Advances in Social Sciences Research Journal – Vol. 12, No. 12
Publication Date: December 23, 2025
DOI:10.14738/assrj.1212.19750.
Bora, G., Gupta, S., & Gupta, S. (2025). Evaluating ESG Scoring Consistency of Large Language Models Using Retrieval Augmented Generation Methods. *Advances in Social Sciences Research Journal*, *12(12).* 218-234.

Evaluating ESG Scoring Consistency of Large Language Models Using Retrieval Augmented Generation Methods

Gaargi Bora
Atholton High School, Maryland, United States

Sashrika Gupta
West Windsor-Plainsboro High School South, New Jersey, United States

Dr. Sudip Gupta, PhD
John's Hopkins University Carey Business School, United States

ABSTRACT

This study evaluates how large language models (LLMs) determine environmental, social, governance (ESG) scores, utilizing retrieval-augmented-generation (RAG) procedures. Three LLMs–Claude-4, ChatGPT-4o, and Gemini-2.5–were used to find the environmental (E), social (S), and governance (G), scores for a total of nine different, publicly traded companies of three different sizes (small, medium, and large) based on market capitalization. The scores of four companies (Morgan Stanley, Goldman Sachs, Berkshire Hathaway, East West Bancorp) were found using one set of prompts, and the other five (BlackRock, PNC, Bank of America, American Financial Group, and GreenDot) were found using a separate set of prompts, utilizing the same criteria and methods with changes in wording. Both sets of prompts, or each trial, found that RAG approaches produce more stable scores that were more consistent with their existing scores found by established rating agencies (Morningstar Sustainalytics, S&P Global, and JUST Capital). These findings suggest that LLMs demonstrate greater consistency in measuring ESG performance when using RAG methods and providing structured data and criteria, exemplifying the growing capabilities and prospective viability for LLM-determined ESG ratings.

Keywords: Environmental, Social, Governance (ESG), Artificial Intelligence (AI), Financial Institutions, Retrieval-Augmented-Generation (RAG)

INTRODUCTION

Environmental, Social, and Governance (ESG) metrics have become integral to the evaluation of corporate sustainability and ethical performance, guiding investment decisions and stakeholder engagement across global markets. Nevertheless, traditional ESG rating agencies, while well-established, have been criticized for inconsistency, opaque methodologies, and subjective weighting of indicators. Concurrently, large language models (LLMs) have emerged as robust instruments for data analysis and synthesis, demonstrating the capability to process both structured and unstructured information at scale. Despite this advancement, there is a dearth of research and systemic inquiry into the applications of LLMs for ESG assessment, especially via RAG techniques that integrate structured data frameworks with contextual interpretation. This study investigates how diverse LLMs analyze and generate ESG scores through RAG frameworks, examining their interpretive accuracy, consistency, and sensitivity

to prompt design. By comparing LLM-generated outputs with those from established ESG rating agencies, this paper aims to identify methodological distinctions and assess the viability of LLMs as complementary tools for ESG evaluation.

LITERATURE REVIEW

Table 1: Normalized ESG Scores from S&P, Sustainalytics, and JUST Capital (Trial 1)

Company	S&P				Sustainalytics	JUST Capital			
	E	S	G	ESG	ESG	E	S	G	ESG
Morgan Stanley	41	41	42	41.3	50.4	38.938	57.203	48.713	48.285
Goldman Sachs	41	31	45	39.0	49.6	67.257	75.269	53.585	65.380
Berkshire Hathaway	8	13	17	12.7	47.6	29.204	20.177	13.277	20.886
East West Bancorp	13	24	39	25.3	53.0	29.204	46.307	56.043	48.851

Note. Data from S&P ESG Scores and Raw Data, Morningstar Sustainalytics ESG Risk Ratings, and JUST Capital 2025 Overall Rankings

Table 2: Normalized ESG Scores from S&P, Sustainalytics, and JUST Capital (Trial 2)

Company	S&P				Sustainalytics	JUST Capital			
	E	S	G	ESG	ESG	E	S	G	ESG
BlackRock	43	42	51	45.333	45.333	53.097	72.05	66.596	63.914
PNC	37	31	40	36	36	46.018	62.102	81.83	63.317
BofA	51	59	59	56.333	56.333	93.805	75.743	79.787	83.112
American Financial	7	21	29	19	19	29.204	47.459	15.989	30.884
Green Dot	--	--	--	--	--	--	--	--	--

Note. Data from S&P ESG Scores and Raw Data, Morningstar Sustainalytics ESG Risk Ratings, and JUST Capital 2025 Overall Rankings

In Table 1 and Table 2, ESG scores from the three rating agencies (S&P, Morningstar Sustainalytics, and JUST Capital) were normalized to scores out of 100 for reliable comparisons. For S&P, the E, S, and G, were left individually as is.

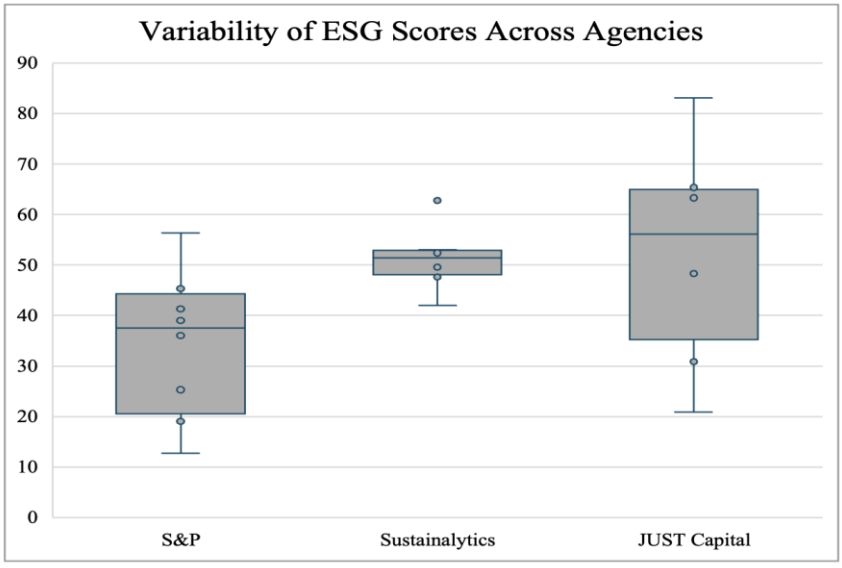

Figure 1: Variability of normalized ESG scores across S&P, Sustainalytics, and JUST Capital.
Note. Data from S&P ESG Scores and Raw Data, Morningstar Sustainalytics ESG Risk Ratings, and JUST Capital 2025 Overall Rankings. Green Dot is not included because S&P Global and JUST Capital do not provide ESG scores for the company.

URL: http://dx.doi.org/10.14738/assrj.1212.19750

The ESG score was found through an average of the three E, S, and G. Sustainalytics's score was found by subtracting the risk rating from 50 and dividing that by 50 (then multiplying that by 100). The normalized scores for JUST Capital were found by separating the scores into 3 categories: E = Environment; S = Customers and Communities; G = Workers, Shareholders and Governance, then dividing each score by the overall best. The average per category (E, S, and G) was multiplied by 100. The overall rating was found by taking the average of the E, S, and G scores to find the overall rating.

Each box plot represents the distribution of overall ESG scores for the same set of companies, highlighting differences in rating dispersion among agencies.

S&P's Corporate Sustainability Assessment (CSA) Score is the S&P Global ESG Score without the inclusion of any modeling approaches. It measures a company's performance on and management of ESG risks, opportunities, and impacts through a combination of company disclosures, media and stakeholder analysis, modeling approaches, and company engagement. The score is relative, measuring a company's performance on and management of ESG risks, opportunities, and impacts compared to their peers within the same industry classification. Conversely, the S&P Global ESG Score uses a double materiality approach–how sustainability issues impact both business (financial materiality) and how the business impacts the environment and society (impact materiality)–where issues are considered material if they present a significant impact on society or the environment and a significant impact on a company's competitive position and long-term shareholder value creation. Ultimately, S&P Global's methodology primarily centered on disclosure and governance, which generally results in lower scores compared to Sustainalytics and JUST Capital. The significant discrepancies seen with companies such as Berkshire Hathaway, as shown in Table 2, suggest that limited data availability and transparency translate to lower scores. However, S&P's approach introduces several potential biases. Firms with the resources to produce comprehensive reports and participate in the CSA tend to perform better, regardless of actual performance. Scores are also evaluated relative to industry peers, meaning that a company in a poorly performing sector can achieve a comparatively favorable score. Additionally, companies that actively collaborate with or invest in S&P may benefit indirectly.

Sustainalytics classifies a company as being in its comprehensive universe or core universe primarily based on its market capitalization and whether it is represented in major global and regional indexes. Sustainalytics' Core Framework covers 20-30 management indicators, whereas its Comprehensive Framework covers over 70 management indicators. Sustainalytics' system includes three key components: controversy ratings, which track ESG-related incidents; the exposure score considers subindustry and company-specific factors such as its business model; and a management score, which assesses how effectively those risks are managed. Exposure refers to the extent to which a company is exposed to its relevant material ESG issues. Management reflects how well a company is managing its relevant ESG issues. Access to a breakdown of the scores' "controversy ratings," however, was blocked for personal accounts (e.g., Gmail, Yahoo!, etc.) and had to be accessed only through a professional email, limiting the ability of individual investors using personal email accounts to view detailed controversy data and perform a full assessment of a company's ESG risk profile. Overall, Sustainalytics' framework measures ESG risk exposure and management, emphasizing how these factors influence enterprise value, rather than broader stakeholder values. As a result, most companies

land in the mid-range, and even firms with lower S&P or JUST Capital scores, seen in Table 2 In the case of Berkshire Hathaway, perform moderately. However, its risk framing treats ESG primarily as a financial concern, not a moral or environmental one. The management score biases reward larger companies with formal documentation and risk systems, compared to smaller companies with less formalized ESG oversight mechanisms. The controversy score penalizes firms that attract media or non-governmental organization (NGO) criticism, even if the issues are minor or already being addressed, introducing a reactive bias.

JUST Capital conducts both qualitative focus groups and quantitative surveys of a representative sample of the American public to determine what issues comprise "just" corporate behavior, how these issues should be defined, and their relative importance (or weight). This organization surveys the American public to determine which ESG issues matter most, gathers behavioral data on Russell 1000 Companies, allows firms to submit evidence-based corrections, and produces rankings that blend public priorities with corporate data. To support their suggested changes, companies are required to provide publicly available sources. JUST Capital creates a thorough company data review, where companies are allowed to review the data collected and provide suggestions for revisions. It also develops a ranking model that implements survey research and company analysis to score and rank companies. An overall ranking of companies is generated, along with industry-level rankings to compare companies' performance to that of their peers. JUST Capital centers on stakeholder and public alignment, focusing on how companies' actions reflect public values and perceptions. Its scores vary widely, with higher-performing firms typically showing strong reputations and visible social or environmental initiatives, even if they underweight internal ESG risks. Nevertheless, a reliance on U.S. public opinion may not represent global ESG priorities; companies with strong public trust and public relations campaigns in the United States may receive inflated scores, even if their actual ESG performance is lacking. Moreover, modeling bias emerges when missing data is filled in using estimates, potentially distorting results. This trend is especially damaging for small or medium-cap companies, where there tends to be less of a focus on factors contributing to higher ESG scores, or where some factors are irrelevant to certain industries.

METHODOLOGY

To maintain consistency and reduce sector-specific variability, this study focuses exclusively on the ESG scores of financial companies. Compared to other industries, the financial sector demonstrates relatively uniform reporting standards and ESG impacts, which can aid in mitigating inconsistencies in data interpretation and scoring. Moreover, ESG ratings for financial companies are among the most frequently analyzed and publicly scrutinized, with the sector occupying leading positions in investment portfolios and market research. This focus also reflects the status quo, as many financial corporations hold stakes in and maintain partnerships with major ESG rating agencies, further underscoring this sector's influence on the evolving landscape of sustainability assessment. Within this sector, companies are further segmented by market capitalization to capture potential differences in disclosure quality, data accessibility, and model interpretability across firm sizes. Larger firms–BlackRock, PNC, Bank of America, Morgan Stanley, Goldman Sachs, and Berkshire Hathaway–tend to have more comprehensive sustainability reporting and established investor relations infrastructures, while smaller firms–East West Bancorp and American Financial Group (mid-cap); Green Dot (small-cap)–may exhibit greater variability and limited ESG transparency. This stratification

enables a more controlled analysis of how large language models handle varying levels of data availability and complexity.

Retrieval-augmented generation (RAG) was utilized in this study to enhance the reliability and contextual grounding of LLM-generated ESG assessments. By integrating a retrieval component, RAG frameworks allow the model to dynamically access and synthesize external information prior to generating a response, mirroring the methodology of human ESG analysts, absent potential bias. Links [Table 4] containing information from the indicators [Table 3] – including sustainability reports, regulatory filings, and verified financial disclosures– were compiled and provided to the LLMs, ensuring that all data retrieved was sourced from credible and relevant references. Attempts 1, 3, and 4 were conducted utilizing this RAG approach. In attempts 2 and 5, the LLM was instructed to find relevant and credible information on its own.

Table 3: ESG Indicators Used in the RAG Prompts

Indicator Type	Environmental (E)	Social (S)	Governance (G)
Quantitative	1. Greenhouse gas emissions (direct and indirect energy-related) in tons 2. Hazardous waste generated (in tons) 3. Non-hazardous waste generated (in tons) 4. Nitrogen oxide emissions 5. Sulfur oxide emissions 6. Particulate emissions (particle pollution) 7. Total energy consumption 8. Total gas or oil consumption 9. Total water consumption 10. Total paper consumption (and density) 11. Packaging materials used for finished products (in tons and per unit, if applicable)	1. Gender breakdown of the total workforce 2. Types of employment by total workforce 3. Age group breakdown of the total workforce 4. Geographical region breakdown of the total workforce 5. Gender-based employee turnover rate 6. Age group-based employee turnover rate 7. Rate of employee turnover by region 8. Turnover rate of employees by type of employment 9. Work-related fatalities that occurred in each of the past three years, including the year under review 10. Injuries at work resulting in lost days 11. Gender-based employee training percentages 12. Employee training percentages by category 13. Gender-specific average training hours completed by employees 14. By employee category, the average number of training hours completed per employee 15. Geographical distribution of suppliers 16. Recalls of products due to safety and health concerns as a	1. Number of executives and directors 2. Proportion of female executives and directors 3. Total amount and number of antitrust fines and settlements 4. Total value of green bonds issues

		percentage of total products sold or shipped 17. Number of complaints about products and services, and how they are handled	
Text	1. Reduction targets for greenhouse gases, dust, and other emissions, and actions taken to meet them 2. Waste reduction targets (hazardous and/or nonhazardous), and actions taken to achieve them 3. Reduction targets for noise and air pollutants, and mitigation steps 4. Targets and actions to improve energy, water, and building material efficiency 5. Reduction goals for packaging and construction materials and steps taken to meet them 6. Significant impacts on soil and water resources, and corresponding management measures 7. Policies addressing physical risks of climate change and mitigation strategies 8. Policies addressing transitional (economic/regulatory) risks of climate change and responses 9. All environmental performance targets and implementation plans 10. Approach to setting, tracking, and achieving environmental goals	1. Policies in place to ensure workplace health and safety, including their implementation and monitoring. 2. Procedures for assessing employment policies to prevent child and forced labor, and actions taken to address and eliminate child or forced labor if discovered 3. Standards used when working with suppliers, the number of suppliers to which these standards apply, and how implementation and compliance are monitored. 4. Methods used to determine environmental and social risks in the supply chain, along with monitoring and enforcement practices. 5. Criteria and practices for selecting suppliers that prioritize environmentally friendly products and services, including how these are applied and tracked. 6. Number and nature of product and service-related complaints received and the process used to resolve them. 7. Measures taken to respect and safeguard intellectual property rights. 8. Quality control procedures and steps followed in the event of a product recall. 9. Consumer data protection and privacy policies in place, along with the mechanisms used to ensure their enforcement and oversight. 10. Training programs provided to directors and staff on preventing corruption. 11. Initiatives and contributions made to support education. 12. Efforts made to address and support environmental sustainability.	1. Actions on ESG related issues 2. Procedures for evaluating board and executive performance 3. Rules and processes for electing board members and senior executives 4. Structures and policies are in place to manage business ethics and prevent corruption 5. Procedures for reporting ethical violations or corruption 6. Management of risks related to ethics and corruption 7. Publicly disclosure of incidents of corruption or ethical misconduct 8. Antitrust or antimonopoly rectification measures 9. Policies and/or management systems addressing risks and ethical issues related to science and technology 10. Investments in projects related to ESG goals

		13. Contributions toward addressing labor market needs or workforce development.	

In Table 3, 82 indicators were compiled and sorted into subcategories of E, S, G, and whether they were text-based or quantitative. The LLMs used a RAG approach to retrieve information solely based on these indicators for each company, and for each indicator.

Table 4: Sources of ESG Data by Company

Company	Sources	Company	Sources
Morgan Stanley	1. 2022 Diversity and Inclusion Report 2. 2023 ESG Report 3. 2025 State of the Workplace Study 4. 2025 Code of Ethics and Business Conduct	**BlackRock**	1. 2024 GHG Emissions Report 2. 2024 Proxy Statement 3. 2024 Climate Report 4. Code of Business Conduct and Ethics
Goldman Sachs	1. 2023 Statista Gender Diversity in the Workforce 2. 2023 Form 10-K 3. 2024 Approach to Stewardship in Asset Management 4. 2024 Privacy Policy 5. DiversIQ Diversity Profile 6. Corporate Governance 7. Security & Fraud Awareness	**PNC**	1. 2023 Climate Response 2. 2024 Corporate Responsibility Report 3. 2025 Proxy Statement 4. Supplemental Financed Emissions and Emission Intensity Disclosure
Berkshire Hathaway	1. 2024 Schedule 14A Information – Definitive Proxy Statement 2. 2025 Reuters Special Report: Buffett's Berkshire Hathaway operates the dirtiest set of coal-fired power plants in the US	**Bank of America**	1. 2024 Sustainability Report 2. 2025 Code of Conduct 3. 2025 Proxy Statement
East West Bancorp	1. 2024 Form 10-K	**American Financial Group**	1. 2024 Corporate Social Responsibility Report 2. 2024 Form 10-K 3. 2025 Proxy Statement 4. Stock Ownership Guidelines
--	--	**GreenDot**	1. 2022 ESG Report 2. 2024 Form 10-K 3. Code of Business Conduct and Ethics 4. Board Risk Committee Charter 5. Nominating and Corporate Governance Committee Charter

Table 4 includes ESG-related documents, proxy statements, and corporate responsibility reports that were provided to the LLMs in the RAG-based attempts. It ensures consistency across firms by drawing from comparable disclosures such as sustainability, ethics, and governance reports.

Table 5: Prompts Used in the RAG Process (Trial 1- Morgan Stanley, Goldman Sachs, Berkshire Hathaway, East West Bancorp)

Method	E/S/G Individual	ESG
With Data	You are an expert in the field of ESG (Environmental, Social, and Governance). Create an [E/S/G] rating for [company] using these indicators: [21/30/14 indicators]. Use external data from the web but also utilize these resources: [links]. If there isn't enough data for an indicator, provide an estimate for each indicator. If there is not enough data for a proper score, start at 50, and then go above or below 50 based on the available data. Additionally, ensure that you are using all available information from the web, not just the provided documents. Provide a score for every single indicator out of 100 and then calculate a mean [E/S/G] score.	You are an expert in the field of ESG (Environmental, Social, and Governance). Create an ESG rating for [company] using these indicators: [82 indicators]. Use external data from the web but also utilize these resources: [links]. If there isn't enough data for an indicator, provide an estimate for each indicator. If there is not enough data for a proper score, start at 50, and then go above or below 50 based on the available data. Additionally, ensure that you are using all available information from the web, not just the provided documents. Provide a score for every single indicator out of 100, and then calculate a mean E, S, and G score, as well as an overall mean ESG score.
Without Data	You are an expert in the field of ESG (Environmental, Social, and Governance). Create an [E/S/G] rating for [company] using these indicators: [21/30/14 indicators]. If there isn't enough data for an indicator, provide an estimate for each indicator. If there is not enough data for a proper score, start at 50, and then go above or below 50 based on the available data. Additionally, ensure that you are using all available information from the web, not just the provided documents. Provide a score for every single indicator out of 100 and then calculate a mean [E/S/G] score.	You are an expert in the field of ESG (Environmental, Social, and Governance). Create an ESG rating for [company] using these indicators: [82 indicators]. If there isn't enough data for an indicator, provide an estimate for each indicator. If there is not enough data for a proper score, start at 50, and then go above or below 50 based on the available data. Additionally, ensure that you are using all available information from the web, not just the provided documents. Provide a score for every single indicator out of 100, and then calculate a mean E, S, and G score, as well as an overall mean ESG score.

In Table 5, the "with data" prompts included news articles from reputable sources pertaining to the indicators, alongside corporate documents–cited in references– which were provided to the LLMs to generate responses using a RAG approach. For prompts generated "without data," the LLMs found data on each of the companies by themselves, without any documents provided. For ChatGPT and Gemini, each chat received a prompt from the "ESG" category for one out of the four companies, summing up to a total of four chats per attempt (ex, Attempt 1: Morgan Stanley ESG, Goldman Sachs ESG, etc.). For Claude, scores for E, S, and G were sent in separate chats per company, summing up to a total of twelve chats per attempt (ex, Attempt 1: Morgan Stanley E, Morgan Stanley S, Morgan Stanley G, etc.). Results from the prompts used in this table are displayed in Tables 7 and 8.

Table 6: Prompts Used in the RAG Process (Trial 2- BlackRock, PNC, Bank of America, American Financial Group, Green Dot)

Method	E	S	G
With data	You are a professional environmental risk analyst. Consider the following environmental risk indicators: [21 environmental indicators]. Find data on each of these indicators and rate each risk on a scale of 0-(100/21). add up all the scores to find an environmental risk score out of 100, with 100 being the best, and 0 being the worst. Omit any indicators that are irrelevant to the industry/company and recalculate the score so all the indicators add up to 100 (example- if indicator #8 is irrelevant, recalculate all the other indicators on a scale of 0-(100/20) each). If the indicator is relevant to the industry but there is no available data, give it a 50% score. Companies of varying sizes will be listed, alongside company information. Search the web for any data that isn't available in the sources provided.	You are a professional social risk analyst. Consider the following indicators: [30 social indicators]. Find data on each of these indicators and rate each risk on a scale of 0-(100/30). Add up all the scores to find a social performance score out of 100, with 100 being the best, and 0 being the worst. Omit any indicators that are irrelevant to the industry/company and recalculate the score so all the indicators add up to 100 (example- if indicator #8 is irrelevant, recalculate all the other indicators on a scale of 0-(100/29) each). Companies of varying sizes will be listed, alongside company information. Search the web for any data that isn't available in the sources provided.	You are a professional governance risk analyst. Consider the following indicators: [14 governance indicators]. Find data on each of these indicators and rate each risk on a scale of 0-(100/14). Add up all the scores to find a governance performance score out of 100, with 100 being the best, and 0 being the worst. Omit any indicators that are irrelevant to the industry/company and recalculate the score so all the indicators add up to 100 (example- if indicator #8 is irrelevant, recalculate all the other indicators on a scale of 0-(100/13) each). If the indicator is relevant to the industry but there is no available data, give it a 50% score. Companies of varying sizes will be listed, alongside company information. Search the web for any data that isn't available in the sources provided.
Without	You are a professional environmental risk analyst. Consider the following environmental risk indicators: [21 environmental indicators]. Find data on each of these indicators and rate each risk on a scale of 0-(100/21). Add up all the scores to find an environmental risk score out of 100, with 100 being the best, and 0 being the worst. Omit any indicators that are irrelevant to the industry/company and recalculate the score so all the indicators add up to 100 (example- if indicator #8 is irrelevant, recalculate all the other indicators on a scale of 0-(100/20) each). If the indicator is relevant to the industry but there is no available data, give it a 50% score. Do this for the	You are a professional social risk analyst. Consider the following indicators: [30 social indicators]. Find data on each of these indicators and rate each risk on a scale of 0-(100/30). Add up all the scores to find a social performance score out of 100, with 100 being the best, and 0 being the worst. Omit any indicators that are irrelevant to the industry/company and recalculate the score so all the indicators add up to 100 (example- if indicator #8 is irrelevant, recalculate all the other indicators on a scale of 0-(100/29) each). If the indicator is relevant to the industry but there is no available data, give it a 50% score. Do this for the	You are a professional governance risk analyst. Consider the following indicators: [14 governance indicators]. Find data on each of these indicators and rate each risk on a scale of 0-(100/14). Add up all the scores to find a governance performance score out of 100, with 100 being the best, and 0 being the worst. Omit any indicators that are irrelevant to the industry/company and recalculate the score so all the indicators add up to 100 (example- if indicator #8 is irrelevant, recalculate all the other indicators on a scale of 0-(100/13) each). If the indicator is relevant to the industry but there is no available data, give it a 50% score. Do this for the

	following companies: [companies].	following companies: [companies].	following companies: [companies].

In Table 6, "with data" prompts included company names alongside corporate documents (listed in Table 4), which were provided to the LLMs to generate responses using a RAG approach. For the prompts "without [data]," the LLMs found data on each of the companies by themselves without any documents provided. These prompts were sent in three separate chats for all three LLMs. For Chat-GPT, once the prompt was sent, each company and its data were sent subsequently in separate messages. For Claude and Gemini, the companies and their respective data were included in the original prompt, so they were able to process one long message as opposed to 6 shorter ones.

RESULTS

Table 7: ESG Scores Generated Using the RAG Model (Trial 1)

Company	LLM	E			S			G			ESG		
		1	3	4	1	3	4	1	3	4	1	3	4
Morgan Stanley	Claude	66.4	82.9	67.6	68.9	82.4	81.7	81.1	84.6	78.5	72.13	83.3	75.93
	Chat GPT	80.95	55.7	61.9	83.33	55.5	57.8	71.43	71.4	67.9	78.5	61	62.53
	Gemini	100	70.24	71.67	100	71.33	71.39	100	79.29	84.29	100	73.62	75.78
Goldman Sachs	Claude	62.4	70.5	62.4	70.9	72.5	67.4	78.2	71	64.3	70.5	71.3	64.7
	Chat GPT	71.43	64.3	58.9	73.33	59.3	61.8	64.29	60.7	65	69.7	61.5	61.9
	Gemini	95.24	70.24	74	83.33	66.83	78.8	92.86	75	80	90.48	71.83	77.6
Berkshire Hathaway	Claude	59.1	34	36.1	62.1	58	57.4	69.6	55.55	68	63.83	49.18	53.83
	Chat GPT	23.8	41.67	48.33	33.35	53.17	49.67	28.57	55.36	51.43	28.6	49.9	49.81
	Gemini	23.81	55.95	53.2	0	54.17	54.7	50	67.86	52.1	24.6	61.14	53.3
East West Bancorp	Claude	46.9	54	50.9	60.3	84.8	76.4	67.1	84.6	76.8	58.1	74.46	68.03
	Chat GPT	29.57	43	25	50	67	40	42.9	63	50	40.5	57.67	36.67
	Gemini	95.24	44.29	55.4	60	49.17	43.1	71.43	69.64	55.2	59.68	57.24	67.9

Table 8: ESG Scores Generated Without the RAG Model (Trial 1)

Company	LLM	E		S		G		ESG	
		2	5	2	5	2	5	2	5
Morgan Stanley	Claude	71.9	61.43	81	73.7	81.1	68.2	75.5	67.77
	Chat GPT	85.71	64.8	86.67	66.2	71.43	72.5	83.7	67.83
	Gemini	95.24	67.6	93.33	63	100	69.17	93.81	66.25
Goldman Sachs	Claude	78.8	64	76	70.8	78.2	56	76.73	63.6
	Chat GPT	66.67	76.4	60	73.2	64.29	78.9	61.3	76.16
	Gemini	95.24	60.71	83.33	59.83	92.86	67.4	90.48	62.56
Berkshire Hathaway	Claude	38.1	52.2	64.1	54.2	69.6	59.1	50.31	55.17
	Chat GPT	19.05	49.7	26.67	52.1	28.57	56.4	22.4	52.73
	Gemini	23.81	83.6	66.7	64.3	50	71.4	19.68	73.1
East West Bancorp	Claude	55.8	53	85.3	75.4	67.1	74.4	75.32	67.6
	Chat GPT	23.81	50.5	40	50.7	42.9	53.2	33.2	51.46
	Gemini	47.62	52.9	60	66.8	71.43	85	59.68	68.23

Table 7 includes results found from attempts 1, 3, and 4, utilizing a RAG approach. Table 8 includes results from attempts 2 and 5, which were found without a RAG approach, in which each LLM found data to create the ESG score on its own. Each table includes the individual E, S, and G scores from each attempt, as well as the overall ESG score, calculated as the mean of all three individual category scores. The results in these tables (7 and 8) were found using the prompts in Table 5, indicators in Table 3, and links in Table 4.

URL: http://dx.doi.org/10.14738/assrj.1212.19750

Table 9: ESG Scores Generated Using the RAG Model (Trial 2)

Company	LLM	E (1)	E (3)	E (4)	S (1)	S (3)	S (4)	G (1)	G (3)	G (4)	ESG (1)	(3)	(4)
Morgan Stanley	ChatGPT-4o	68	72	64.87	78.7	77.5	61.2	91.28	89.3	51.2	79.16	79.6	59.09
	Claude-4	71.4	82	68.1	72.8	76.67	78.4	78.5	79.7	85.7	74.23	79.46	77.4
	Gemini-2.5	27.44	55.75	60.5	62.61	63.3	49.99	72	75	60.7	54.02	64.68	57.06
Goldman Sachs	ChatGPT-4o	67.8	67	62.17	69	70.4	66	91.36	75	68.4	76.05	70.8	65.52
	Claude-4	80.4	74	80	82.8	85.33	88.2	73.6	73.1	78.6	78.93	77.48	82.27
	Gemini-2.5	31.01	58.5	75	44.82	66.85	60.36	80	81	67.9	51.94	68.78	67.75
Berkshire Hathaway	ChatGPT-4o	79.32	74	59	85.7	73.2	63	96.6	81.8	78.01	87.21	76.33	66.67
	Claude-4	86.5	83	50	89.1	74.33	83.4	72.9	80.8	57.1	82.83	79.38	63.5
	Gemini-2.5	52.28	61	56.5	72.66	63.3	54.43	63	90.82	71.4	62.65	71.71	60.78
East West Bancorp	ChatGPT-4o	40	52	45.38	46.6	46.6	58	89.3	66.4	47.7	58.63	55	50.36
	Claude-4	38.7	53	50	73.6	72	61.5	67.7	72.1	50	60	65.7	53.83
	Gemini-2.5	49.21	50	47.5	40.7	56.3	54.06	80	90.35	50	56.64	65.55	50.52

Table 8: ESG Scores Generated Without the RAG Model (Trial 2)

Company	LLM	E (2nd)	E (5th)	S (2nd)	S (5th)	G (2nd)	G (5th)	ESG (2nd)	ESG (5th)
BlackRock	ChatGPT-4o	42	68.9	17	73.21	28	61.38	29	67.83
	Claude-4	45	71.4	75.8	78.5	65	75	61.93	74.97
	Gemini-2.5	68.2	78.57	64.82	53.9	89.88	67.86	74.3	66.78
PNC	ChatGPT-4o	51	75.8	56	71.43	88	70.7	65	72.64
	Claude-4	52	65.2	73.6	78	78	74.5	67.87	72.57
	Gemini-2.5	85.5	77.27	69.88	67.1	80.2	82.14	78.53	75.5
BofA	ChatGPT-4o	38.5	72.1	52.5	92.86	78	76.64	56.33	80.53
	Claude-4	38	73.8	78.1	82.5	85.6	74	67.23	76.77
	Gemini-2.5	70.6	71.43	72.18	71.1	91.12	75	77.97	72.51
AFG	ChatGPT-4o	14	50.7	54	69.64	56	66.7	41.33	62.35
	Claude-4	58	52.4	68.5	71.2	69.5	67.5	65.33	63.7
	Gemini-2.5	39.5	65.52	57.19	85.5	74.7	71.43	57.13	74.15
GreenDot	ChatGPT-4o	17	50.7	51	69.64	40	63.43	36	61.26
	Claude-4	48	57.1	65.2	65.9	62.5	61.5	58.57	61.5
	Gemini-2.5	22.8	2.38	59.32	55.3	70.78	46.43	50.97	34.7

Table 9 includes results found using a RAG approach. Attempts 1, 3, and 4 all used a RAG approach for each of the companies in each LLM. Table 10 includes results from attempts 2 and 5, which were found without a RAG approach, where each LLM found data to create the ESG score on its own. Each table includes the E, S, and G scores from each attempt, as well as the overall ESG score, calculated as the average of the three individual scores. The results in these tables (9 and 10) were found using the prompts in Table 5 and the indicators in Table 3.

DISCUSSION

By LLM (Trial 1)

Fig. 2. Average (mean) of the total ESG scores found across all attempts, separated into No-RAG (attempts 2 and 5) and RAG (attempts 1, 3, and 4). Error bars represent one standard deviation from the mean, indicating the variability of scores across attempts.

Claude consistently assigned lower ESG scores compared to ChatGPT-4o and Gemini, maintaining a relatively narrow scoring range across all indicators. It rated companies on a 100-point scale, providing specific justifications including direct links to all the data retrieved

for each indicator. Claude's scores were typically "middle of the road" and displayed strong stability across multiple attempts, making it the most consistent model, regardless of whether data was provided or absent. LLM's methodology aligned closest with Sustainalytics, which uses a risk-based approach to ESG assessment. Similarly, Claude appeared to automatically assign baseline or moderate scores when data was unavailable, rather than penalizing companies heavily for missing information.

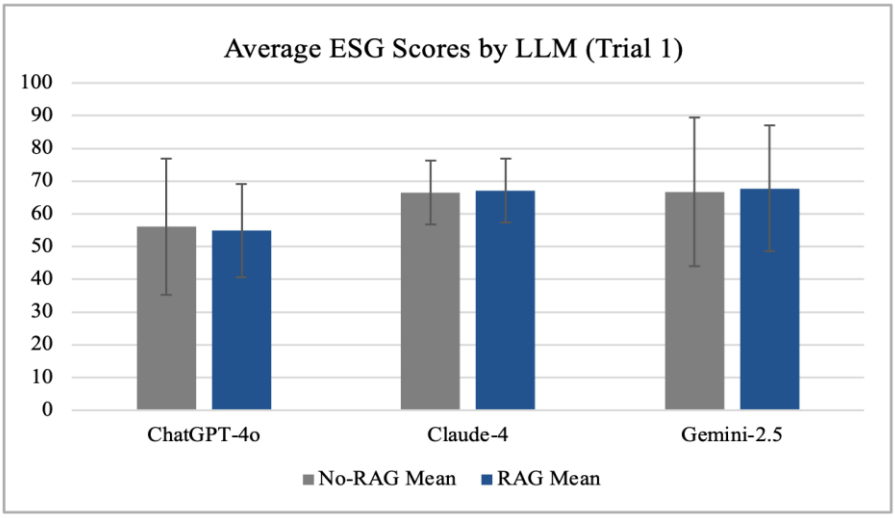

Figure 2: Average ESG Scores by LLM (Trial 1)

ChatGPT-4o based its ESG scores primarily on data availability and disclosure levels, aligning closely with S&P's methodology. It rated environmental (E) and governance (G) factors more stringently than social (S) criteria, and explicitly noted when insufficient data prevented scoring. Overall, ChatGPT-4o's scores were generally lower than Gemini's and displayed greater stability compared to Claude's fluctuations. However, when updated to GPT-5, the model demonstrated notable changes– rating Morgan Stanley and Goldman Sachs significantly lower, while Berkshire Hathaway and East West Bancorp received much higher scores. This shift reflected GPT-5's enhanced retrieval-augmented generation (RAG) capabilities, allowing more efficient integration of external data. Additionally, responses under the updated model were more concise, at times only providing a singular point of justification under quantitative and text categories for each E, S, and G scores.

Gemini consistently assigned the highest ESG scores across all companies and exhibited large variation between attempts. Despite explicit instructions to incorporate external data, the model often relied solely on the documents provided, reflecting an initial inability to extract external data. Nonetheless, Gemini offered the most detailed justifications and methodological transparency, explaining the rationale for each score in depth. It frequently produced extreme highs and lows, inflating strong performances while heavily penalizing weak ones, especially in earlier attempts. Its scoring approach corresponded most closely with JUST Capital, emphasizing stakeholder perception and public-facing ESG commitments, rather than internal risk exposure. However, as a result, scores were extremely sensitive to potential scandals and controversies, inflating performance for companies with more robust public relations departments. Moreover, Gemini reflects JUST Capital's reliance on U.S. public opinion, with

some scores being attributed to public views solely in America, despite all four companies being international.

In general, the LLMs initially relied almost exclusively on the supplied files, despite being prompted to use and evaluate external resources. Nevertheless, in later rounds–when explicitly instructed again, they demonstrated a greater use of retrieval and interpretation, highlighting the importance of clear, structured prompts in optimizing LLM performance.

By Company (Trial 1)

Morgan Stanley received the highest ESG scores across the board. Scores generally increased from attempt 1 to attempt 2 but sometimes dropped in attempt 3. Moreover, this company was the most consistent across all rating systems. Scores were consistent when data was provided but dipped lower without provided data.

Goldman Sachs exhibited mid-to-high scores and had the lowest average variability across E, S, and G. This company received lower scores across disclosure-based rating systems (S&P and Chat GPT), average scores under risk-based systems (Sustainalytics and Claude), and received higher scores under stakeholder/public-driven systems (JUST Capital and Gemini). Scores were somewhat consistent when data was provided but dipped lower without provided data.

Berkshire Hathaway received the lowest scores across the board, likely due to the LLM's reliance on internal documents that reflected a negative environmental impact. The company received a higher score in the round with ChatGPT-5, as compared to the first two rounds, demonstrating that the LLM was utilizing external sources more. Due to this, it had high variability in E (19→41) and G (21→53). Additionally, the company had the largest discrepancy across rating systems. Scores showed the largest divergence in the absence of provided data, likely because of its sparse public ESG reporting.

East West Bancorp also scored relatively low, potentially due to its limited disclosure levels compared to other companies. It had the highest variability, especially across E (95→44) and G (4→63).

Overall, scores with the provided data tended to be more moderate and clustered, showing strong alignment across models. In contrast, scores without provided data were more volatile, with wider spreads between attempts and LLMs. Scores from large-cap companies (Morgan Stanley, Goldman Sachs, Berkshire Hathaway) varied significantly, as all available data provided different implications. Mid-cap company East West Bancorp, however, displayed lower ESG scores across the board, due to limited disclosure levels.

By LLM (Trial 2)

Fig. 3. Average (mean) of the total ESG scores found across all attempts, separated into No-RAG (attempts 2 and 5) and RAG (attempts 1, 3, and 4). Error bars represent one standard deviation from the mean, indicating the variability of scores across attempts.

Claude-4 generally had the highest overall score for ESG overall. It was also one of the most consistent across attempts. Additionally, it recognized that different companies have different goals, so a smaller fintech company like GreenDot has less of a focus on ESG due to its size, and

when discussing ESG issues, it only discusses those that relate to its industry. Claude-4's methodology came closest to S&P's CSA scoring system, as both most heavily relied on disclosure, industry comparability, and weighted consistency. Like S&P, Claude-4 also gave large, well-disclosed companies high, stable scores, mirroring S&P's disclosure bias. Though companies possessing resources to issue long reports or participate in CSA are favored, Claude did not replicate other CSA-related distortions, like peer-relative scoring or the tendency for companies with close ties or investments in S&P to gain potential, indirect advantages. However, Claude often lacked S&P's explicit double materiality—its focus was on what the company disclosed and not on coordinating company impact and exposure—resulting in somewhat consistent outcomes at the surface level, overemphasizing disclosure sufficiency without looking deeper into financial or societal outcomes.

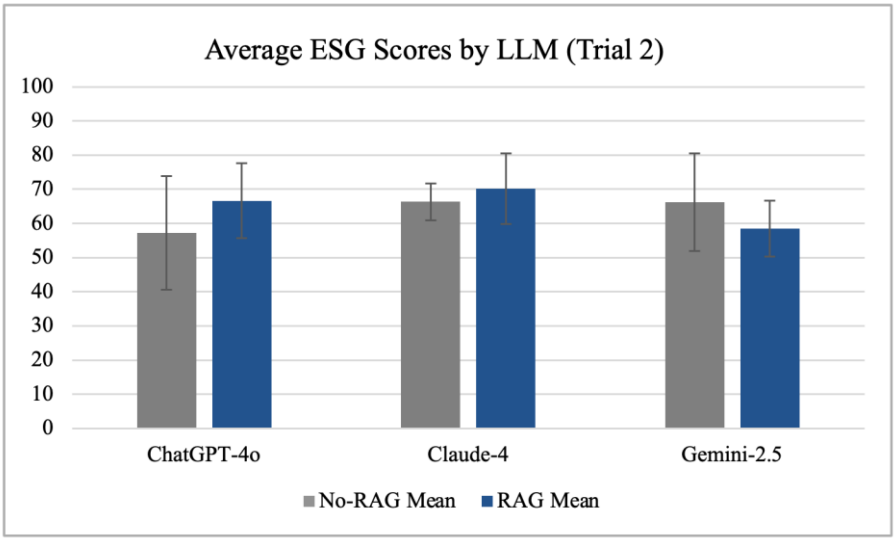

Figure 3: Average ESG Scores by LLM (Trial 2)

On the other hand, Gemini-2.5 had the most fluctuation and typically gave the lowest scores. Its deep research tool gave significant explanations per indicator, with substantially more context than the other LLMs. However, scoring practices were inconsistent across attempts, and sometimes, even across companies within the same attempt, leading to a large fluctuation in scores despite the same parameters and data provided. Gemini-2.5's approach was most similar to Sustainalytics, particularly for volatility and risk-based disclosure of ESG performance. Sustainalytics is built on exposure, management, and controversy scores, and therefore, results would be extremely sensitive to scandals or risks at industry levels, and Gemini mimicked this volatility through irregular scoring schemes across tries—occasionally purely indicator-based and occasionally grouped. This mirrors the structural biases within Sustainalytics, where risk-framed scoring gives higher scores to large firms with formalized ESG systems and penalizes smaller companies or those facing even minor controversies. Gemini's inconsistent weighting produced similar downshifts despite its otherwise strong ESG reporting. Ultimately, Gemini, like Sustainalytics, was excellent at contextualized commentary but demonstrated intermittent scoring that made results less predictable and less reproducible. ChatGPT-4o fell between Claude-4 and Gemini-2.5 for both the overall score and fluctuation, except for the governance score, which tended to be the highest and had the most fluctuation. ChatGPT-4o's disclosure-driven, conservative scoring approach also resulted in larger

URL: http://dx.doi.org/10.14738/assrj.1212.19750

companies with extensive reporting, such as BlackRock, PNC, or Bank of America, scoring higher, whereas smaller fintech competitors such as GreenDot were penalized heavily for limited disclosures, sometimes more than was proportional to their size or industry relevance. ChatGPT-4o scored closest to JUST Capital. Both used structured, indicator-based methodologies that were highly transparent but very disclosure-dependent and often over-penalized smaller firms. This pattern reflects the same modeling bias seen in JUST Capital's framework, where missing or incomplete data is filled with modeled estimates that can distort scores, particularly for small or mid-cap companies. ChatGPT consistently segmented scores equally by indicator based on explicit numeric values, similar to the application of public data, surveys, and modeled estimates by JUST in building rankings. The result, like JUST, was formal, disclosure-focused, and transparent, producing averages that clearly explained deductions but also highlighted the rigidity of the framework.

By Company (Trial 2)

BlackRock consistently achieved the highest scores across LLMs. Scores tended to be the most stable under Claude, with little variability across attempts. Though performance slightly declined in attempts without structured data, the overall consistency suggests transparent ESG reporting. PNC also demonstrated moderate to high scores consistently, particularly under ChatGPT and Claude. Gemini introduced greater fluctuation among E and S scores, though overall results remained balanced. PNC's performance was steadier when structured data was provided to the LLMs, indicating a dependence on clear disclosure for accurate results.

Bank of America exhibited considerable variation across models, with strong S and G scores but lower E ratings. Gemini showed the greatest inconsistency, while Claude's assessments were generally higher. Performance declined when structured data wasn't available to the LLMs, indicating that ESG consistency heavily relies on data accessibility.

American Financial Group received much lower scores across all models, with Gemini producing marginally higher results than ChatGPT. High variability, especially among G scores, suggests inconsistent recognition of governance-related information, likely due to limited public reporting because of company size.

GreenDot recorded the lowest and most inconsistent scores across models. Even between attempts, variability was high, demonstrating the influence of data availability and LLM retrieval limitations on ESG evaluations. GreenDot has the smallest market cap among these five companies, correlating with the lowest scores received.

Overall, structured input using a RAG approach produced more moderate and consistent scores across LLMs, while unstructured attempts led to volatility and disparities between models. Smaller companies, including American Financial Group and Green Dot, received lower and more fluctuating scores due to their lack of available ESG-related data. These results highlight the central role of data transparency in ensuring reliable ESG assessment using LLMs.

Differences and Similarities in Results

Indicators:

Across all models, environmental (E) indicators displayed moderate variability, while social (S) indicators exhibited the greatest variability, with large spikes and drops depending on the LLM.

Governance (G) remained the most stable category overall, with few outliers and consistent performance across attempts—especially when data was provided in attempts 1, 3, and 4.

Agencies:

The two analyses produced contrasting readings of how each LLM most closely mapped against established ESG rating agencies. In the first set, Claude was found to align most closely to Sustainalytics, using a risk-based scoring model that gave moderate base values in the presence of limited data. However, the second analysis concluded that Claude mapped most closely to S&P's Corporate Sustainability Assessment (CSA), where high significance was attached to disclosure depth, industry comparability, and consistency of indicators. Additionally, while the first analysis placed ChatGPT-4o with S&P, the second matched it with JUST Capital, citing its indicator-driven methodology and disclosure-based transparency. Finally, the two analyses connected Gemini based on systems aimed at stakeholder or risk sensitivity, even though the first compared it with JUST Capital for stakeholder engagement, while the second compared it to Sustainalytics for its risk and volatility-based variability. These variations emphasize that the methodological convergences of the LLMs—and thus their ESG scoring approach—vary depending on the criteria of interpretation and evaluation focus.

LLMs:

Each test produced a different depiction of the overall trends for each LLM. In the first set, Claude was found to consistently produce lower ESG scores across the board, with low fluctuation in scoring ranges. Conversely, in the second set, Claude assigned the highest overall ESG scores, while remaining the most consistent model. Moreover, while the first set determined that Gemini assigned the highest ESG scores across all companies, the second set determined the opposite, finding that Gemini consistently assigned the lowest scores. However, both sets noted that Gemini displayed inconsistent scoring practices across attempts, displaying the highest variability. Finally, both sets determined that ChatGPT-4o fell between Claude-4 and Gemini-2.5 for both overall score and fluctuation. Despite the select similarities, these major variations highlight that the difference in outcomes reflects the distinct methods of the LLMs, demonstrating that their approaches to ESG scoring are influenced by varying interpretive criteria and evaluation priorities.

CONCLUSION

This study evaluated how LLMs determine the E, S, and G scores of nine different, publicly traded companies of three different sizes (small, medium, and large) based on market capitalization. A total of 82 indicators, with 21 environmental, 30 social, and 18 governance, were used by the three LLMs (ChatGPT-4o, Claude-4, and Gemini-2.5) to determine the scores. The LLMs used a mix of SEC filings and company reports to evaluate the indicators. Across all of the companies, regardless of size, Claude-4 produced the most consistent results, with ChatGPT-4 coming close as well. Gemini-2.5, however, had the most fluctuation and inconsistency across companies and pillars. Large-cap companies generally had the highest scores among each pillar, and each LLM, with some exceptions like Berkshire Hathaway. Small and medium-cap companies had varying fluctuation of scores, primarily dependent on the availability of data and its comprehensiveness. Overall, these findings suggest that LLMs more consistently measure ESG performance when provided with structured data and using a RAG approach. Similarity across models suggests there is potential for AI-led ESG analysis to augment established scoring systems. However, their output remains heavily reliant on the

URL: http://dx.doi.org/10.14738/assrj.1212.19750

weighting of qualitative data in the model and the depth of available public information. Additional research needs to explore how alignment of LLM-based ESG assessments with current standards can be increased so that AI-powered assessments are transparent, reliable, and equitable with regard to company size and sector.

References

1. DiversIQ. "Goldman Sachs Company Profile | DiversIQ." *DiversIQ*, 13 Aug. 2024, diversiq.com/company-profiles/goldman-sachs/.
2. Garz, Hendrik, and Claudia Volk. *The ESG Risk Ratings*. Morningstar Sustainalytics, Oct. 2018.
3. Goldman Sachs. "Corporate Governance." *Goldmansachs.com*, 2024, www.goldmansachs.com/investor-relations/corporate-governance.
4. ---. "Goldman Sachs Privacy Policy." *Goldmansachs.com*, 28 Feb. 2024, www.goldmansachs.com/privacy-and-cookies/global-privacy-policy.
5. ---. "Security & Fraud Awareness." *Goldmansachs.com*, 2024, www.goldmansachs.com/security.
6. Goldman Sachs Global Stewardship Team, and Goldman Sachs Asset Management. *Our Approach to Stewardship*. 2021, www.gsam.com/content/dam/gsam/pdfs/common/en/public/miscellaneous/our-approach-to-stewardship.pdf?sa=n&rd=n.
7. JUST Capital. *JUST Capital Ranking Methodology*. Feb. 2025, com-justcapital-web-v2.s3.us-east-1.amazonaws.com/pdf/JUST_Capital_2025_Ranking_Methodology_250129.pdf.
8. McLaughlin, Tim, and Joshua Schneyer. "Buffett's Berkshire Hathaway Operates the Dirtiest Set of Coal-Fired Power Plants in the US." *Reuters*, 14 Jan. 2025, www.reuters.com/investigations/buffetts-berkshire-hathaway-operates-dirtiest-set-coal-fired-power-plants-us-2025-01-14/.
9. Morgan Stanley. *2023 ESG Report*. 2023.
10. ---. *Diversity and Inclusion Annual Report*. June 2022.
11. ---. *Morgan Stanley Code of Ethics and Business Conduct*. Apr. 2025.
12. ---. *State of the Workplace 2025 Financial Benefits Study*. June 2025.
13. Morningstar Sustainalytics. *ESG Risk Rating License*. 2024.
14. ---. *Preparation Guide: ESG Risk Rating Full Update*. July 2024.
15. S&P Global. *ESG Scores and Raw Data*. 2024, www.spglobal.com/esg/solutions/esg-scores-data.
16. Statista Research Department. "U.S.: Gender Diversity at Goldman Sachs by Job Type 2021." *Statista*, June 2024, www.statista.com/statistics/1317452/us-gender-diversity-goldman-sachs-by-job-category/.
17. United States Securities and Exchange Commission. "Berkshire Hathaway, Inc. SCHEDULE 14A INFORMATION. 2024." *Sec.gov*, 2023, www.sec.gov/Archives/edgar/data/1067983/000119312524069107/d512828ddef14a.htm.
18. ---. "East West Bancorp. Form 10-K. 2024." *Sec.gov*, 2024, www.sec.gov/Archives/edgar/data/1069157/000106915725000025/ewbc-20241231.htm.
19. ---. "The Goldman Sachs Group, Inc. Form 10-K/a 2023." *Sec.gov*, 2023, www.sec.gov/Archives/edgar/data/886982/000088698224000012/gs-20231231.htm.
20. Volk, Claudia, et al. *ESG Risk Ratings*. Morningstar Sustainalytics, 2024.