AI-Assisted Compliance: Evaluating Corporate Governance Reporting through Large Language Models

Thesis – MSc Economics

Lucas Breuker

SNR: 2150588

ANR: 849276

University Supervisor: Prof. Dr. Jan Boone

Internship Supervisor: Dr. Michiel Bijlsma & Joost Witteman

Tilburg University

17 June 2025

Abstract

This study explores the use of Large Language Models (LLMs) to automate compliance checks of the Dutch Corporate Governance Code (DCGC) by analyzing annual reports of Dutch-listed companies. A custom workflow was made using the Retrieval Augmented Generation (RAG) approach combined with embedding indexing and the use of OpenAI's GPT-4o model. After prompt refining the compliance of a sample of 53 companies with 19 provisions was assessed. Two hypotheses were tested—external alignment (H1) and internal robustness (H2). The results from the workflow showed that LLMs can provide a reliable aggregated compliance percentage (Average aggregate compliance = 80%). However, when looking at individual provisions it shows it can struggle with more complex provisions due to limitations in context retrieval. Statistical tests, including Monte Carlo simulations and bootstrapping, confirmed internal reliability, but weak correlation with historical human-based assessment (Pearson $r \approx 0.15$), highlighting the continued need for human oversight. Overall, the research demonstrates that LLMs can drastically improve efficiency and consistency of compliance monitoring, but results should be carefully checked before usage for complete accuracy.

Table of Contents

1. Introduction

The rapid advancement of AI, specifically in the domain of Natural Language Processing, has significantly transformed the way in which organizations are able to analyze large textual data. In particular, Large Language Models (LLMs) such as OpenAI's GPT-4o, have created new possibilities for automating complex cognitive tasks that were thought to be relying heavily on human expertise before. One area potentially benefitting from these developments is corporate governance regulations, where companies must adhere to specific reporting standards outlined by regulatory bodies. The use of LLMs for automating the analysis of large and complex textual datasets has shown promising results across various domains, indicating their potential effectiveness in regulatory compliance assessments (Grohs et al., 2023). The Dutch Corporate Governance Code (DCGC), established to ensure transparency and accountability within listed companies in the Netherlands, exemplifies such a regulatory standard, requiring companies to adhere to stated provisions within their annual reports (Monitoring Commissie Corporate Governance Code, 2025).

Assessing compliance with corporate governance codes through annual reports poses significant challenges due to the substantial length, complexity, and variability in language used across reports. Furthermore, the heterogeneity of these companies, and their corresponding reports, can lead humans to interpret the same governance rules differently depending on the specific context. Traditional manual methods of compliance assessment are labor-intensive, prone to human error, and limited by cognitive and temporal constraints. As such, manual analyses struggle to maintain consistency and accuracy, making it difficult to ensure reliable compliance evaluation across numerous companies.

To address these limitations, this research uses advanced NLP techniques provided by different models from OpenAI, aiming to automate the assessment of compliance with the DCGC. In earlier research the capability of LLMs to effectively exclude irrelevant information and summarize large textual data has already been demonstrated in analyzing corporate documents (Kim et al., 2024). A custom-developed Python workflow utilizing the extensive OpenAI API will be developed to systematically extract, analyze and assess information from extensive annual reports relevant to compliance with the DCGC. Given the variability LLMs possess and the ethical concerns regarding biases, the research includes rigorous statistical testing, specifically, Monte Carlo simulations and bootstrapping, to evaluate the stability and reliability of automated assessments. Furthermore, the workflow will be evaluated against the benchmark of human-made assessments from earlier years to analyze the accuracy of the results.

This study contributes to the emerging field of automated textual analysis by addressing the research question:

> *To what extent can LLMs reliably assess compliance with the Dutch Corporate Governance Code from corporate annual reports?*

Previous literature has shown LLMs have inherent variability, even when setting its temperature to zero, and advocated for an increase in benchmark studies trying to quantify uncertainty (Blackwell et al., 2024). Combined with the increasing capabilities of state-of-the-art LLMs, this caused me to formulate the following two hypotheses:

H1: The overall compliance percentages produced by the LLM workflow will have moderate to high correlation with historical human expert assessments.

H2: The variability observed in the LLM workflow's compliance assessments will be limited and will not compromise the validity or practical utility of the workflow's results.

The outcomes of this study aim to evaluate the reliability, accuracy and effectiveness of LLMs in assessing corporate compliance with the DCGC based on annual report disclosures. By addressing the methodological performance of LLMs, this study contributes to a deeper understanding of AI-assisted textual analysis specifically in the context of corporate governance.

2. Literature Review

In this section I will explore the background literature connected to my research. Starting with literature about the effectiveness and application of corporate governance regulations. Then, I will explain more about readability and classification issues which cause the need for the deployment of LLMs. Third, the emergence and workings of LLMs and generative AI. Finally, some ethical concerns and potential issues regarding LLMs will be discussed.

2.1 Corporate governance

Corporate governance is a broad field containing regulations, practices and theories aimed at enhancing transparency, accountability and ethical management. Its impact and significance extend beyond organizational management, reaching into legal frameworks, economic efficiency, and empirical research. The complex dynamics and significant consequences associated with corporate governance practices make it an important subject of study to understand its broader implications within economy and society. This section examines the effectiveness of corporate governance and

application examples of governance regulations as well as the importance of corporate governance across the legal field and economics as well as for empirical research.

2.1.1 Effectiveness and application of Governance Regulations

Corporate governance regulations emerged in response to significant corporate scandals and failures, aiming to establish clear rules and practices to enhance transparency, stakeholder engagement and trust and accountability within organizations. High-profile collapses such as Enron, WorldCom, and Parmalat in the early 2000s highlighted the need for extensive regulatory frameworks that could mitigate risks associated with unethical management practices and financial misconduct (Clarke, 2007; Coffee, 2006).

The main purpose of corporate governance regulations is to align interests of management with interest of stakeholders, ensuring effective oversight and responsible decision making focused on the benefit of the company and not that of the management personally. This includes clearly defining roles and responsibilities, establishing rigorous financial reporting standards required to be audited by an independent auditor, suggesting transparent disclosure practices, and promoting ethical corporate behavior (Aguilera & Cuervo-Cazurra, 2009). The implementation of these regulations is not only important for preventing misconduct, but it enhances corporate reputation, reduces agency costs and improves the efficiency of organizations.

The Dutch Corporate Governance Code (DCGC) is an example of these efforts. It provides guidelines focused on transparency, management accountability, remuneration policies, risk management, corporate culture and shareholder engagement (Monitoring Commissie Corporate Governance Code, 2025). These regulations present substantial challenges regarding application, interpretation and consistent enforcement across diverse contexts. Variations in company size, industry practices, and managerial approaches cause issues with consistent assessment of regulatory compliance, often requiring significant human judgement to interpret the quality of compliance with provisions in varying contexts.

To improve the effectiveness of these regulations, ongoing research and innovation with respect to methods of compliance assessment, such as automation using artificial intelligence, should be undertaken. Technology like LLMs show considerable promise concerning the accurate interpretation of large bodies of text, such as annual reports, and being able to filter out irrelevant information (Chiang & Lee, 2023; Kim et al., 2024). This could potentially lead to improved consistency, accuracy and efficiency in regulation assessment.

2.1.2 Interdisciplinary implications of corporate governance

Corporate governance is interesting for multiple academic disciplines due the its variety of implications. In the legal field for instance, corporate governance has shown to have several important connections. Persons (2005) found that companies with an audit committee solely comprised of independent directors have a significantly lower chance of committing fraud. Furthermore, another study by Velte (2023) found indications that board expertise and gender diversity in top management of a company decreases financial misconduct. This clearly shows the importance of good corporate governance rulemaking for regulatory institutions as it can have a significant impact on overall compliance with the law.

Research on corporate governance found several implications in the field of economics as well considering financial performance. The results from several studies indicate that board diversity is positively associated with financial performance, suggesting that good corporate governance practices contribute to company performance (Aggarwal, 2013; Erhard et al., 2003). Moreover, Kiel & Nicholson (2003) argue that board size has a positive effect on firm value after controlling for firm size. These results highlight further evidence of the importance of corporate governance and its positive effects.

In addition to its practical implications in legal and economic domains, corporate governance research, such as the automated compliance assessments explored in this study, provides valuable groundwork for future empirical academic research. By utilizing technologies like LLMs, this research generates reproducible, systematic and scalable insights in a fraction of the time it would take a human expert. This enables subsequent studies to explore relationships with corporate governance with improved accuracy as has already been done in other areas (Webersinke et al., 2021). Thus, my study not only adds to previous knowledge related to corporate governance, but also offers insights into a new way of generating robust empirical data for future research.

2.2 Readability and complexity issues

There has been a significant amount of previous research into readability and complexity of financial disclosures. A research by Smith & Taffler (1992) already mentioned low readability and understandability of accounting disclosures, with only a small target audience being able to grasp all the information. Furthermore, Guay et al. (2016) showed that new accounting rules are make disclosures increasingly complex, thus decreasing readability. They show that managers try to mitigate this by use of voluntary disclosures where the focus can be on one important aspect that the stakeholders need to be informed of. This is supported by Dyer et al. (2017) who analyzed US 10-K disclosures. They showed an upward trend in the length of disclosures, redundancy and boilerplate and

decrease in specificity and readability together with a considerable reduction in specificity and clarity. These characteristics have a significant impact on stakeholders' ability to understand and act on key corporate information effectively, influencing quality of decision-making.

Regulators have realized these problems and launched several initiatives to encourage more concise, easy-to-read and informative disclosures. Efforts by the Security and Exchange Commission illustrate ongoing efforts to improve readability and clarity of information (SEC, 2013). Despite these efforts, Cohen et al. (2020) showed that even today investors are not able to properly digest financial disclosures.

2.3 Large Language Models (LLM) and their usage

Large Language Models (LLMs) represent immense innovation in the fields of Natural Language Processing (NLP) and Artificial Intelligence (AI). These models are recognized for their vast amounts of parameters and training on all sorts of textual data. In recent years they have rapidly grown to be of major importance in everyday life, both in professional and academic spheres.

2.3.1 Evolution and technological advances

LLMs began as simple n-gram models, which took the previous n number of words to predict the next word (Brown et al., 1992). Later, when the internet became more used, researchers began basing their statistical language models on bigger web-based datasets. However, these models were still very much rule-based and encountered issues trying to capture long-range dependencies. With the development of Recurrent Neural Networks (RNN) combined with Long Short Term Memory (LSTM) this became a lot less of an issue. RNN's architecture allowed it to take the output of the last state, also called "hidden state", as an input in the next state. This enabled RNNs to capture patterns as well as interdependencies between specific words much better than earlier models. However, the real breakthrough happened following research by Vaswani et al. (2017) who came up with a new model called the Transformer. This model was significantly more computationally efficient as well as providing much better results than any earlier model. This was also the model that was then used for the release of BERT, the state-of-the-art LLM of the time in 2018. Later, OpenAI came with their own LLM, GPT-1, and its later updates up to GPT-4o, which is the most used LLM as of today.

2.3.2 Usage of LLMs

The advanced capabilities of modes like BERT and the GPT series prove them reliable and accurate enough to start being used across many different fields. The financial sector, for example, has already started using these models for various tasks (Li et al., 2023). LLM's ability to summarize, analyze

and extract key insights from large financial data helps financial institutions make informed investing choices (Zhao et al., 2024). They can, for example, be used for tasks such as risk and volatility prediction of financial markets, increasing effectiveness in forecasting both volatility and variance(Cao et al. 2024). Furthermore, they can be used to detect fraudulent transactions and even detect complex fraud schemes that went undetected by earlier conventional methods (Korkanti, 2024). Financial service providers can also make use of LLMs by creating advanced chatbots that can handle a vast range of different kinds of customer questions in real-time. They can reduce the workload for human workers, while a the same time being able to answer more accurately and in a more timely manner (Hivarhizov, 2024).

Beyond finance, LLMs are deployed in various other fields. In healthcare the technology is used to improve the precision of diagnostics, predict patient outcomes and advance personalized healthcare (Rithika, 2024). In education, LLMs are being integrated into personalized learning environments as well as automatic grading and tutoring services. Furthermore, in the legal domain LLMs are being used to automate various legal tasks like judgment prediction and both document analysis and writing (Sun, 2023). However, with all these applications ethical concerns arise about potential biases and privacy issues.

2.4 Ethical concerns

Although LLMs provide powerful new opportunities, their increasing usage in every-day life asks for a thorough examination of their possible ethical flaws and issues. Key concerns include potential biases, privacy issues, environmental costs and explainability, asking for careful consideration during development and deployment.

2.4.1 Bias and explainability

LLMs are trained on datasets that often contain injustices or biases shown in human behavior. This can cause the same issues to be continued into their output. For example, if training data contains bias regarding race or gender, the LLM might generate responses that prolong these issues, potentially creating further inequality in society (Bender et al., 2021; Gallegos et al., 2024). This can also be seen in other examples like research by Blodgett & O'Connor (2017), which showed that language by minorities and females was sometimes analyzed poorly compared to that of white males. Furthermore, LLMs are used in high-stake fields like healthcare, where a wrong diagnosis based on a bias inherent in the model, could lead to wrong treatment and significant real-world consequences.

Another risk that should be taken into account when using LLMs is explainability. The explainability of a model is defined by the degree to which users and developers are able to understand

the underlying mechanisms being used by the model to come to its conclusions (Du et al., 2008). While the models show impressive accuracy and efficiency, they lack in transparency about the underlying decision-making process that is being utilized. The nature of LLMs being almost like a “black-box”, meaning it is very difficult unravel the how and why behind a decision, complicates the task of diagnosing errors and validating results. The absence of model transparency can lead to generation of hurtful responses or even hallucinations (Weidinger et al., 2021). If explainability is increased, this can aid developers when identifying unintended biases, risks and other areas where performance can be improved. Moreover, it would be beneficial for general users, since they would be able to better understand potential flaws and limitations in the responses they are getting. This would help them in recognizing these flaws and ultimately utilizing the models in a better way.

2.4.2 Privacy issues and environmental costs

Now that models like the GPT series and BERT have become reliable and accurate enough to be used by a widespread userbase, environmental issues start to arise. The large scale usage of LLMs requires an immense amount of computational power, both for training as well as reasoning done during every-day usage. Training these state-of-the-art LLMs demands considerable energy consumption, thus leading to high carbon emissions. Another issue is the fresh water that is needed to cool the datacenter where the computations are performed (Mytton, 2021). Although exact data on energy cost per prompt is still lacking, at this moment in time most companies spend more resources operating their LLMs than training them. Both NVIDIA and Amazon Web Services claimed that about 90 percent of their total workload is spent on reasoning and only a small percentage on training (Patterson et al., 2021). Thus, it seems that the more accurate and usable these LLMs become, the more the userbase will grow and with it the total effect on the environment.

Another issue that is becoming more apparent as LLMs grown in size is privacy. LLMs are being trained on an increasingly vast body of data that runs the risk of containing sensitive private information that is then memorized and potentially exposed by the model. In a research by (Carlini et al., 2020) it became apparent that training data extraction attacks can successfully be performed on LLMs. These attacks are able to extract individual training examples from the training data, including personal information like email, name, phone number and address. Most worryingly, they showed that these attacks sowed more success when being performed on larger models compared to smaller models. Thus, as LLMs continue to expand it is important to keep assessing the completeness of existing protocols that deal with the privacy of these models and update them wherever seems necessary (Das et al., 2024).

3. Methodology

This chapter outlines the design for this research and the methods used to assess the compliance with the Dutch Corporate Governance Code (DCGC), as determined by an LLM workflow using annual reports. The approach will utilize state-of-the-art LLMs and a RAG technique to extract relevant information from annual performance PDFs and apply statistical resampling techniques to assess variance and performance robustness. This approach is designed to test the two main hypotheses by combining compliance benchmarking, Monte Carlo simulations, and bootstrapping techniques.

In this section, first I will explain what choices and deliberations lead to the choosing of my sample. Then, I will explain the data pre-processing steps, which entails extracting the plain text from all annual reports in my sample and embedding it into indexes ready for retrieval. Furthermore, I will elaborate on the design of the workflow and the pipeline each iteration goes through. Lastly, I will explain which statistical tests will be performed to test the validity and accuracy of my results.

3.1 Data collection and preparation

3.1.1 Company sample selection

The sample that will be used in this research consists of 53 companies falling under the DCGC, meaning they are a listed company with a Dutch legal entity. The sample consists of different types of companies listed in the Netherlands. Both smaller Dutch-listed companies as well as bigger companies are analyzed. This ensures the results give an accurate representation of the performance of the LLM workflow for various types of companies and annual reports.

3.1.2 Provision sample selection

The DCGC is a comprehensive list of standards which contains both ‘reporting provisions’ and ‘behavioral provisions’. Behavioral provisions are assumed to be complied with unless otherwise stated in the annual report of the company. It is not possible to assess these provisions using the annual reports and thus they are not included in the scope of this research. The reporting provisions are mostly to be either reported in the annual report or on the website. The set of provisions that need to be reported in the annual report are the provisions focused on in this paper and used for the final correlation and reliability analysis. Therefore, I end up with analysis of the following 19 provisions: 1.1.3, 1.1.4, 1.3.6, 1.4.2, 1.4.3, 2.1.2, 2.1.3, 2.1.6, 2.1.10, 2.2.8, 2.3.5, 2.3.11, 2.4.4, 2.5.4, 2.7.4, 2.7.5, 3.4.1, 4.2.6 and 5.1.5.

3.1.3 PDF preprocessing

Prior to the assessment of the individual provision in the DCGC, the PDF's of the annual reports need to be converted to plain text. This is a crucial first step as it is essential for the Natural Language Processing (NLP) tasks that will have to be performed later on in the workflow. For this purpose I will be using a package called PyMuPDF, specifically the version PyMuPDF4llm, which is designed specifically for converting PDFs to markdown text suitable for LLMs[1]. Markdown is a lightweight markup language that uses plain-text syntax to denote formatting like tables or headings. Since I will be mostly using OpenAI's ChatGPT GPT-4o model, which can understand and sometimes even benefits from markdown text format this should be beneficial for the accuracy of the results. It can detect standard text and tables as well as identifying bold, italic and mono-spaced text and formatting them accordingly. This high-quality text extraction preserves the content, and even formatting like tables or bold text is translated into correct markdown format. This significantly improves subsequent processes such as embedding creation and indexing, as more high quality information will be available. This approach avoids relying on an LLM's image/vision capabilities to read PDF's and corresponds with best practices in NLP workflow design.

3.2 The design of the main workflow

The workflow uses the following pipeline to generate its end results. First, it extracts all text and creates indexes from the PDF for several sections: the remuneration report, the supervisory board report, the management board report and the full pdf excluding the financial statements. Next, the LLM is given a prompt with the question to look for the section containing any information about codes the company does not comply with as well as a prompt asking the LLM to then analyze the reasons given by the company for not complying. These reasons need to be sufficient to explain their deviation from the code. After these initial processing steps, the main excel file containing the prompts for the different provisions is ingested and provisions are analyzed for compliance. The subchapters below will explain the main sections of the workflow in more detail.

3.2.1 Embedding Indexing

Given the length of most annual reports, the methodology employs an embedding-based retrieval technique to provide relevant context to the LLM. This corresponds with the Retrieval Augmented Generation (RAG) approach, where text is broken into chunks, encoded into vector embeddings, and stored in an index. Embeddings are dense vector representations of text that capture semantic meaning, allowing for efficient similarity-based retrieval of relevant pieces of text. For this

[1] From: https://pypi.org/project/pymupdf4llm/

specific case, after some testing I decided to set the size of the chunks used for embedding at 700 tokens with an overlap of 200 tokens. This means the annual report will be indexed about half a page at a time. When a prompt asks the LLM to assess a specific provision, it first retrieves the most relevant text chunks, in my case six, using semantic similarity, and only feeds those into the model. In initial testing, I found that semantic retrieval alone often missed shorter, more explicit answers, particularly when the correct sentence was buried in longer context with lower overall semantic similarity. To address this, I adapted the retrieval method for such provisions by adding a lightweight keyword filter. This hybrid search approach significantly improved the chances of retrieving the correct passage for provisions that are typically answered in just one or two sentences. Eventually the LLM will end up with six pieces of context of 700 tokens, which corresponds to about six times half a page, which it needs to base its conclusion on. I chose these parameters as this seemed to be the best compromise between having enough relevant context, while making sure the LLM was not given too much irrelevant context. The embeddings are generated using OpenAI's "text-embeding-3-large" model. I chose this model as my workflow would be using OpenAI's LLMs as well, and this model is seen as one of the more powerful text embedding models.

Literature on RAG approaches has shown that combining the language model with a vector index of knowledge can greatly improve accuracy on queries aimed at a specific knowledgebase (Lewis et al., 2021). In my case, creating an embedding index of each company's report, even creating separate indexes for specific sections like the remuneration or supervisory board report, helps focus the LLM on the context most important for the task given. By narrowing the context in this manner, the chance that the LLM gets "lost" in irrelevant text is reduced dramatically. Essentially, this method allows the LLM to search within the document for the answer, without needing to read the whole document, which is a robust approach for compliance related tasks that depend on specific statements and paragraphs to be found in the annual reports.

3.2.2 Prompt engineering

A key aspect of working with LLM's is effective prompt engineering. The term prompt engineering refers to carefully producing and refining the instructions and questions given to the LLM to produce the desired output. Instead of altering model parameters, prompt engineering uses task-specific questions to guide the model's focus (Sahoo et al., 2024). The provisions, rules set out by the DCGC, are all turned into one or more prompts depending on the extensiveness of reporting necessary. Literature has shown that a good prompt is able to guide the LLM to the relevant section of the report and has a significant impact on the end result provided by the LLM (Lupo et al., 2023). In particular, clear and structured prompts, in this case including definitions or criteria from the DCGC, showed to

significantly improve the LLM's capability to find relevant context and analyze it correctly. Similar to the paper by Lupo et al. (2023), I iteratively refined the prompts to ensure the model understands each compliance question and uses the correct context from the annual report. Aligning with best practices I ended up moving towards a chain-of-thought method of prompting, as I noticed this allowed the model to deal better with extensive, and sometimes irrelevant context.

The different prompts for each provision will then be ran using the embedded text as a context index. I will test out and refine numerous prompts that instruct the APIs to find the information relevant to each individual provision, to then assess compliance using this relevant text. I will be running the same workflows several times for each document to assess if the variation between answers is to such an extent that it causes reliability issues. It will be interesting to see how the variation between these answers compares to variations between human expert answers, given the same question and text. Furthermore, this allows for later resampling to analyze variance and accuracy.

3.2.3 Running the individual provisions

To assess the different provisions, an excel file is made where each provisions has its own instruction and prompt as well as relevant section where the context should be found, if applicable. For example, provision 1.1.3 of the code gets, an instruction telling the LLM what "role" it should take on while answering the prompt, the prompt telling the LLM what its task is and what the criteria are that should be assessed and the relevant section, in this case the Supervisory Board Report. First filtering the index to only contain these sections helps the LLM in narrowing the scope of possible context, giving it a higher chance of finding the right context and consequently providing the correct assessment.

Some provisions depend on other provisions, either directly or indirectly. For example, provision 2.3.11 depends directly on compliance with six other provisions, or 2.1.6 indirectly depends on compliance with 2.1.5. To handle these edge cases, a different excel file has been created that contains all dependencies. This excel file contains a list of dependencies and a column telling the workflow what type of dependency it is as well as the other provisions it depends on. This information together with the code in the workflow allows the program to handle these edge cases correctly.

3.3 Results of the workflow

The workflow returns several results related to the compliance assessment. First of all, the workflow keeps track of elementary things like which company, and which provision is currently being assessed and return this with the results. The results then consist of the conclusion of the assessment, either non-compliance or compliance, and the reason for its assessment. Furthermore, the page

numbers of the context used for its conclusion are returned to increase transparency. These results can then be used to calculate compliance percentages per provision and per company for further statistical analysis.

After completing the compliance assessment for all provisions the workflow returns several results. The program returns a list of results for each provision including the conclusion, either non-compliance or compliance, and the reason for its conclusion as well as the pages where the information used was found and the company that was being assessed.

3.4 Benchmarking against previous years

To test the external validity of the assessment given by the workflow, I will be analyzing the correlation with compliance scores from previous years. I will be calculating a compliance percentage per provision and/or per company, which can be compared to data from 2022 which is publicly available in a report on the website of SEO[2]. These results are from a different year and unfortunately do not contain all provisions assessed by my workflow, but this is the most recent data available and the code has not changed in between these years, making it the best data available. The data will not provide a perfect comparison, but helps to give a rough estimation of the accuracy and usefulness of the results. The first statistic that will be evaluated is the Pearson correlation coefficient to measure the strength of the correlation to get a first idea. Then, the Intraclass Correlation Coefficient (ICC) will be calculated as this provides an extra check for systematic bias in the correlation. For example, if the LLM workflow consistently has a lower compliance percentage compared to the older human expert assessment, this will not show up in the Pearson correlation, but it will in the ICC. These metrics will serve as a measure of external validity for the workflow to combat any possible inaccuracies or systematic biases.

3.5 Resampling with bootstrapping and Monte Carlo simulations

LLMs are known to have variability in their answers. Even when setting the temperature of an LLM to zero to minimize the creativity and aim for deterministic outputs, there is a small amount of variation in its responses. If this variability can cause the conclusion to be flipped the other way, this can significantly impact the reliability of the results. To assess this uncertainty and its effect on my results I will use two resampling techniques to optimize my workflow and test the robustness of its results.

[2] https://www.seo.nl/wp-content/uploads/2021/11/2021-112-Corporate-Governance-Code.pdf

3.5.1 Monte Carlo simulations

As mentioned earlier, even with the temperature parameter of the LLM set to zero, there is still room for small variability between answers due to non-deterministic behavior. When I give a prompt to my LLM workflow, the LLM will not return the exact same answer on every iteration. This stochastic behavior introduces uncertainty into the compliance assessment, which must be accounted for to accurately assess the robustness of my workflow, especially since some prompts might trigger more variability in answers than others(Reiss, 2023). To combat this, I will run each prompt the exact same way multiple times creating an error distribution for each different provision per company. Then I will employ Monte Carlo simulations to capture the size of the internal variance on provision-level compliance using these error distributions. Subsequently, I will simulate assessment for each of my provisions by use of random noise based on the error distribution found. Assuming random variation, this noise will be introduced in the following way:

$$ComplianceSim_i = \min(1, \max(0, p_i + \varepsilon_i)), \qquad \varepsilon_i \sim N(0, \sigma_i^2) \tag{1}$$

Where p_i is the observed compliance percentage per provision i and σ represents the standard deviation of the compliance rate found for each provision i. I will run this simulation, let's say a thousand times to create a realistic distribution of simulated compliance scores reflecting internal variance.

This truncation of the simulated compliance scores ensures that values remain valid within the range of 0% to 100%, reflecting the real-world bounds and avoiding mathematically implausible values due to random noise. The simulation provides us with two key insights. First of all, the internal variance per provision, highlighting the provisions that possibly cause instability in LLM interpretation. Second of all, which provisions consistently have a stronger correlation with the benchmark of the scores from previous years.

3.5.2 Bootstrapping

Next, bootstrapping will be used to measure the degree of variability, and thus stability, of the found compliance percentages. I will randomly sample a certain amount of companies from the total dataset with replacement. Subsequently, for this bootstrap sample, I will compute the compliance percentage per provision and the correlations with the previous years. I will repeat this process a large number of times, let's say a thousand iterations, to obtain a distribution of the variability of my correlation coefficient. This will allow me to derive confidence intervals around my results and test the robustness of my correlation when the sampled dataset varies.

4. Results

This chapter presents an extensive and thorough examination of the empirical results obtained from my investigation on the reliability of Large Language Models (LLMs) in evaluating corporate compliance with provisions detailed in the Dutch Corporate Governance Code (DCGC). A complete analytical framework, which combines statistical analysis and robust evaluation methodologies, allows my investigation to address the central research question:

> *To what extent can LLMs reliably assess compliance with the Dutch Corporate Governance Code from corporate annual reports?*

With respect to answering the research question, the chapter addresses overall compliance accuracy, explores provision-specific reliability, quantifies uncertainty in compliance evaluations, and critically compares LLM-derived assessments against human expert benchmarks, thus contributing significantly to understanding the potential and limitations of LLMs in automated compliance assessments.

4.1 Overall compliance assessment

Initially an aggregated overview of provision-level compliance was generated from the assessment performed by the workflow. This roughly shows the rudimentary capabilities of the LLM and whether it was able to provide sensible conclusions concerning the different provisions for different types of companies. These results were computed and averaged over multiple runs to ensure robustness in the preliminary assessment. Table 1 shows the average compliance scores and the corresponding standard deviations across iterations on a provision level. In the appendix a more extensive table including sub-provisions will be attached to give a complete initial overview of results.

Table 1: Provision-level compliance summary.

Provision	Compliance Score (% ± SD)
1.1.3	96.2 ± 17.0
1.1.4	77.9 ± 31.1
1.3.6	89.3 ± 28.2
1.4.2	95.4 ± 12.9
1.4.3	93.9 ± 16.1
2.1.10	92.4 ± 26.2
2.1.2	82.7 ± 26.8
2.1.3	97.8 ± 14.1
2.1.6	92.8 ± 20.4
2.2.8	73.2 ± 33.0
2.3.11	36.5 ± 44.7
2.3.5	67.5 ± 33.5
2.4.4	76.8 ± 40.7
2.5.4	70.3 ± 33.3
2.7.4	99.7 ± 2.0
2.7.5	100.0 ± 0.0
3.4.1	75.5 ± 23.5
4.2.6	97.5 ± 13.3
5.1.5	74.5 ± 28.5

Note: Provision 2.3.11 is dependent on compliance with several other provisions. Therefore, the low score is explained by non-compliance with other provisions.

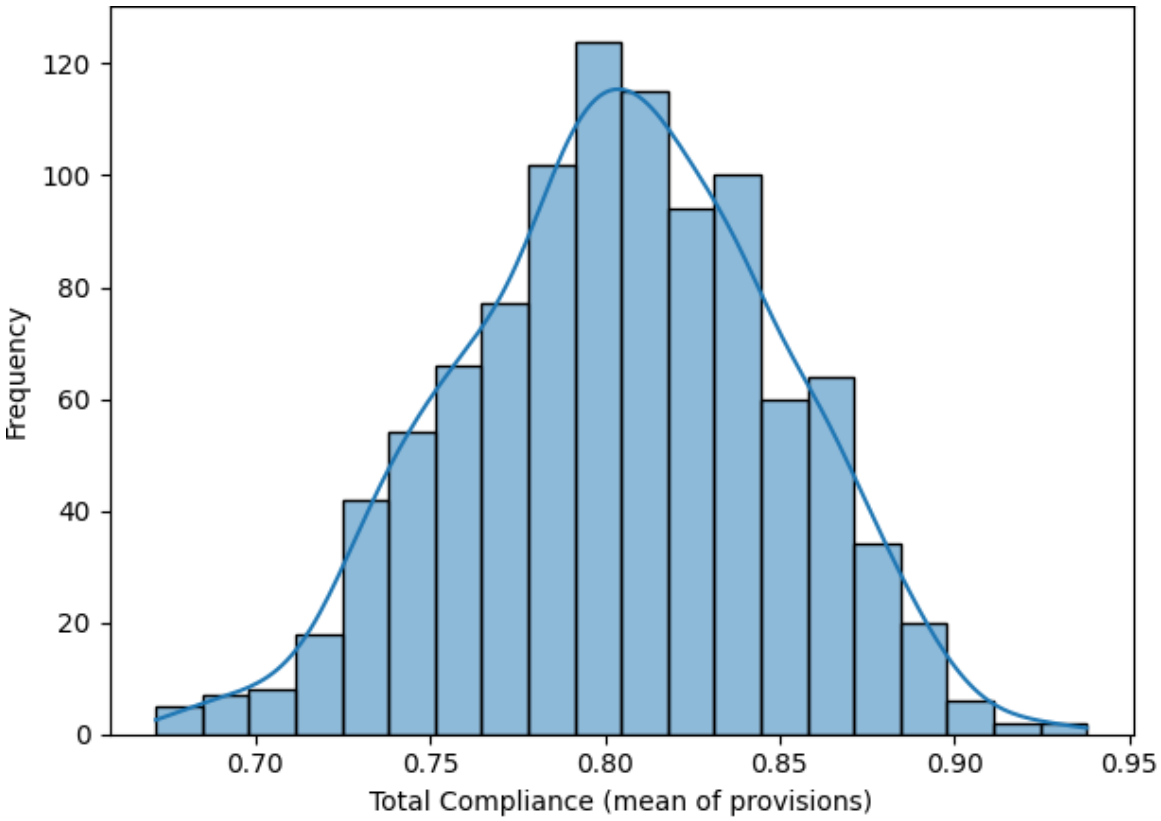
The table demonstrates significant variation among the provisions examined. Provisions such as 1.1.3, 1.4.2, 2.7.4, 2.7.5 and 4.2.6 consistently demonstrate high compliance scores, reflecting the clarity and explicit reporting concerning these provisions, which aids straightforward interpretation by the LLM. Conversely, provisions like 2.5.4, and 3.4.1 show significantly lower compliance scores with greater variability. These are provisions that are more extensive and require more reporting as well as higher quality explanations to satisfy the criteria for compliance. Companies struggle with these aspects and have already exhibited lower scores on these provisions in previous years (Witteman et al., 2022). However, the LLM workflow shows it struggles with these provisions as well. The more extensive reporting as well as the increased difficulty of the criteria that need to be met for compliance cause it to sometimes wrongly assess a provision as non-compliance, causing a noteworthy decrease in compliance compared to the data from 2022.

4.2 Analysis of internal reliability using Monte Carlo simulations

To quantify the reliability of my workflow and assess uncertainty inherent in LLMs I conducted several Monte Carlo simulations. For these simulations the observed variation in compliance assessment across several identical runs was used. These simulations then generated probabilistic distributions of compliance outcomes, offering valuable insights into provision-level reliability and

uncertainty. I ran Monte Carlo simulations for total compliance averaged over all provisions and all companies which resulted in the results in figure 1.

Figure 1: Distribution of total compliance (%) across 1,000 Monte Carlo iterations.

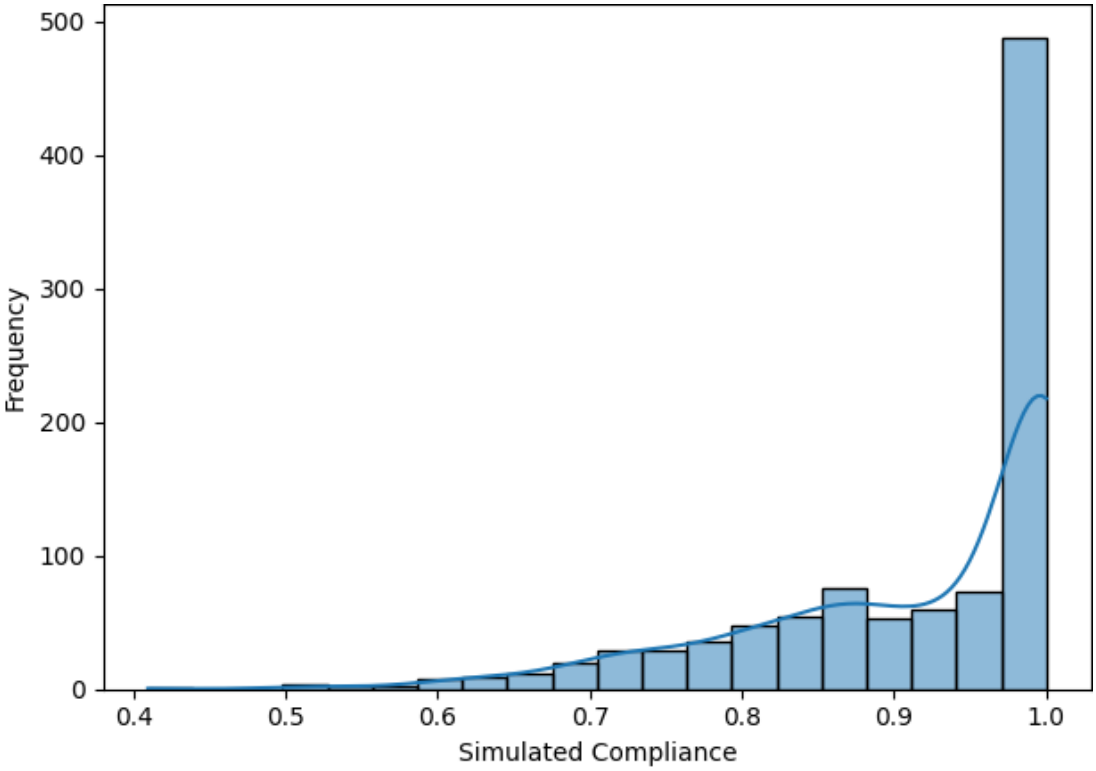


The figure reveals a normal distribution suggesting that possible errors on the provision level tend to cancel each other out. The total compliance percentage converges to a stable mean around 80%, with a relatively modest amount of outliers. This shows that the LLM can provide robust overall compliance estimates, supporting possible utilization for reliable preliminary checks. In the next subchapters I will delve deeper into individual provision reliability to analyze which provisions might possibly be better suited for LLM assessment and which provisions tend to cause more issues.

4.2.1 High reliability provisions

A subset of provisions, including 1.1.3, 1.4.2, 1.4.3, 2.1.3, 2.7.4, 2.7.5 and 4.2.6, revealed highly skewed Monte Carlo distributions with a strong concentration around full compliance. Figure 2 below shows this skewness clearly for provision 1.1.3. Figures for the other provisions will be provided in the appendix. For these provisions the figures confirm a high reliability of the compliance assessment by the LLM. This is most likely due to their explicit criteria as well as the straightforwardness of the reporting needed to comply.

Figure 2: Distribution of compliance with 1.1.3 (%) across 1,000 Monte Carlo iterations.

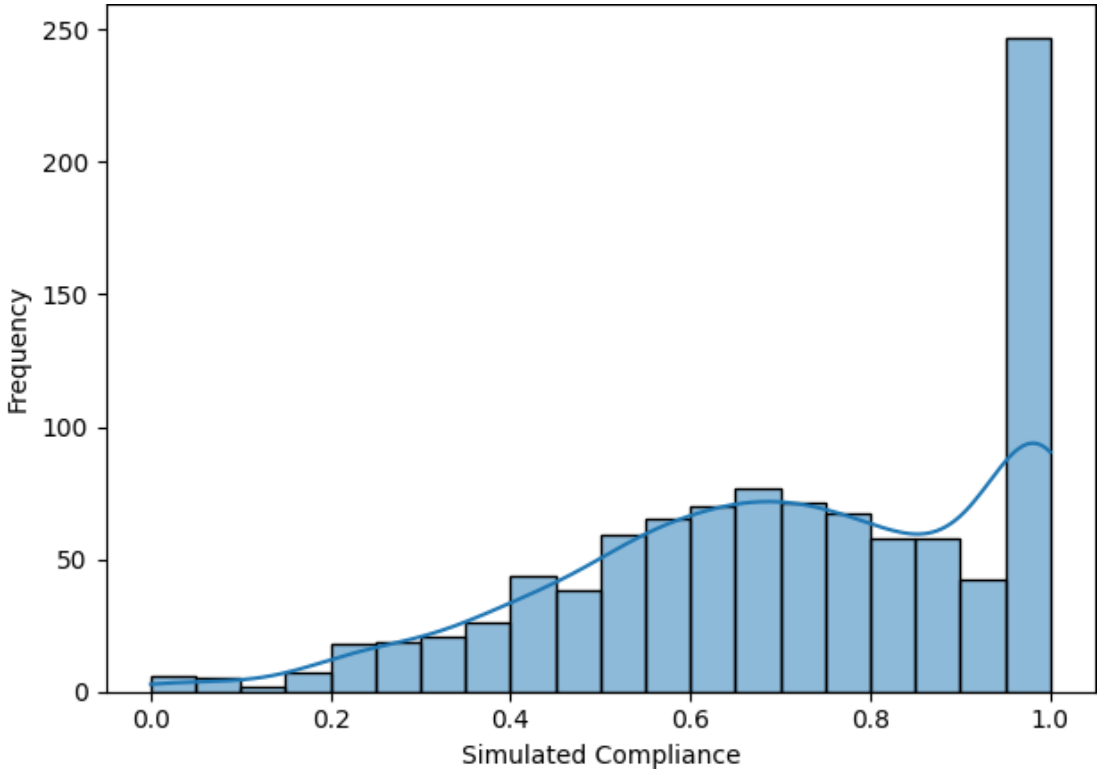


The minimal variability, with most simulations ending up in the 90 to 100% compliance part, underlines the ability of LLMs to reliably handle these type of textual analysis tasks. Thus, clearly showing that for these type of provisions, with clear governance reporting criteria, LLMs can significantly enhance productivity and efficiency while saving costs and potentially increasing accuracy in routine compliance monitoring.

4.2.2 Moderate reliability and ambiguous provisions

Provisions such as 1.1.4, 1.3.6, 2.1.2, 2.1.6, 2.1.10, 2.4.4, and 5.1.5 exhibit moderate reliability. Their Monte Carlo distributions are more dispersed with values spread over a wider range. Some of the graphs show gently sloping bell shapes or bimodal distributions, indicating some variance between test runs for the LLM results. Figure 3 shows provision 5.1.5 where you can clearly see the bimodal shape with a peak of simulations around the 70% level and a peak around the 100% level.

Figure 3: Distribution of compliance with 5.1.5 (%) across 1,000 Monte Carlo iterations.

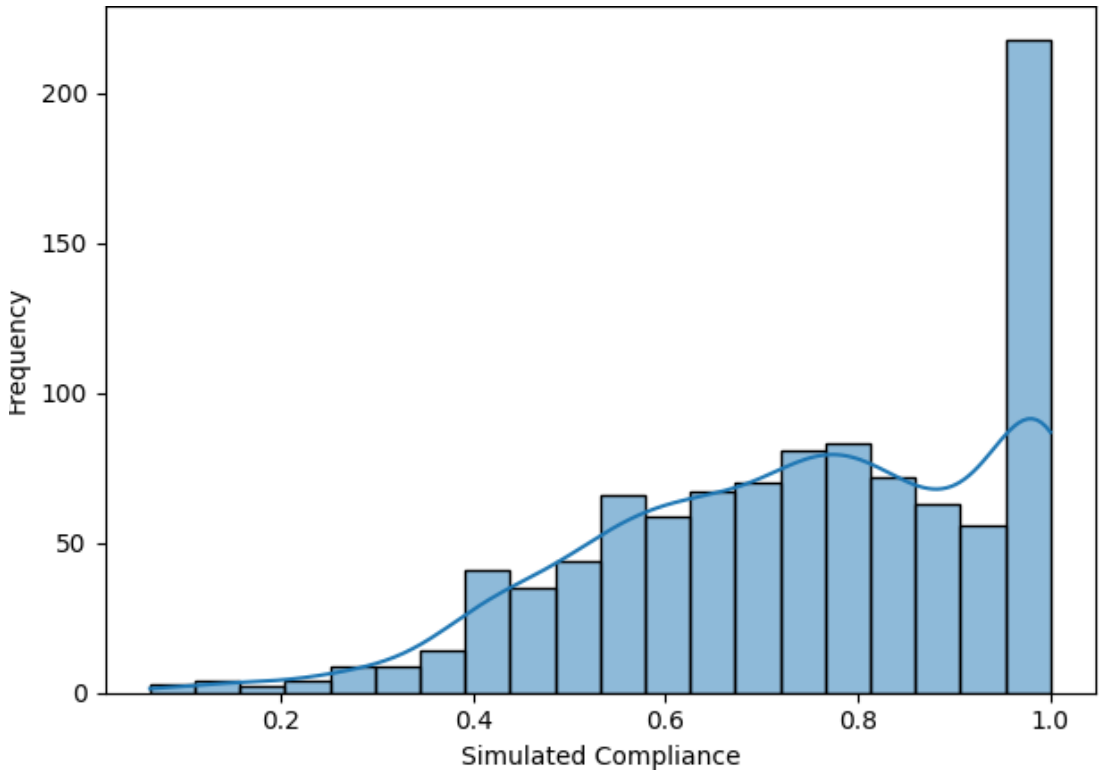


These results suggest that while the LLM can generally assess compliance well with these provisions, certain ambiguities in phrasing in the annual reports can introduce inconsistencies. This group includes provisions like 2.1.2 and 2.4.4 which are reported in different formats like tables or figures which have a higher chance of causing issues for the LLM. Furthermore, provisions like 2.1.6 and 1.1.4 require extensive reporting about strategies and policies. Since our RAG approach only has a context of maximum six chunks of 700 tokens, this can sometimes cause the LLM not to be provided with all the context it needs to correctly assess the provision.

4.2.3 Provisions indicating low reliability

The most noticeable uncertainty showed in provisions 2.2.8, 2.3.5, 2.3.11 and 3.4.1. These provisions showed wide Monte Carlo distributions approaching uniformity or displaying significant density at multiple points in the distribution. Take provision 3.4.1 for example, which shows both a strong peak around the 100% level as well as a significant amount of simulations falling within the 60% to 90% range, as shown in figure 4.

Figure 4: Distribution of compliance with 3.4.1 (%) across 1,000 Monte Carlo iterations.



Such distributions indicate that the LLM has substantial interpretive issues when assessing these provisions. This is probably due to large context needed for correct assessment or vague assessment criteria. The variability indicates that my LLM workflow does not have the capabilities yet to fully give a fully reliable assessment of these types of provisions, highlighting the need for manual checking by humans or more sophisticated prompt engineering.

4.3 Benchmark comparison through correlation analysis

To validate the reliability of the assessment results I compared the compliance results per provision to the benchmark data. This data is found in the earlier monitoring done in 2022 by SEO and publicly published together with their main conclusions on their website(Witteman et al., 2022). A simple Pearson correlation showed a small positive correlation ($r = 0.146$, $p=0.731$), indicating a weak, insignificant, relationship between LLM-based compliance scores and those produced by human experts.

Furthermore, I estimated the ICC to further analyze consistency between the benchmark results and the LLM results. Given the design of this study which only has one way of scoring by the LLM and having only one set of results from previous years, ICC(3,1) was selected as the most appropriate variant. This version of the ICC assumes a fixed set of raters (in this case 2, the LLM and the original data) and specifically quantifies the consistency of the ranking order of the LLM relative to

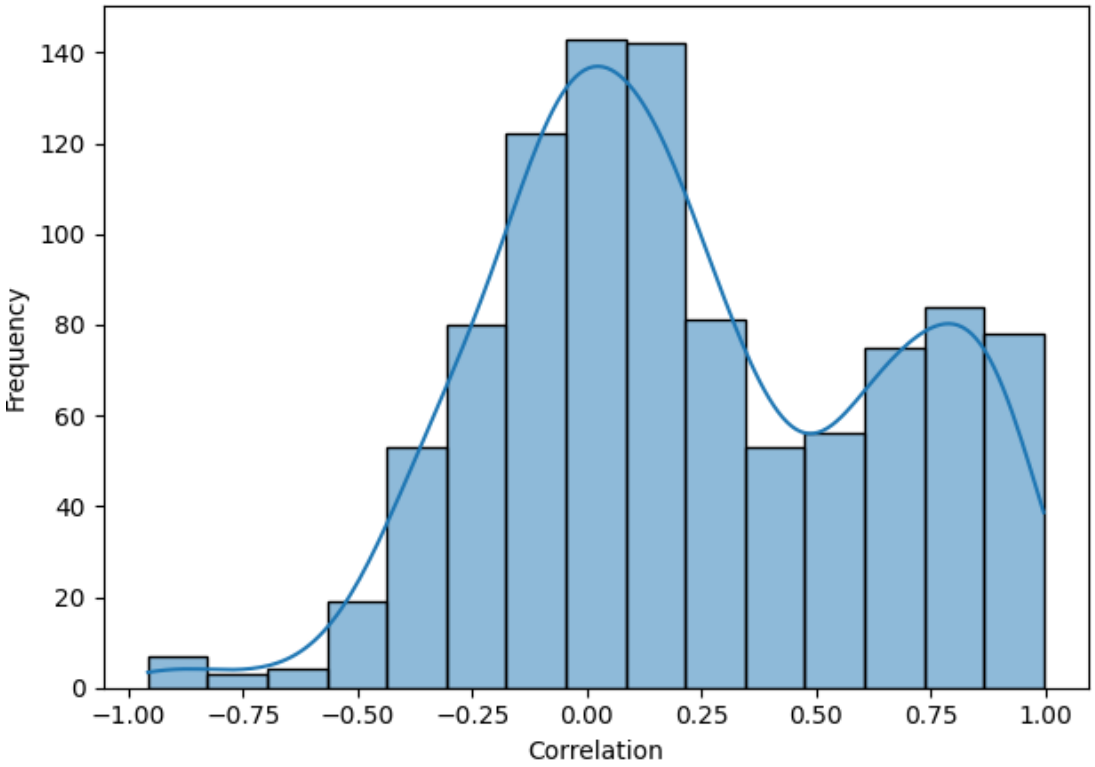
the benchmark. This correlation coefficient gives me a robust estimate of the extent to which the LLM workflow is able to replicate the relative ordering by human experts of provisions by compliance level, while taking into account any systematic differences in the absolute sizes of the scores. The result gave us a correlation coefficient of 0.006 ($p = 0.5$), showing barely any correlation in the ranking, even when taking the systematic differences into account. This value suggests that the LLM's ranking of provisions shows close to no consistency with the benchmark. An important note with this result is the lack of comparison data from previous years. Unfortunately, the data from 2022 only contains results for eight provisions which were analyzed in the same way by my workflow and thus suitable for comparison. This drastically increased the impact of a single provision being in a different order compared to the benchmark.

This finding aligns with the observed Pearson correlation coefficient, confirming the conclusions that while the LLM demonstrates high reliability for clearly formulated provisions, its ability to replicate expert assessment for more complex or subjective provisions remains limited.

4.4 Robustness analysis through bootstrapping

Next to using Monte Carlo simulations for analysis of the consistency of the assessment by the LLM, I also used bootstrapping to analyze the consistency of the correlation results when using different samples of companies. The bootstrapped Pearson correlation coefficients displays clear variability and a noticeable bimodal distribution with peaks around the values of 0 and 0.8, as shown in figure 5.

Figure 5: Distribution of Pearson correlation coefficients across 1,000 bootstrapping iterations.



This variability indicates the fragility of the correlation results, with only a few provisions being compared as well as the possibility of different types of companies being included that might have a bigger chance of not complying with several provisions. Furthermore, smaller companies might have a lower quality annual report which causes more issues for the LLM analysis resulting in wrongful non-compliance assessments. As the figure shows, if some more of these companies are left out of the sample the correlation coefficient has the possibility of drastically increasing with a significant amount of simulations ending up around the 0.8 mark.

5. Conclusion and discussion

In this study I explored the capability of Large Language Models (LLMs) to reliably assess compliance with the Dutch Corporate Governance Code (DCGC) through automated analysis of corporate annual reports. I analyzed corporate annual reports through a carefully designed workflow which uses NLP techniques including Retrieval Augmented Generation, embedding indexing, and prompt refinement. The results demonstrated significant potential for the use of LLMs in corporate governance code assessment.

The results indicated some provisions exhibit high reliability while others are significantly more prone to variation in assessment by the LLM. Provisions such as 1.1.3 and 1.4.2 consistently showed

minimal variance and high compliance accuracy. This is most likely due to multiple factors. First of all, these are provisions that require relatively less context to be assessed, as they can be complied with using a small explanation of only a few sentences. This means my LLM workflow has less chance of not being given the full context necessary for its assessment. Second of all, these provisions are generally easier for the companies to comprehend and comply with, which increases the overall quality of the explanation for compliance, in turn simplifying the assessment for the LLM. However, my results also showed substantial challenges faced by the LLM when assessing provisions that involve excessive context such as 2.3.5 and 3.4.1. For these provisions the LLM showed a greater variability, reducing the reliability of the conclusion. This is most likely due to missing context due to which the LLM does not have the right information to provide a clear assessment.

The above shows that on the provision-level there is still some variability with specific provisions. However, I also analyzed the results at the aggregated provision level, calculating the aggregate compliance percentage over all companies. Monte Carlo simulations in figure 1 showed that the average total compliance percentage was normally distributed around the 80% level, with a relatively small amount of outliers suggesting sufficient internal validity. This suggests that the assessment results from the workflow form a reliable and comprehensive first test of total compliance. although human checks are still recommended as reliability on the provision level is not always guaranteed.

For the analysis of external validity I compared my results to benchmark results from compliance research four years ago. I did this both for the full company sample as well as a bootstrapped subset of the companies in my sample with replacement. The results showed a weak Pearson correlation with the full set of results and a negligible correlation when calculating the Intraclass Correlation Coefficient. This confirms that the results on its own are not suitable to replace human assessment and further fine tuning of the prompts and human checking is needed. Interestingly, the bootstrapping results showed something slightly different. The distribution in figure 5 clearly indicates that, although the correlation found most often is around 0.15, there is a second peak of results which have a correlation of approximately 0.8. This suggests that the LLM approach has the potential to be very accurate for a certain subset of companies. An explanation for this weak correlation for my sample might be that my sample of companies is not representative of the full list of companies which was used to get the data form

With respect to the first hypothesis, that the workflow's overall compliance scores would show moderate to high alignment with historical human expert assessments, the workflow's overall compliance scores barely showed any alignment with historical human expert assessments, thus not

supporting my hypothesis. Important to note is that the bootstrapping distribution in figure 5 showed that there is potential for the correlation to be significantly higher when using a different sample of companies. This indicates that my company sample might not have been representative of the full sample used for the data from 2022. With respect to my second hypothesis, that the internal variability would remain limited and not compromise the workflow's validity, the internal variability of the LLM's assessment remained sufficiently low on the aggregate level as shown in figure 1, indicating stable and practically useful results and confirming my hypothesis at the aggregate level. Important to note is that on an individual provision level there are certain provisions which demonstrate significant variation, which should be interpreted with extra caution.

In conclusion, LLMs show substantial potential for automating corporate governance code assessment, offering improved efficiency, reduced costs and enhanced accuracy. However, their limitations in assessing more extensive and complex provisions underscore the continued need for human oversight and refinement. Future research should focus on enhancing the ability of the LLM to find the relevant context as well as its ability to deal with vague and ambiguous text to improve reliability.

6. Limitations and recommendations

Although my results demonstrate considerable potential, several methodological limitations should be mentioned. First of all, the method used for extracting the context from the annual report PDFs does not extract any images as well as sometimes struggling to extract tables correctly. Other packages like MinerU or Mistral might perform better at this, but are more difficult to integrate and come with higher costs. Second, the embedding and indexing techniques applied require choices to be made about chunk sizes, overlap and embedding models. I did do some refinement and testing to improve the results, but other choices could be have been made potentially leading to better context being found by the LLM. Currently these constraints potentially resulted in incomplete retrieval of relevant context, causing more extensive provisions to display heightened variability in assessment by the LLM. This directly impacted the provision-level interpretation of my second hypothesis, and calls for further research into the differences in performance between different types of provisions.

Another limitation is the limited availability and narrow scope of the benchmark data. With historical compliance data only containing a subset of eight provisions suitable for comparison, the ability to test my first hypothesis, assessing external validity against human evaluations, was limited. Furthermore, the subjective nature of certain provisions causes additional complexity, as variability in human judgment in turn challenges the robustness and consistency of the fully automated LLM

assessments. Furthermore, during the creation of my workflow I used OpenAI's GPT-4o model for most of my assessment as this was the most advanced model available for a reasonable price. However, since writing my code and finishing my tests, OpenAI has announced a major price reduction of their GPT-o3 model, which is arguably an even stronger model as it is better at handling large context and tasks that require reasoning steps. Availability of this model would have impacted my approach and possibly increased the accuracy and reliability of my results.

To mitigate these limitations, future research should focus on experimenting with new context retrieval methodologies, such as enhancing embedding approaches and Retrieval Augmented Generation techniques, to increase the accuracy and reliability of assessments, particularly in case of provisions requiring large amounts of context. Further studies should pursue to test their results against larger more comprehensive benchmark datasets to improve the quality of external validation and revisit the first hypothesis with improved circumstances. This would allow a more detailed and accurate understanding of strengths and limitations of the LLM results and targeted refinement with respect to certain provisions.

Lastly, I recommend to integrate structured human oversight when using automated workflows for assessing these types of rulesets. Although the results look promising, reliability is not always guaranteed and the results given by the LLM approach should be seen as a method of producing a preliminary check which requires refinement and validation by human experts. This hybrid approach uses the efficiency offered by automation while protecting accuracy and ensuring human experts are still the final decision-maker.

7. Bibliography

Aggarwal, P. (2013). *Impact of Corporate Governance on Corporate Financial Performance* (Vol. 13, Issue 3). www.iosrjournals.org

Aguilera, R. V., & Cuervo-Cazurra, A. (2009). Codes of good governance. *Corporate Governance: An International Review*, *17*(3), 376–387. https://doi.org/10.1111/j.1467-8683.2009.00737.x

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Blackwell, R. E., Barry, J., & Cohn, A. G. (2024). *Towards Reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores*. http://arxiv.org/abs/2410.03492

Blodgett, S. L., & O'Connor, B. (2017). *Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English*. http://arxiv.org/abs/1707.00061

Brown, P. F., deSouza, P. V, Mercer, R. L., Della Pietra, V. J., & Lai, J. C. (1992). *Class-Based n-gram Models of Natural Language*.

Cao, Y., Chen, Z., Pei, Q., Dimino, F., Ausiello, L., Kumar, P., Subbalakshmi, K. P., & Ndiaye, P. M. (2024). *RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data*. http://arxiv.org/abs/2404.07452

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2020). *Extracting Training Data from Large Language Models*. http://arxiv.org/abs/2012.07805

Chiang, C.-H., & Lee, H. (2023). *Can Large Language Models Be an Alternative to Human Evaluations?* http://arxiv.org/abs/2305.01937

Clarke, Thomas. (2007). *International corporate governance : a comparative approach*. Routledge.

Coffee, J. C. (2006). *Gatekeepers: The Professions and Corporate Governance*. Oxford University PressOxford. https://doi.org/10.1093/oso/9780199288090.001.0001

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy Prices. *Journal of Finance*, *75*(3), 1371–1415. https://doi.org/10.1111/jofi.12885

Das, B. C., Amini, M. H., & Wu, Y. (2024). *Security and Privacy Challenges of Large Language Models: A Survey*. http://arxiv.org/abs/2402.00888

Du, M., Liu, N., & Hu, X. (2008). *Techniques for Interpretable Machine Learning*.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). *The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation*. https://ssrn.com/abstract=2741682

Erhard, N. L., Werbel, J. D., & Shrader, C. B. (2003). Board of director diversity and firm financial performance. In *Corporate Governance: An International Review* (Vol. 11, Issue 2, pp. 102–111). Blackwell Publishing Ltd. https://doi.org/10.1111/1467-8683.00011

Gallegos, I. O., Rossi, R. A., Barrow, J., Research, A., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3). https://doi.org/10.1162/coli

Grohs, M., Abb, L., Elsayed, N., & Rehse, J.-R. (2023). *Large Language Models can accomplish Business Process Management Tasks*. http://arxiv.org/abs/2307.09923

Guay, W., Samuels, D., Taylor, D., Badertscher, B., Barth, M., Bushee, B., Chapman, K., Hail, L., Hanlon, M., Heinle, M., Hepfer, B., Holthausen, B., Kepler, J., Rawson, C., Schipper, K., Schrand, C., Tsui, D., Van Buskirk, A., Verrecchia, R., … Stice-Lawrence, L. (2016). *Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure*.

Hivarhizov, I. (2024). DEVELOPMENT AND IMPLEMENTATION OF SYSTEMS BASED ON LLM IN FINANCE. *State and Regions. Series: Economics and Business*. https://api.semanticscholar.org/CorpusID:271810600

Kiel, G. C., & Nicholson, G. J. (2003). Board composition and corporate performance: How the Australian experience informs contrasting theories of corporate governance. *Corporate Governance: An International Review*, *11*(3), 189–205. https://doi.org/10.1111/1467-8683.00318

Kim, A. G., Muhn, M., & Nikolaev, V. (2024). *Bloated Disclosures: Can ChatGPT Help Investors Process Information?*

Korkanti, S. (2024). Enhancing Financial Fraud Detection Using LLMs and Advanced Data Analytics. *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, 1328–1334. https://doi.org/10.1109/ICSSAS64001.2024.10760895

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.

Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large Language Models in Finance: A Survey. *ICAIF 2023 - 4th ACM International Conference on AI in Finance*, 374–382. https://doi.org/10.1145/3604237.3626869

Lupo, L., Magnusson, O., Hovy, D., Naurin, E., & Wängnerud, L. (2023). *Towards Human-Level Text Coding with LLMs: The Case of Fatherhood Roles in Public Policy Documents*. http://arxiv.org/abs/2311.11844

Monitoring Commissie Corporate Governance Code. (2025). *Dutch corporate governance code.* https://mccg.nl

Mytton, D. (2021). Data centre water consumption. In *npj Clean Water* (Vol. 4, Issue 1). Nature Research. https://doi.org/10.1038/s41545-021-00101-w

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon Emissions and Large Neural Network Training*.

Persons, O. S. (2005). *The Relation Between the New Corporate Governance Rules and the Likelihood of Financial Statement Fraud* (Vol. 4).

Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark*. http://arxiv.org/abs/2304.11085

Rithika, R. (2024). Recent Advances in Large Language Models: An Upshot. *International Journal of Research Publication and Reviews*, *5*(6), 137–143. https://doi.org/10.55248/gengpi.5.0624.1403

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. http://arxiv.org/abs/2402.07927

SEC. (2013). *Report on Review of Disclosure Requirements in Regulation S-K*.

Smith, M., & Taffler, R. (1992). Readability and Understandability: Different Measures of the Textual Complexity of Accounting Narrative. *Accounting Auditing & Accountability Journal*, *5*(4), 84–98.

Sun, Z. (2023). *A Short Survey of Viewing Large Language Models in Legal Aspect*. http://arxiv.org/abs/2303.09136

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*.

Velte, P. (2023). The link between corporate governance and corporate financial misconduct. A review of archival studies and implications for future research. *Management Review Quarterly*, *73*(1), 353–411. https://doi.org/10.1007/s11301-021-00244-7

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). *ClimateBert: A Pretrained Language Model for Climate-Related Text*. http://arxiv.org/abs/2110.12010

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Com>, <lweidinger@deepmind. (2021). *Ethical and social risks of harm from Language Models*.

Witteman, J., Koeman, N., Rougoor, W., & Verheuvel, N. (2022). *CORPORATE GOVERNANCE CODE BOEKJAAR 2020*.

Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., Xu, S., Dai, H., Zhao, L., Jiang, H., Pan, Y., Chen, J., Zhou, Y., Mai, G., Liu, N., & Liu, T. (2024). *Revolutionizing Finance with LLMs: An Overview of Applications and Insights*. http://arxiv.org/abs/2401.11641