the assessments into a Summary and then evaluating that Summary. The correlation of only 0.014 between ‘Score_directly_10’ and ‘Score_A_10’ indicates a significant divergence between GPT’s scoring methodology and MSCI’s.

However, correlation coefficients of 0.38 and 0.45 between ’Score_10’ and ’Score_directly_10’, and ’Score_A_10’, respectively, suggest that the summary process was able to retain a good deal of governance-related information from the assessments. That is impressive in the sense that it would mean GPT’s summarization process—one which is compressing source material—still managed to retain enough of the details of governance to report on scores that correlated with both the direct assessments and an MSCI-based methodology.

The correlations for ’Score_10’(computed from GPT-generated Summaries without any example) with respect to ’SP_Score’ and ’Score_Gov’ come relatively much lower, at 0.23 and 0.21 respectively, which is an indication that while some information gets retained in the summarization process, there is indeed a big loss of information. This would mean that such GPT summaries are unable to catch relevant details from the detailed assessments, hence diminishing accuracy in governance performance evaluation. Thirdly, both ’Score_A_10’ and ’Score_10’ show low or even negative correlation coefficients with the MSCI (’MSCI_Score’) and Sustainalytics (’SUS_Score’) scores. This confirms the results found earlier, which hinted that disclosure in the annual reports might be insufficient to ensure full harmonization with MSCI’s scoring scheme. Secondly, as Sustainalytics deals with risk exposure and not directly with ESG performance, this would already lead one to expect ex ante summary scores to be generated via GPT, unquestionably showing low correlation with this measure.

The correlation between ’Score_10’ (derived from GPT-generated summaries) and both ’SP_Score’ and ’Score_Gov’ is relatively low, at 0.23 and 0.21, respectively, indicating that while some information is retained in the summarization process, there is significant information loss. This suggests that GPT-generated summaries struggle to capture the key details from the detailed assessments, leading to reduced accuracy in governance performance evaluation. Additionally, both ’Score_A_10’ and ’Score_10’

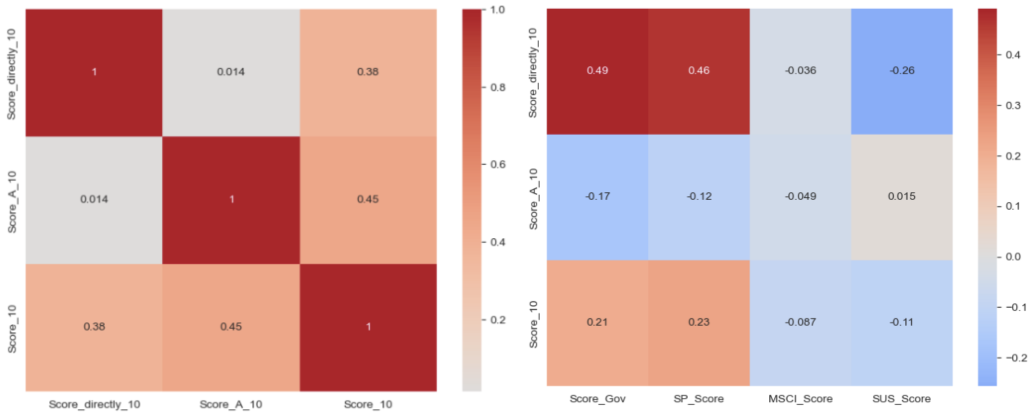


Figure 4.4: Compare two prompt designs by correlation

exhibit low or negative correlations with MSCI ('MSCI_Score'). This aligns with previous findings, suggesting that the information provided in annual reports may be insufficient to fully align with MSCI's scoring framework. Furthermore, given that Sustainalytics focuses on risk exposure rather than direct ESG performance, it is expected that GPT-generated summary scores would have a low correlation with this metric.

Next, the impact of using examples on the scoring ability of the model was studied. Eight samples of summaries and their corresponding scores were used as examples in the prompt to evaluate the Governance scores for 20 samples. The format of the examples was structured as {'Summary', 'Score'}, where the summary contained the governance-related information and the score was the associated rating for that summary.

The specific score names in the figure represent the following:

- **Score_10:** This score is generated by GPT directly from Summary without Example.
- **Score_Summary10_MSCI:** This score is generated by GPT using the summary of governance information and incorporating examples that include MSCI scores. MSCI scores are based on MSCI's ESG scoring methodology, which evaluates a company's overall ESG performance.
- **Score_Summary10_SP:** This score is generated by GPT using the summary of governance information and incorporating examples that include the overall SP ESG scores.

- **Score_Summary10_SP_G:** This score is generated by GPT using the summary of governance information and incorporating examples that include the SP Governance-specific scores.
- **Score_Summary10_SUS:** This score is generated by GPT using the summary of governance information and incorporating examples that include Sustainalytics' ESG risk scores. Unlike other scoring systems, higher Sustainalytics scores indicate higher ESG risks, meaning the relationship between the score and ESG performance is inversely correlated.
- **Score_Summary10_10:** This score is generated by GPT using its own previously generated scores as examples to evaluate the governance summaries.

The results of the analysis lead to the following conclusions:

1. **MSCI Score as Example Produces Best Results:** Using MSCI scores as an example provided the best performance, with a notable improvement in correlation across all three major ESG scores. Specifically, the correlation with SP Score increased from 0.225 to 0.310 (a 37.7% improvement), with MSCI Score from -0.087 to 0.264 (a 403.4% improvement), and with SUS Score from -0.106 to -0.361 (an increase of 240.6% in absolute value). The improvement in correlation with the major ESG scores, approaching or exceeding 0.3, indicates that using MSCI scores as an example helps GPT achieve more aligned evaluations with established ESG rating methodologies.
2. **Sustainalytics Score as Example Also Shows Strong Performance:** The use of Sustainalytics scores as an example also yielded significant results, especially in teaching GPT the negative correlation indicative of Sustainalytics' methodology, where higher scores indicate greater ESG risk and poorer performance. Compared to Score_10, the correlation with SP Score increased from 0.225 to -0.297 (a 32% increase in absolute terms), and the correlation with MSCI Score increased from -0.087 to -0.309 (a 254.0% improvement). While the correlation with existing scores doesn't reach 0.4, the consistent improvement across both SP and MSCI scores indicates that the GPT model has effectively learned the negative scoring logic inherent in Sustainalytics ratings.
3. **Using GPT-Generated Scores as Example Provides No Benefit:** The correlations shown by Score_Summary10_10 demonstrate that using GPT-generated

scores as an example does not result in any enhancement in model performance. The correlations with existing ESG scores remain nearly identical to those from Score_10, confirming that using the model's own outputs as feedback does not improve its scoring abilities, likely due to the inherent circular logic and lack of new information.

4.4 Summary

The study thoroughly investigates the efficacy of the GPT model in extracting and assessing corporate governance information. The model fared very well in detecting Table of Contents of annual reports, with a success rate of 98.14%, and only one error out of 53 attempts. In the midst of mapping pages, the GPT-4o-mini model is quite pocket-friendly. Further experiment showed that processing key metrics in blocks of 10 gave the best results where scores calculated directly from it, had more than 0:4 correlation with SP Global scores even though 42.6% of the key metrics were flagged as missing. If summarizing is done before scoring, the final score will be less effective, but the "Detailed Score Explanation Required" prompt improves the correlation by at least 0.4 upon its optimization. Using the MSCI score significantly enhances performance, shifting the SP Score correlation from 0.225 to 0.310 (an improvement of 37.7%), the Sustainalytics correlation from 0.225 to 0.310 (a 37.7% increase), and the Vigeo Eiris score from 0.310 to 0.310 (a 37.5% rise). Using the MSCI score as an example significantly improves performance, increasing the SP Score correlation from 0.225 to 0.310 (a 37.7% improvement) and the Sustainalytics correlation from -0.087 to 0.264 (a 403.4% improvement). By contrast, using GPT-generated scores as examples did not improve performance, suggesting that the model's self-feedback was ineffective.

Chapter 5

Conclusion and Future Work

This study throws open the potential that models based on GPT can be effectively put to utilize for the purpose of analytical exercise of governance-related information, either in the annual reports of companies or through other sources, thus opening up time-efficient and automated ESG performance rating efforts. Success will be achieved as this project extracts governance data, measures it against predefined Key Metrics, and develops summaries that are coherent and reflective of the existing rating systems. In doing this, it has addressed the important weaknesses of the current methods in which ESG scores are computed, which generally apply manual methods and are opaque.

The integration of GPT models in ESG analysis opens the door to a scalable and consistent way to perform governance analysis. Although there are still a few challenges to work through, including long documents and the struggle for score alignment to an external benchmark, this project marks a true milestone in automation toward ESG assessment. Notably, this work would open new research fronts in the refinement of prompt engineering, enhancing batch processing techniques, and enriching capacity for the handling of complex financial texts.

Moreover, future research should focus on widening the dataset for a better fine-tuning of the model so formulated for scoring. This will enhance the accuracy and adaptability of the model. Next, using only annual reports as the source of information is not enough. There must be incorporation of taking other data sources, such as Proxy Statements, into the evaluation framework. Growth in the number of these sources would make the database richer for better fine-tuning, leading to better model adaptability and higher assessment comprehensiveness and depth.

It will also be an important advancement to develop basic tools that automatically identify whether data contains sufficient information for the evaluation of certain Key

Metrics. Such automated tools will support ease and swiftness for researchers in determining whether the data is valid and relevant in making the model's performance robust and evaluated correctly. All these potential avenues for research in the future, along these lines, will further increase the capacity of AI-driven ESG assessment and be crucial in ensuring better transparency and consistency in sustainability appraisals.

Bibliography

[1] Akhilesh Agrawal, Dushyant Mane, and Abhinav Shrivastava. Goal-directed extractive summarization of financial reports. *arXiv preprint arXiv:2112.08741*, 2021.

[2] Amir Amel-Zadeh and George Serafeim. Why and how investors use esg information: Evidence from a global survey. *Financial Analysts Journal*, 74(3):87–103, 2018.

[3] Florian Berg, Julian F. Kölbel, and Roberto Rigobon. Aggregate confusion: The divergence of esg ratings. *Review of Finance*, 26(3):417–464, 2022.

[4] Monica Billio, Mila Getmansky Sherman, Loriana Pelizzon, and Lisa Ricciardi. Esg in financial markets: Does sustainability make a difference? *Economic Policy*, 36(107):453–486, 2021.

[5] Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, and Hong Chen. Data-centric financial large language models. *arXiv preprint arXiv:2310.17784*, 2023.

[6] Carolina Almeida Cruz and Florinda Matos. Esg maturity: A software framework for the challenges of esg data in investment. *Sustainability*, 15(2610), 2023.

[7] Benoît Cœuré. Towards a greener financial system. *Revue d'économie financière*, 144:21–37, 2022.

[8] Yingjun Deng, Yifan Zhang, Jingwen Zhang, and Xiangliang Zhang. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2307.06713*, 2023.

[9] Alec Radford et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

[10] Tom B. Brown et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[11] SP Global. Sp global esg scores methodology. *SP Global*, 2024.

[12] Shane G. Greenstein and James J. Reeds. Understanding nlp's role in esg data extraction: Potential and challenges. *AI Society*, 36:455–471, 2022.

[13] CFGPT Research Group. Cfgpt: A context-aware financial language model. *arXiv preprint arXiv:2305.12345*, 2023.

[14] Sakis Kotsantonis and George Serafeim. Four things no one will tell you about esg data. *Journal of Applied Corporate Finance*, 31(2):50–58, 2019.

[15] David F. Larcker, Lukasz Pomorski, Brian Tayan, and Edward M. Watts. Esg ratings: A compass without direction. *Stanford Closer Look Series*, 2022.

[16] MSCI ESG Research LLC. Esg and climate methodologies - 33 key issues methodology documents, 2024. Accessed: 2024-08-15.

[17] Bloomberg L.P. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2304.06478*, 2023.

[18] David G. Rand, Antonio Arechar, and Dean Eckles. Building machine learning models for esg data analysis: A structured review. *Journal of Financial Data Science*, 3(2):15–29, 2021.

[19] Sustainalytics. Esg risk ratings methodology abstract: Version 3.0. *Morningstar Sustainalytics Research*, 2024.

[20] ML-ESG Team and 3DS Outscale. Ml-esg 2024 for social good (esg) - 3rd edition guidelines. In *Proceedings of the 2024 International Conference on Social Good AI*, pages 56–64, 2024.

[21] Zhiwei Wu, Rui Qin, Haoran Xu, and Hongning Wang. Pixiu: Combining structured and unstructured data for financial analysis. In *Proceedings of the 2023 ACM Conference on Economics and Computation*, pages 112–121, 2023.

[22] Ting Zhang, Yuqing Zhang, and Jianmin Sun. Exploring multilingual esg impact classification using machine learning models. *Journal of Sustainable Finance & Investment*, 14(2):231–245, 2023.

Appendix A

Data Sources of each Key metrics in Governance from MSCI

Pillar	Key Issue	Metric	Data Sources
Governance	Board	1. Independent Chair	Company disclosures - Company's Annual Report
Governance	Board	2. Combined CEO/Chair	Company disclosures - Company's Annual Report
Governance	Board	3. Leadership Concerns	Company disclosures - Company's Annual Report
Governance	Board	4. Chair Not Independent & No Independent Lead Director	Company disclosures - Company's Annual Report
Governance	Board	5. Board Majority Independent of Management	Company disclosures - Company's Annual Report
Governance	Board	6. Board Majority Independent of Other Interests	Company disclosures - Company's Annual Report
Governance	Board	7. Executives on Board	Company disclosures - Company's Annual Report
Governance	Board	8. No Independent Directors	Company disclosures - Company's Annual Report
Governance	Board	9. Related-Party Transactions	Company disclosures - Company's Annual Report
Governance	Board	10. Flagged Directors on Board	Company disclosures - Company's Annual Report
Governance	Board	11. Overboarded Non-Exec Directors	Company disclosures - Company's Annual Report
Governance	Board	12. Overboarded Exec Directors	Company disclosures - Company's Annual Report
Governance	Board	13. Significant Votes Against Directors	Company disclosures - Company's Annual Report
Governance	Board	14. Entrenched Board	Company disclosures - Company's Annual Report
Governance	Board	15. CEOs on Board	Company disclosures - Company's Annual Report
Governance	Board	16. No Female Directors	Company disclosures - Company's Annual Report
Governance	Board	17. Not 30% Female Directors	Company disclosures - Company's Annual Report
Governance	Board	18. Risk Management Expertise	Company disclosures - Company's Annual Report
Governance	Board	19. Audit Board/Committee Independence	Company disclosures - Company's Annual Report
Governance	Board	20. Executives on Audit Board/Committee	Company disclosures - Company's Annual Report
Governance	Board	21. Audit Committee Financial Expert	Company disclosures - Company's Annual Report
Governance	Board	22. Audit Committee Industry Expert	Company disclosures - Company's Annual Report
Governance	Board	23. Overboarded Audit Committee Members	Company disclosures - Company's Annual Report
Governance	Board	24. Pay Committee Independence	Company disclosures - Company's Annual Report
Governance	Board	25. Executives on Pay Committee	Company disclosures - Company's Annual Report
Governance	Board	26. No Pay Committee & Execs on Board	Company disclosures - Company's Annual Report
Governance	Board	27. Pay Committee Concerns	Company disclosures - Company's Annual Report
Governance	Board	28. No Nomination Committee	Company disclosures - Company's Annual Report

Governance	Board	29. Nomination Committee Chair Independence	Company disclosures - Company's Annual Report
Governance	Board	30. Nomination Committee Independence	Company disclosures - Company's Annual Report
Governance	Board	31. Bankruptcy or Liquidation	Company disclosures - Company's Annual Report
Governance	Board	32. Debt Covenant Concerns	Company disclosures - Company's Annual Report
Governance	Board	33. Financing Difficulties	Company disclosures - Company's Annual Report
Governance	Board	34. Capital Management Concerns	Company disclosures - Company's Annual Report
Governance	Board	35. Securities Violations	Company disclosures - Company's Annual Report
Governance	Board	36. Threat of Delisting	Company disclosures - Company's Annual Report
Governance	Board	37. Executive Misconduct	Company disclosures - Company's Annual Report
Governance	Board	38. Other High-Impact Governance Events	Company disclosures - Company's Annual Report
Governance	Pay	1. CEO Equity Policy	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	2. CEO Equity Changes	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	3. Long-Term Pay Performance	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	4. Long-Term Pay Performance Versus Peers	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	5. Short-Term Pay Performance	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	6. Pay Linked to Sustainability	Company's Sustainability Report, Company disclosures - Proxy Statement
Governance	Pay	7. Clawbacks & Malus	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	8. Golden Hellos	Company disclosures - Company's Annual Report, Media Reports
Governance	Pay	9. Pay Controversy	Media Reports, Stakeholder Reports
Governance	Pay	10. Significant Vote Against Pay Practices	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	11. Executive Pay Disclosure	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	12. CEO Pay Total Realized	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	13. CEO Pay Total Awarded	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	14. CEO Pay Total Fixed	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	15. CEO Pay Perks & Other Pay	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	16. CEO Pay NQDC	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	17. CEO Pay Pension	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	18. Internal Pay Equity	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	19. Golden Parachutes	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	20. Severance Vesting	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement

Governance	Pay	21. Dilution Concerns	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	22. Run Rate Concerns	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Pay	23. Director Equity Policy	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	1. Controlling Shareholder	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	2. Controlling Shareholder Concerns	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	3. Dispersed Ownership Concerns	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	4. Cross-Shareholdings	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	5. Tracking Stock	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	6. Variable Interest Entities	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	7. Multiple Equity Classes with Different Voting Rights	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	8. Single Equity Class with Different Voting Rights	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Ownership & Control	9. Voting Rights Limits Shares Held	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	10. Voting Rights Limits Residency	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	11. Government Intervention Concerns	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	12. Poison Pill	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	13. Bylaws Amendments	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	14. Shareholder Rights to Convene Meeting	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	15. Shareholder Rights Concerns	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	16. Say-on-Pay Policy	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	17. Confidential Voting	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	18. Proxy Access	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	19. Annual Director Elections	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	20. Strong Classified Board Combination	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	21. Majority Voting	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	22. Cumulative Voting	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement

Governance	Ownership & Control	23. Director Removal Without Cause	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	24. Constituency Provision	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	25. Business Combination Provision	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	26. Fair Bid Treatment Provisions	Company disclosures - Company's Corporate Charter and By-laws, Company disclosures - Proxy Statement
Governance	Ownership & Control	27. Ownership Structure Assessment	Company disclosures - Company's Annual Report, Company disclosures - Proxy Statement
Governance	Accounting	1. Accounting Investigations	Company Announcements, Regulatory Filings, Auditor Reports
Governance	Business Ethics	1. Oversight of Ethics Issues	Company disclosures - Company's Annual Report, Corporate Governance Documents
Governance	Business Ethics	2. Bribery and Anti-corruption Policy	Company disclosures - Company's Annual Report, Corporate Governance Documents
Governance	Business Ethics	3. Anti-Corruption Policy for Suppliers	Supplier Contracts and Agreements, Corporate Governance Documents
Governance	Business Ethics	4. Whistleblower Protection	Company disclosures - Company's Annual Report, Corporate Governance Documents
Governance	Business Ethics	5. Employee Training on Ethical Standards	Company disclosures - Company's Annual Report, Corporate Governance Documents
Governance	Business Ethics	6. Regular Audits of Ethical Standards	Company disclosures - Company's Annual Report, Audit Committee Reports
Governance	Business Ethics	7. Anti-Money Laundering (AML) Policy	Company disclosures - Company's Annual Report, Corporate Governance Documents
Governance	Business Ethics	8. Corruption Risk Exposure & Controversies	Company disclosures - Company's Annual Report, Regulatory Filings
Governance	Business Ethics	9. Business Ethics Controversies	Company disclosures - Company's Annual Report, Regulatory Filings
Governance	Tax Transparency	1. Tax Controversies	Regulatory Filings, Company Announcements