Investment Institute

WORKING PAPER 178 | OCTOBER 2025

Automating Insight Extraction from Oil And Gas Sector Climate Disclosures with AI

Amundi
Investment Solutions
Trust must be earned

Automating Insight Extraction from Oil And Gas Sector Climate Disclosures with AI

Sonja TILLY
Amundi Investment Institute
sonja.tilly@amundi.com

Aaron MCDOUGALL
Amundi UK
aaron.mcdougall@amundi.com

Tristan CHAILLOU
Amundi France
tristan.chaillou@amundi.com

Théo LE GUENEDAL
Amundi Technology
theo.leguenedal-ext@amundi.com

Sofia SAKOUT
Amundi Technology
sofia.sakout@amundi.com

Takaya SEKINE
Amundi Investment Institute
takaya.sekine@amundi.com

Abstract

Environmental, social, and governance (ESG) reporting has become a cornerstone of corporate transparency and accountability, especially within high emission sectors such as oil and gas. However, the traditional methods of extracting meaningful insights from ESG data are time-consuming and are in general processed manually. In this study, the authors introduce a Retrieval-Augmented Generation (RAG) pipeline, which automates the extraction and evaluation of information across large volumes of general and sustainable reporting, enabling analysts to efficiently process and synthesize data from multiple years and companies. The authors propose evaluation metrics that mimick human assessment. The methodology's scalability and adaptability make it a promising solution for automating the analysis of corporate ESG disclosures on a large scale, thus providing a robust framework for future research and practical applications in corporate sustainability assessment and climate engagement.

Keywords: Natural Language Processing (NLP), Large Language Models (LLM), RAG Pipeline, Agentic AI

JEL classification: C63, C88, O33, M14, Q56.

Acknowledgement

The authors are very grateful to Rami Mery and Tom Calamai from Amundi Technology, Jiali Xu from Amundi Investment Institute and Caroline Le Meaux and all participating members of the Amundi ESG Research & Engagement team for their helpful comments. The opinions expressed in this research are those of the authors and are not meant to represent the opinions or official positions of Amundi Asset Management.

About the authors

Sonja TILLY

Sonja Tilly, PhD, CFA joined Amundi as a Quantitative Analyst in 2024. Commencing her career in 2008 in London, Sonja started out as an Investment Analyst, then transitioning to Quantitative Analyst roles. Her time at Stanhope Capital, Aberdeen Asset Management, and Hiscox honed her skills in investment analysis, financial asset modelling, economic scenario development, and stress-testing portfolios.

While working as Quantitative Researcher at Qoniam Asset Management, Sonja created a filtering method based on deep learning for processing extensive news text, extracting signals that were transformed into a systematic equity trading strategy. Sonja's experience working in traditional Finance is complemented by her insights into the crypto space, gained as Quantitative Researcher at decentralised finance start-up Allora Network, where she led the creation of loan terms for fully automated NFT-backed on-chain loans.

Sonja holds a PhD in Computer Science from University College London. Her research focuses on the impact of news narrative on the economy and financial markets, blending methodologies from data science and econometrics. Further, she is a CFA Charterholder.

Aaron MCDOUGALL

Aaron is the Head of Climate in Amundi's Responsible Investment Research & Engagement division. His work includes developing and implementing policy which sets expectations of the climate performance and transition planning of corporate entities within Amundi's investment universe. To proactively address the dynamic nature of climate thematic investment, Aaron advises on pragmatic product innovation across Amundi's investment platforms. Externally, Aaron is a core member of the European Commission DG FISMA's Climate Benchmark developer community, contributed to the UK Transition Plan Taskforce, holds working group positions within the IIGCC and IEA, and leads CA100+ engagements on several major US firms. Aaron is an alumnus of Newcastle University, Strathclyde University, and University College Dublin, where he worked on the GreenWatch project with the Technical Expert Group of the European Commission's Platform on Sustainable Finance.

Tristan CHAILLOU

Tristan is an ESG analyst in Amundi's Responsible Investment Research & Engagement division, specialized on climate transition planning across major emitting sectors, notably energy and heavy industries. He has been developing for several years methodologies and tools in order to deepen, standardize and leverage the climate research and engagement efforts at Amundi. On top of supporting daily climate research and engagement efforts, Tristan is currently in charge of conducting the ESG research and engagement for the building products, construction materials, and construction & engineering sectors. Externally, Tristan worked to establish common guidance on corporate transition planning, such as with the French Association of Asset Managers (AFG), and contributes to several CA 100+ engagements. Tristan studied at Bordeaux and Paris-Dauphine universities, holding master degrees in International Economics and Finance, Asset Management, and Mathematics and Data Science applied to Financial Markets.

Théo LE GUENEDAL

Théo Le Guenedal is the Head of Prospective and Quantitative Solutions at the Innovation Lab of Amundi Technology. Prior to this, he worked in the Quantitative Research department of the Amundi Institute since 2018, starting with a project on the performance of ESG investing in the equity market. Since then, he has been involved in an extensive research project on incorporating ESG factors, alternative signals and climate risks into asset allocation strategies. In 2020, he co-authored a paper titled "Credit Risk Sensitivity to Carbon Price," which was awarded the GRASFI Best Paper Prize for Research on Climate Finance, a prestigious honor sponsored by Imperial College London. He also made significant contributions to the academic field of physical risk assessment by developing the Tropical Cyclone Generation Algorithm. Théo completed his Ph.D. thesis, "Financial Modeling of Climate-related Risks" in Applied Mathematics at the Institut Polytechnique in December 2023, covering both transition and physical risks. Recently, he has focused on integrating advanced climate metrics, stress tests, and analytics into investment tools at Amundi Technology's Innovation Lab.

Sofia SAKOUT

Sofia Sakout, PhD is Lead Data Scientist in Natural Language Processing and Generative AI within the Innovation Lab of Amundi Technology. Sofia plays a central role in harnessing open source text data and in the development of NLP pipelines, with a particular focus on question-answering (QA) systems. Her focus at Amundi Technology is to channel innovation, data science, artificial intelligence toward practical applications in Amundi.

Sofia holds an engineering degree from École Mohammadia d'Ingénieurs, a Master's degree from École des Ponts ParisTech, and a PhD from the Sciences, Ingénierie et Environnement Doctoral School in partnership with École Polytechnique.

Takaya SEKINE

Takaya Sekine, CFA is the Deputy Head of Quant Portfolio Strategy within Amundi Investment Institute (formerly known as the Quantitative Research Team of Amundi). In this role, he works on the practical implementation of quant research, artificial intelligence and alternative data for investment strategies.

He joined Amundi in 2000 and is in his current position since July 2018. Prior to that, he was Deputy CIO at Amundi Japan (between 2011 and 2018) with a focus on global quantitative strategies, Head of Index and Multi-Strategies at Amundi Japan (between 2010 and 2011), Fund Manager (between 2007 and 2010) and Financial Engineer (between 2001 and 2007). He has been involved in macro and policy related investment strategies for both retail and institutional clients. Takaya began his career as an IT Manager at Amundi Japan's predecessor company (between 2000 and 2001).

Takaya is a CFA charterholder since 2005. He received the Ingénieur Civil des Mines degree from Ecole des Mines de Nancy in 2000.

Key takeaways

- Automated Corporate Information Extraction: The study introduces a RAG pipeline that automates the extraction and evaluation of ESG-focused information, reducing manual effort.
- Scalable and Accurate Tool: The NLP pipeline combines retrieval and generation techniques to offer a scalable and accurate solution for ESG assessments.
- Greenwashing Detection: The framework can aid in identifying greenwashing by cross-referencing corporate claims with verified data, ensuring more transparent and truthful ESG reporting.

1 Introduction

In recent years, sustainable reporting has become a cornerstone of corporate transparency and accountability, especially within high emission sectors such as oil and gas. The growing urgency surrounding climate change has pushed both companies and financial analysts in Europe and Asia to focus on how firms engage with sustainability initiatives, climate risk, and long-term environmental stewardship. As a result, analysts, investors, and stakeholders have developed frameworks to assess the authenticity and quality of ESG disclosures, often seeking to gauge a company's alignment with climate targets, its strategies for mitigating environmental risks, and its overall contributions to sustainability goals (Eccles *et al.*, 2014). Moreover, the climate investment trajectories between major regions have diverged noticeably in recent years, particularly under the influence of shifting US policy. In Europe and Asia, investor commitment to climate-aligned investments remains strong and bolstered by stable policy frameworks and regulatory support. For instance, in 2024, approximately 62% of European and 59% of Asia-Pacific investors continued to regard climate change as central or significant to their investment strategies, far above the mere 23% in North America, where uncertainty under the Trump administration has dampened climate focus (Week, 2025). In North America, volatility in US climate policy, amplified by Trump's return to power and his withdrawal from international agreements has triggered a noticeable pullback from climate investing. As an illustration, a 2025 Robeco survey shows that US holdings in global climate-themed funds dropped from 69% to 61%, with a corresponding shift of capital toward Europe (increase from 19% to 25%) and Asia-Pacific (rising from 9% to 13%) (Robeco, 2025).

However, despite the widespread adoption of ESG reporting frameworks, significant concerns remain regarding the quality and consistency of these reports. Companies vary widely in how they disclose information, with some providing detailed, verifiable data on climate-related risks and others offering vague or incomplete accounts of their environmental impacts (Grewal *et al.*, 2021). The lack of standardized reporting frameworks has made it difficult for analysts to consistently evaluate and compare companies' climate engagements, particularly when the reported data may lack rigor or sufficient detail (Gutman *et al.*, 2024; Whelan *et al.*, 2021). Consequently, these inconsistencies have made it challenging for investors to make informed decisions based on climate engagement (CFA Institute, 2024).

This variability in sustainability reporting has given rise to a growing concern: greenwashing. Greenwashing refers to the practice of misleading stakeholders about a company's environmental efforts or achievements, often by overstating the positive impacts of sustainability initiatives or selectively reporting information that presents the company in an overly favorable light (Lyon & Montgomery, 2015). The phenomenon has garnered increasing attention in both academic literature and regulatory discourse, as greenwashing distorts the true environmental performance of companies, undermines investor confidence, and potentially slows progress toward global climate goals (Calamai *et al.*, 2025; Cartellier *et al.*, 2023; Delmas & Burbano, 2011).

To address these challenges, analysts and researchers have developed a range of methodologies aimed at systematically evaluating the integrity and validity of sustainability disclosures. One such methodology, central to this paper, employs a Retriever-Augmenter-Generator (RAG) pipeline, which automates the extraction of relevant information from large volumes of corporate ESG reports (Lewis *et al.*, 2020; Stagnol *et al.*, 2023). This automation allows for the efficient processing of textual data that would otherwise require manual analysis, eliminating human processing of large volumes of data, reducing inconsistencies, and ensuring more robust assessment across multiple years and companies. By replacing previously time-consuming manual tasks, RAG enhances the accuracy of evaluations, prevents over-

sights, and makes the engagement process more consistent and scalable.
Furthermore, this same methodology provides an additional layer of insight by enabling a comprehensive assessment of the overall quality of sustainability reporting, particularly in terms of identifying potential greenwashing. This quality assessment allows analysts to not only check whether companies meet specific engagement criteria, but also to evaluate whether the disclosures are transparent, and free from misleading claims. By applying the RAG pipeline to five years' worth of ESG reports from 50 major oil and gas companies, this paper demonstrates how such automated systems can be employed to detect greenwashing candidates, offering a robust and scalable framework for evaluating corporate climate engagement.
This research contributes to the literature by advancing methods for automating the extraction of information from large-scale reporting, a task that has traditionally been labor-intensive. While previous studies (e.g. by Folke and Ivan Erik Troedsson (2025) and Ontiveros *et al.* (2025)), explore the use of large language model (LLM)-based question-answering pipelines to address ESG-related questions, these approaches often focus on a small set of specific queries or predefined criteria, limiting their scalability and adaptability to diverse and evolving ESG disclosures. In contrast, this study introduces the Retriever-Augmenter-Generator (RAG) pipeline, which automates the extraction and evaluation of a broader range of ESG information across large datasets, enabling analysts to efficiently process and synthesize data from multiple years and companies. By combining retrieval, augmentation, and generation techniques, the RAG framework enhances both the accuracy and consistency of ESG content analysis. Moreover, it integrates a quality assessment layer that evaluates the extracted ESG content for completeness, consistency, and potential misleading claims, thus providing a more comprehensive evaluation tool. While this evaluation also contributes to detecting greenwashing, its primary focus is to extract information from a large volume of corporate general and sustainable reporting. In doing so, the study addresses a gap in current sustainability evaluation practices by combining state-of-the-art information extraction with quality assessment, offering a more scalable and accurate tool for reporting analysis.

Structure of the paper

This paper is structured as follows. Section 2 reviews related works. Section 3 sets out the methodology. Section 4 presents the results, applying the methodology to general and sustainable reporting issued by 50 oil and gas companies. Section 5 extends this research by applying the methodology to identify potential greenwashing. Section 6 discusses the overall findings. Finally, section 7 concludes the paper.

2 Related Works

The Retrieval-Augmented Generation (RAG) architecture has emerged as a promising solution for question answering (QA) tasks, particularly by enhancing the abilities of large language models (LLMs) through the integration of external knowledge (Lewis *et al.*, 2020). RAG combines two key components: a retrieval mechanism and a generative model, which works together to improve the accuracy and relevance of responses.

2.1 RAG Architecture for Question Answering

RAG pipelines have seen widespread application across various sectors, with numerous studies highlighting their domain-specific benefits. Mao *et al.* (2024) showcases how RAG frameworks can enhance the quality of text generation by feeding external knowledge into the LLM, which is particularly beneficial for domain adaptation. The integration of domain-specific knowledge improves the relevance and accuracy of generated answers, proving the adaptability of RAG systems in specialized fields. Alan *et al.* (2025) supports this idea, exploring the application of RAG architectures in question answering, applying it to the religious domain. The authors emphasize the potential of RAG to enhance question answering by retrieving pertinent external knowledge, thus providing a richer context for the LLM to generate more precise answers. Balaguer *et al.* (2024) focuses on the agricultural sector, where RAG systems are used in conjunction with LLMs fine-tuning with specific agricultural knowledge. This application showcases how RAG can be used to tailor models for specialized industries, improving the precision of the LLM's responses in these contexts. Islam *et al.* (2024) extends the use of RAG by applying an open RAG framework, which significantly enhances the reasoning capabilities of LLMs. OpenRAG transforms a dense openâsource LLM into a parameter, efficient sparse mixtureâofâexperts (MoE) model, enabling dynamic expert selection and adaptive retrieval to significantly enhance reasoning capabilities, especially for complex and multiâhop tasks within a Retrieval Augmented Generation architecture. The authors' use of a parameter-efficient sparse mixture of experts model allows for a more scalable and efficient implementation of RAG. Jeong *et al.* (2024) propose an adaptive retrieval-augmented generation (RAG) framework that dynamically selects the appropriate retrieval strategy based on query complexity. By training a lightweight classifier to predict whether a query requires no retrieval, single-step retrieval, or iterative retrieval, their method optimizes both accuracy and efficiency across a range of question types.

2.2 Evaluation of RAG Pipelines

Evaluating Retrieval-Augmented Generation (RAG) systems requires multidimensional assessment across factuality, relevance, groundedness, and fluency. QAFactEval introduces a QA-specific factuality evaluation method using large language models (LLMs) to determine whether a generated answer accurately responds to the input question, offering a fine-grained alternative to traditional fact-checking (Feng *et al.*, 2023). Similarly, RAGAs proposes a framework for evaluating groundedness in RAG pipelines, leveraging both human annotations and automated metrics to assess whether responses are supported by retrieved context (Es *et al.*, 2024). From an information retrieval perspective, the BEIR benchmark remains a cornerstone for evaluating document relevance, employing dense vector similarity (e.g., cosine similarity) to quantify how well retrieved documents align with the query (Thakur *et al.*, 2021). Cherief *et al.* (2025) measure semantic alignment by calculating the cosine similarity between a vector representation of an embedded definitions of a concept and clusters of terms. Beyond factuality and relevance, FEQA emphasizes the importance of fluency and conciseness in QA outputs, recognizing that high-quality responses must also be coherent and easy to interpret (Durmus *et al.*, 2020). Together, these frameworks underscore the need for comprehensive evaluation methodologies that go beyond accuracy to incorporate contextual alignment, retrieval quality, and linguistic clarity in RAG systems. Krishna *et al.* (2024) address the need for comprehensive evaluation of retrieval-augmented generation (RAG) systems by introducing FRAMES (Factuality, Retrieval, And reasoning MEasurement Set), a dataset designed to assess LLMs' factual accuracy, retrieval effectiveness, and reasoning in an end-to-end manner. Unlike prior evaluations that consider these abilities separately, FRAMES features challenging multi-hop questions requiring synthesis from multiple sources. Baseline results reveal that even state-of-the-art LLMs perform modestly without retrieval (0.41 accuracy), while a multi-step retrieval pipeline boosts accuracy significantly to 0.66.

2.3 RAG Pipelines to Reduce Hallucinations

Hallucinations, the generation of incorrect or fabricated information by LLMs, remain a significant challenge in natural language processing (NLP), especially when these models are tasked with answering questions in knowledge-intensive domains. The integration of a retrieval phase in RAG systems has shown to mitigate such issues by grounding the generation process in real, verifiable information.
Niu *et al.* (2023) propose the RAGTruth corpus, which addresses the need to identify LLM hallucinations by providing nearly 18,000 manually annotated responses from diverse LLMs, with detailed word-level hallucination intensity labels across multiple domains. This dataset enables benchmarking of hallucination frequencies and evaluation of detection methods. The authors find that fine-tuning a smaller LLM on RAGTruth achieves competitive hallucination detection compared to prompt-based methods using frontier class models like GPT-5, and effectively mitigates hallucinations in LLM outputs. Veturi *et al.* (2024) discusses how the retrieval mechanism plays a crucial role in minimizing hallucinations by providing concrete information from trusted sources, thus reducing the model's reliance on potentially fabricated internal representations. Song *et al.* (2024) introduce a hallucination-aware tuning method for RAG, training hallucination detection models that generate detection labels and describe the detected hallucinations, resulting in improved response quality. Furthermore, Ran *et al.* (2025) propose the use of a "re-ranking" step in RAG systems to further reduce hallucinations. Additionally, Taguchi *et al.* (2025) explores adaptive k-retrieval, a method that dynamically adjusts the number of retrieved passages based on the distribution of similarity scores between the query and the retrieved text. This technique enhances the retrieval process, ensuring that the LLM is provided with the most relevant and high-quality information for answering questions, which can be directly leveraged through efficient prompting to improve response quality. Joshi (2025) provides a systematic review of technologies to detect and mitigate hallucinations in large language models (LLMs). The study categorizes methods into detection-, prevention-, and correction-based approaches, evaluating them on accuracy, latency, and complexity. The author finds that hybrid retrieval-augmented generation (RAG) reduces errors by 35-60%, while neurosymbolic techniques like automated reasoning and multi-agent validation perform well in high-stakes settings.

2.4 Efficient Prompting as Part of a RAG Pipeline

Efficient prompting within a RAG pipeline is crucial for ensuring that the system's performance is optimized. Prompting directly influences how well the LLM utilizes the retrieved information to generate accurate and relevant responses. Several studies have examined ways to enhance the prompting mechanisms to improve the efficiency and effectiveness of RAG pipelines. Roychowdhury *et al.* (2024) presents a framework for evaluating RAG systems, emphasizing the importance of designing effective prompts that guide the LLM towards generating accurate answers. The study suggests that prompt refinement is an essential part of optimizing the retrieval process and ensuring that the LLM can effectively process the retrieved information to generate coherent responses. Recent work by Chua (2025) introduces the COSTAR prompt framework, a structured approach to prompt design aimed at improving the consistency and effectiveness of interactions with large language models. This framework emphasizes key elements such as Context, Objective, Style, Tone, Audience, and Response format, offering a practical methodology for prompt engineering in real-world applications. Merth *et al.* (2024) propose superposition prompting, a novel method that enables parallel processing of input documents through multiple prompt paths, discarding irrelevant ones without requiring fine tuning. This approach significantly improves both efficiency and accuracy in RAG tasks, achieving up to a 93 times reduction in compute time and a 43% accuracy gain on benchmarks like NaturalQuestions Open. As LLMs gain the ability to process longer input sequences, RAG can incorporate more retrieved information to improve outputs. Jin *et al.* (2024) find that beyond a certain point, adding more passages harms quality due to irrelevant facts retrieved. To address this, the authors explore training-free retrieval reordering and fine-tuning methods with intermediate reasoning to enhance robustness in long-context RAG. The study concludes that key factors like data distribution, retriever choice, and context length are critical for optimizing performance with longer inputs.

2.5 Greenwashing

This research uses climate claims extracted from corporate general and sustainable reporting.
Climate claims refer to statements made by organizations regarding their efforts to mitigate climate change, for instance by reducing carbon emissions or adopting environmentally friendly practices (Stammbach *et al.*, 2022). When such claims are exaggerated, vague or unsupported, this can result in greenwashing, which is a deceptive practice where businesses present an environmentally friendly image without following through on their claims (Bernini *et al.*, 2023; Dumitrescu *et al.*, 2022; Yang *et al.*, 2020). The validation of climate claims is crucial to holding corporate entities accountable, ensuring transparency and fostering genuine progress in the fight against climate change.
Approaches to validating climate claims proposed in existing literature primarily focus on the semantic analysis of publicly available corporate documents such as annual reports and other filings. Calamai *et al.* (2025) review how natural language processing is being used to detect corporate greenwashing, focusing on techniques for identifying misleading environmental claims in text. They analyze existing datasets, models, and detection tasks, while highlighting current limitations and suggesting paths for future research. Ghitti *et al* identify greenwashing claims by contrasting ex-ante intentions with ex-post outcomes on environmental policies, also taking into account the dispersion of ESG ratings across providers and actual violations committed by a firm (Ghitti *et al.*, 2024). He *et al* measure the distance between a firm's green talk through call transcripts and specific environmental performance and initiatives through corporate reporting (He *et al.*, 2024). Blanco *et al* leverage S&P500 firms' sustainability disclosures to extract features and to predict greenwashing (Ruiz-Blanco *et al.*, 2022).

3 Methodology

We implement a Retrieval-Augmented Generation (RAG) pipeline to systematically analyze environmental claims within 6,448 reports from 50 oil and gas sector companies, spanning the period from 2019 to 2024.
Amundi's climate ambitions [1] translate into a detailed Climate Engagement Framework. A Climate Engagement Framework is a structured set of criteria and guidelines used to evaluate the climate - and sustainability-related aspects of a company's operations and disclosures. It serves as a tool for systematically assessing whether companies are aligning with best practices in sustainability and climate responsibility. This assessment is typically conducted

[1] https://about.amundi.com/article/engagement-report-2024; accessed on 27/08/2025.

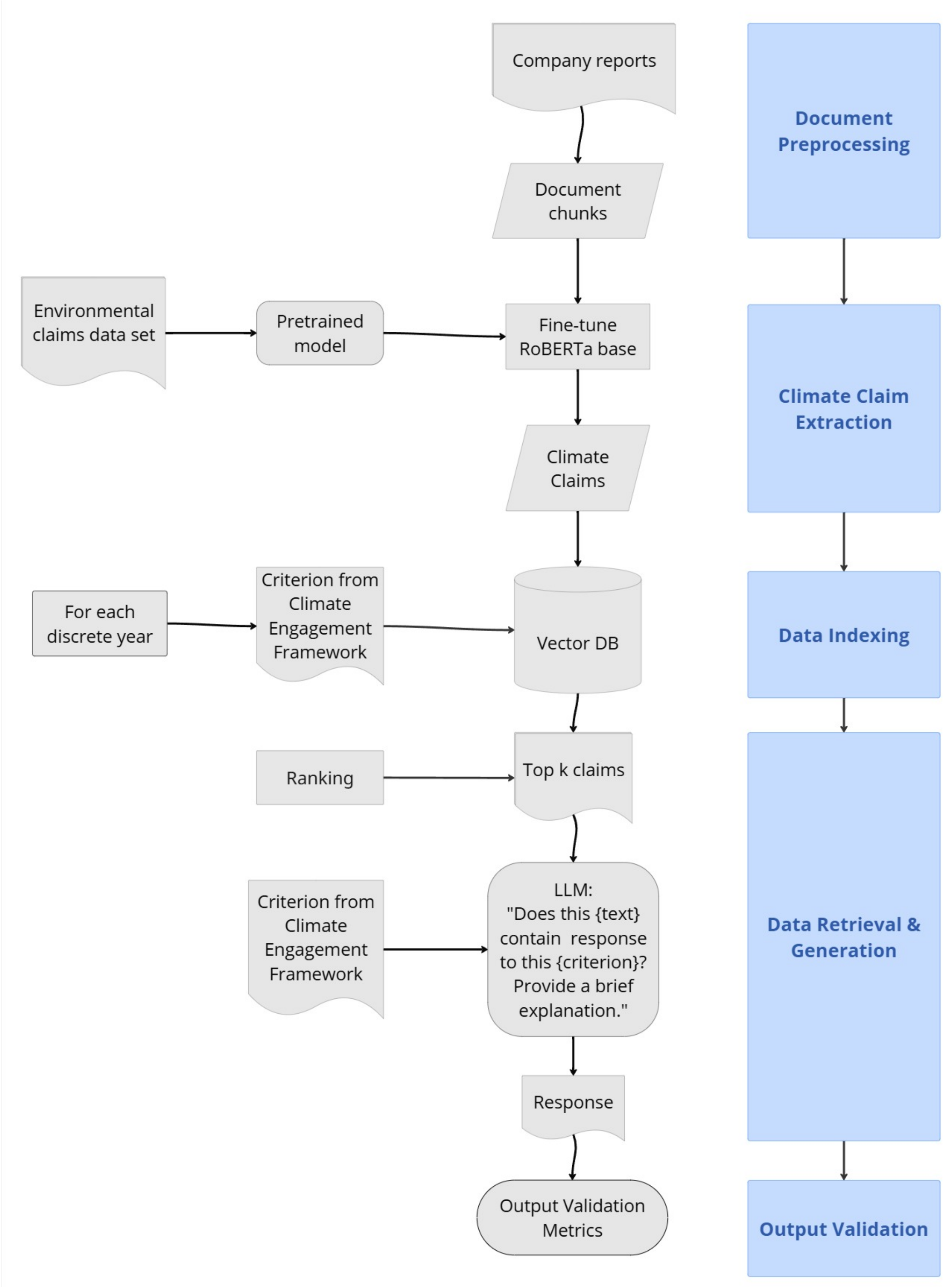
on a periodic basis and includes key indicators such as the presence of emission reduction targets, transparency in climate-related reporting, and whether executive remuneration is linked to climate objectives.

When a company does not meet specific criteria outlined in the framework, these gaps are flagged. In the case of institutional investors, the results of this evaluation inform engagement strategies: the investor actively reaches out to companies to encourage alignment with the framework's expectations. This engagement process aims to drive improvements in corporate climate governance, disclosure, and performance through direct dialogue and shareholder influence.

Algorithm 1 RAG Pipeline with Automated Evaluation for ESG Report Analysis

Require: ESG reports in PDF format for 50 oil and gas companies over the past 5 years
Ensure: Validated climate-related assessments per Climate Engagement Framework criterion
Document Preprocessing:
for each ESG report **do**
    Extract text from PDF
    Chunk document into semantically coherent sections
end for
Climate Claim Extraction:
Load Environmental Claims Dataset (ECD)
Fine-tune `RoBERTa_base` model on ECD with class weights
for each text chunk **do**
    Use the fine-tuned model to identify and extract climate-related claims
end for
Data Indexing:
for each year and each criterion in the Climate Engagement Framework **do**
    Embed climate claims using a sentence embedding model
    Store in a vector database indexed by year and criterion
end for
Data Retrieval and Generation:
for each criterion **do**
    Rank documents by relevance to criterion
    Retrieve top 25% most relevant claims from the vector store
    Prompt the LLM with retrieved claims to assess if the criterion (including guidance) is addressed
    Generate a structured response and a brief explanation
end for
Automated Evaluation:
for each generated output **do**
    Compute **Answer Correctness** (1–5): LLM rates how well the answer addresses the query
    Compute **Faithfulness** (1–5): LLM evaluates if the answer is fully grounded in retrieved context
    Compute **Conciseness** (1–5): LLM assesses if the answer is direct and focused
    Store all scores and brief justifications
end for
return Validated assessments and evaluation scores

Figure 1: Methodology overview



Source: Amundi Investment Institute.

3.1 Document Preprocessing

We collect 6,448 reports published by 50 oil and gas companies from 2019 to 2024. Each report is preprocessed through a pipeline that included text extraction and chunking. The documents are segmented into semantically coherent text chunks of 300 words suitable for input into NLP models, ensuring both granularity and contextual integrity for downstream tasks.

3.2 Climate Claim Extraction

To identify and extract climate-related claims, we a data set created by Stammbach *et al.* (2022), which contains climate discourse in corporate texts. We then fine-tune $RoBERTa_{base}$ on this dataset to strengthen domain-specific accuracy. We then apply the fine-tuned model to the chunked ESG documents in order to extract discrete climate-related claims.
We fine-tune the $RoBERTa_{base}$ model (Liu *et al.*, 2019) following the classification task outlined by Stammbach *et al.* (2022), using the Environmental Claims Dataset (ECD)[2]. While Stammbach *et al.* (2022) employ a $DistilBERT_{base}$ variant (Sanh *et al.*, 2019), we opt for $RoBERTa_{base}$ (Church *et al.*, 2021) and account for class imbalance - 1982 True vs. 665 False labels prior to train-test split by incorporating class weights, as recommended by Rajaraman *et al.* (2022). Our grid search explores two learning rates and varied epoch counts, with ten runs per configuration. Based on the tighter variance in evaluation accuracy, we select the model trained with a learning rate of $3 \cdot 10^{-5}$ for its consistent performance across epochs.

3.3 Data Indexing, Retrieval and Generation

For retrieval-augmented generation (RAG), we rely on the LangChain toolkit, an open-source framework that facilitates the orchestration of large language models (LLMs) with external tools such as retrievers, vector databases, and memory components (LangChain, 2025). LangChain allows to construct a modular pipeline for retrieval and reasoning over sustainability content.

On a company by company basis, we construct a vector database for each calendar year represented in the dataset and for each criterion in the Climate Engagement Framework.
Embedding vectors are stored using a lightweight, RAM-based storage mechanism suited to medium-sized corpora where low-latency access to documents is essential. Documents are retrieved using an information retrieval algorithm that uses BM25 ranking to return documents based on token-level similarity scores. BM25 is used as it gives higher weight to documents that contain the exact query terms, especially domain-specific keywords (Crestani *et al.*, 1998) (see appendix A.2 for details).
Using the vector database, we perform semantic search to retrieve the most relevant climate claims for each engagement criterion. Following empirical evaluation, we heuristically select the top 25% of retrieved claims as inputs for response generation. For each year and each criterion, an LLM is prompted to determine whether the retrieved claims addressed the criterion. This structure permits to perform targeted, criterion-specific information retrieval.

[2] https://github.com/dominiksinsaarland/environmental_claims, accessed June 7th, 2024

3.4 Response Generation

We use ChatGPT 4.0 Turbo, a high performance variant of OpenAI's GPT 4 series, for language generation tasks (OpenAI, 2024). The temperature is set at 0.7, which introduces a moderate level of randomness into the model's output, balancing creativity and determinism in the generation process. In addition, we leverage the the COSTAR prompt framework for prompt design (Chua, 2025). The COSTAR framework includes the following key components:

- **Context:** Provides the relevant background information.
- **Objective:** Clearly defines the goal or the problem to be addressed.
- **Strategy:** Specifies how to approach the task or problem.
- **Tactics:** Specific actions taken to address the strategy.
- **Action:** Application of the strategy to solve the problem.
- **Result:** The response.

This structure ensures that the model is guided through a well-defined process to answer complex questions. This is how the prompt looks like:

Prompt Structure

Rate from 1 to 5: Does this context {context} contain an answer to this question: {query}? Provide a brief factual explanation for ESG analysts.

For each year, we prompt the LLM using each engagement criterion alongside the corresponding analyst guidance (see Appendix A.3 for an example), requesting a brief explanation to determine whether the retrieved context adequately addresses the criterion.
The retrieved climate claims represents the context and the query includes a criteria from the climate engagement framework and its corresponding analyst guidance in question format.

3.5 Performance Evaluation

To assess the performance of the language model outputs, we compare the model-generated scores against expert human annotations (see appendix 5 for annotation guidelines), which serve as the ground truth. These comparisons are made at the level of ordinal categorical ratings on a 5-point scale (1 to 5), covering aspects such as correctness, faithfulness, and conciseness.
The following performance evaluation metrics are designed to emulate human assessment, particularly in cases of ambiguity, where most responses are inherently incomplete due to the limitations of the retrieved data. As the retrieved context is often partial or fragmented, the evaluation framework accounts for these gaps, offering nuanced scores and natural language explanations that reflect the challenges of assessing incomplete information. This approach ensures that evaluations are meaningful even when the data does not fully support a comprehensive response. The prompt structure for each of the following LLM-based metrics can be found in section A.5.
We employ a combination of quantitative metrics that are well-suited for ordinal and rating-based data. The Quadratic Weighted Kappa (QWK) is appropriate for ordinal scales, where

the distance between ratings carries semantic meaning (Cohen, 1968). The QWK penalizes larger disagreements more heavily than smaller ones.

QWK is given by

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{1}$$

where the quadratic weights are defined as follows:

$$w_{ij} = \frac{(i-j)^2}{(k-1)^2} \tag{2}$$

O_{ij} is the observed agreement matrix (confusion matrix), E_{ij} is the expected agreement matrix (based on the assumption of independence) and k stands for the number of score levels.

The Mean Absolute Error measures the average absolute difference between predicted and true scores and is given by

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

where y_i is the human (ground truth) score, $\hat{y}_i$ is the model-predicted score and n is the total number of samples.

To provide a more interpretable, visual summary of model performance, we generate confusion matrices comparing the distribution of scores assigned by the LLM and human annotators:

$$C_{ij} = \text{Number of times a true score of } i \text{ was predicted as } j \tag{4}$$

4 Findings

Table 1: Performance metrics comparing LLM scores against human ground truth across evaluation dimensions.

Score Pair	**QWK**	**MAE**
Correctness	0.7589	0.2329
Faithfulness	0.8011	0.0823
Conciseness	0.7578	0.1118

The metrics shown in Table 1 are based on a subset of criteria and guidance drawn from the Climate Engagement Framework. These scores reflect performance on selected evaluation dimensions that were applied consistently across years. This subset was chosen to reduce the number of calls to the LLM and to manage overall computational time, while still ensuring a representative and meaningful assessment. The evaluation metrics are detailed in Section 3.5, and the specific criteria used in this analysis are listed in Appendix A.3.

Table 1 summarizes the performance of the language model across three evaluation criteria: Correctness, Faithfulness, and Conciseness. We report two key metrics: Quadratic Weighted Kappa (QWK) and Mean Absolute Error (MAE), both of which assess agreement between the model's predictions and the human ground truth. QWK values range from 0 to 1, where higher values indicate stronger agreement, with scores above 0.75 demonstrating substantial concordance. The model achieves consistently high QWK scores across all categories, with Faithfulness showing the highest agreement (0.80), suggesting the model is particularly effective at capturing the factual alignment of its outputs. The MAE values, which measure the average magnitude of errors without regard to direction, further illustrate the model's prediction accuracy. However, the relatively higher MAE observed in correctness (0.23) compared to the other categories may indicate greater variability or dispersion in the model's predictions for this aspect, suggesting that while generally accurate, the model's correctness scores exhibit slightly more deviation from human judgments.
Together, these metrics highlight the methodology's robust performance in approximating human judgments across multiple dimensions of evaluation. Although annotations produced by LLMs may differ from human annotations, this does not inherently imply they are less accurate or reliable. In this task, as with much real-world data, a definitive ground truth is not available. Therefore, the focus of annotation is less on absolute accuracy and more on providing informed and meaningful interpretation (Kaikaus *et al.*, 2023).
The confusion matrices shown in figure 2 reveal that correctness is the category with the greatest disagreement between LLM predictions and human ratings. This is illustrated by a higher number of off-diagonal values, indicating mismatched scores. In particular, the LLM often assigns lower scores than humans for correctness as illustrated in figure 3, where the average human ratings for correctness are consistently higher than those of the LLM. The greater dispersion suggests that correctness may be more subjective or harder for the model to assess reliably, assigning more conservative scores than the human annotators, compared to faithfulness and conciseness, where the agreement is stronger.

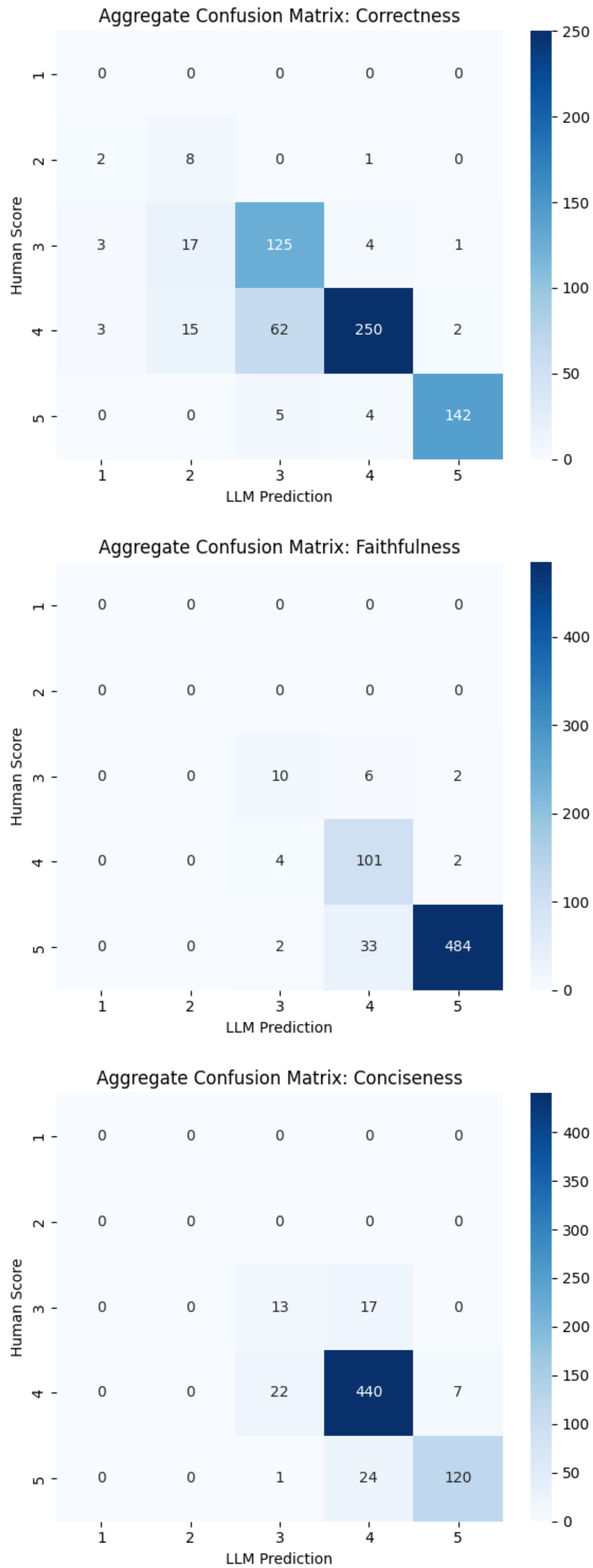


Figure 2: Confusion Matrices by Score

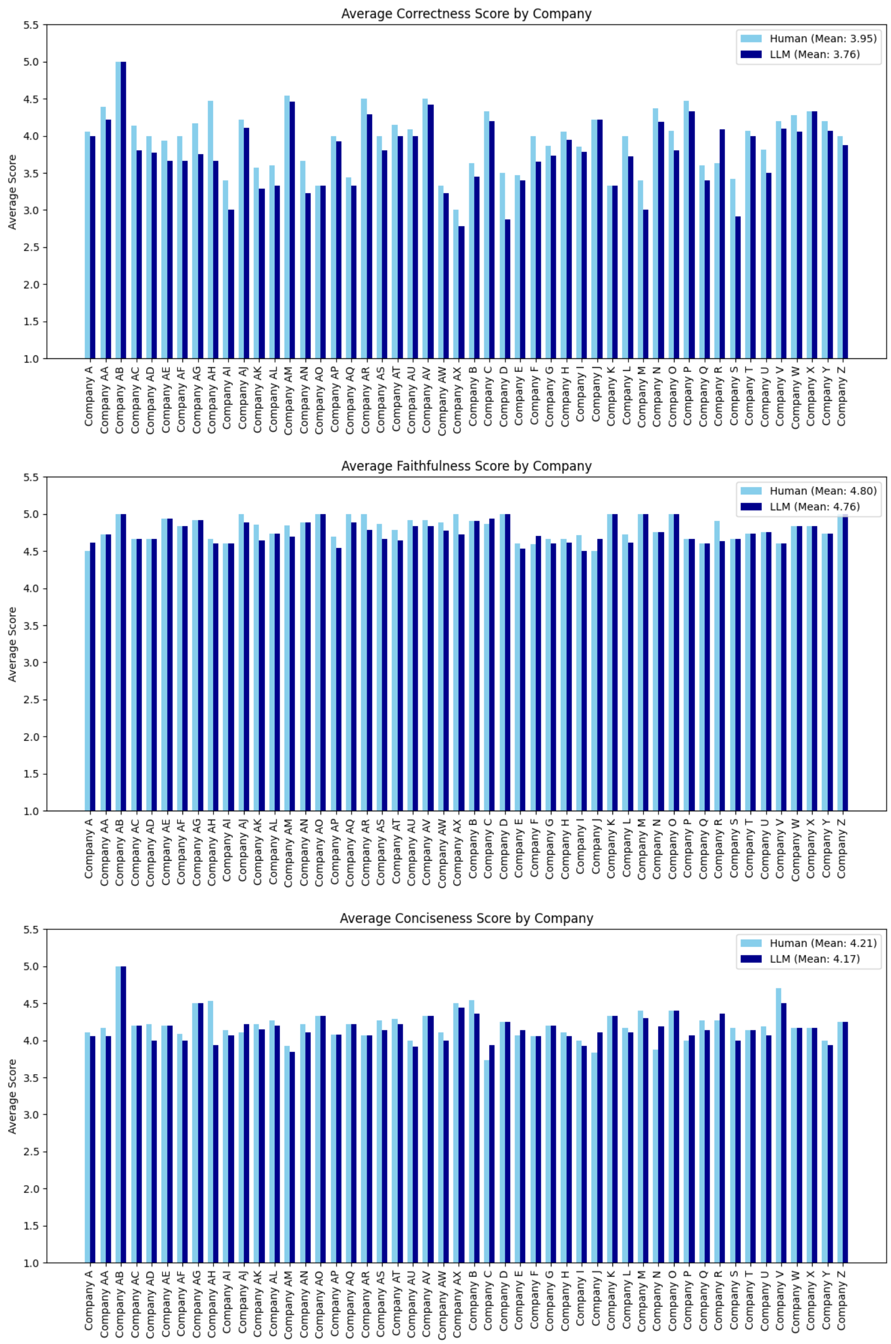


Figure 3: Average Scores by Company

The histograms are based on a subset of criteria and guidance drawn from the Climate Engagement Framework. These scores reflect performance on selected evaluation dimensions, applied consistently across years. The evaluation metrics are detailed in Section 3.5, while the specific subset of criteria used for this analysis is listed in Appendix A.3.

5 Identification of greenwashing candidates

In addition to finding answers to the framework criteria, it is important to note that the awareness and identification of potential greenwashing remain an ongoing concern in the evolving landscape of corporate sustainability claims. As companies continue to refine their sustainability messaging, the ability to detect subtle or deceptive practices becomes even more critical for ensuring the integrity of ESG disclosures.

In section 2.5, we discuss research that attempts to define greenwashing. Further, we test the LLM on its understanding of greenwashing (see appendix A.6). The LLM's explanation of greenwashing is in line with definitions by Bernini *et al.* (2023) and de Freitas Netto *et al.* (2020).

To evaluate the ability of large language models (LLMs) to detect potential greenwashing in corporate climate-related claims, we manually annotate a total of 1,000 statements. Each statement was rated on a 5-point scale, where 1 indicates no greenwashing and 5 indicates definite greenwashing. We then tasked the LLM with generating comparable greenwashing scores for each claim.

As with much real-world data, a definitive ground truth is not available in this task. Therefore, the focus of annotation is less on absolute accuracy and more on providing informed and meaningful interpretation of a climate claim in a given context (Kaikaus *et al.*, 2023).

Table 2: Performance metrics for LLM-generated greenwashing ratings

Metric	**Score**
Quadratic Weighted Kappa (QWK)	0.9009
Mean Absolute Error (MAE)	0.1410

The results indicate strong agreement between LLM-generated ratings and human annotations, as evidenced by a high QWK score of 0.90 and a very low MAE of 0.14. These findings suggest that the LLM is capable of capturing nuanced patterns in greenwashing language.

Interestingly, closer inspection of the confusion matrix (Figure 4) reveals that the LLM does not assign the maximum rating of 5, even though human annotators occasionally do. This suggests a degree of conservatism in the model's scoring, particularly for the most severe cases of greenwashing.

6 Discussion

The results presented highlight the promising capabilities of large language models (LLMs) in approximating human judgments across multiple dimensions relevant to evaluating corporate sustainability claims. The performance metrics, particularly the consistently high Quadratic Weighted Kappa (QWK) scores demonstrate substantial agreement between the model's predictions and human annotations, especially in the category of Faithfulness. This suggests that LLMs are particularly adept at capturing the factual alignment of a statement with reference material or context, a crucial element when assessing the integrity of corporate disclosures.

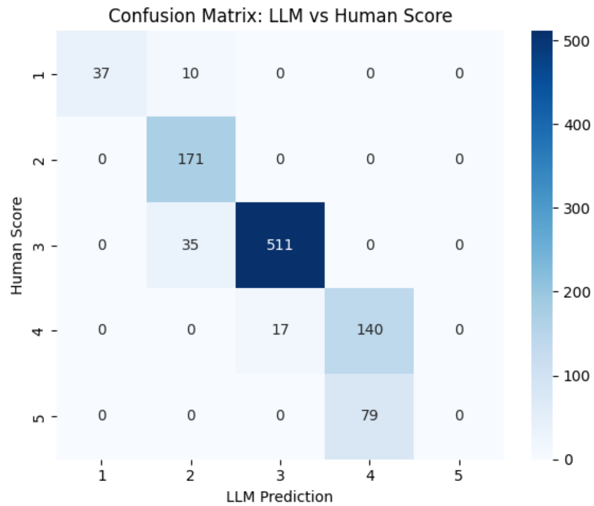


Figure 4: Confusion matrix comparing human and LLM greenwashing ratings

However, the Correctness dimension shows slightly more variability, as reflected by both a lower QWK (though still robust) and a higher Mean Absolute Error (MAE). The confusion matrices further support this observation, with more off-diagonal values indicating disagreement, and a tendency for the model to assign lower scores than humans. This suggests that the model adopts a more conservative or risk-averse stance in rating correctness, potentially due to a lack of contextual nuance or the subjective nature of the task.

It is important to acknowledge that the lack of a definitive ground truth in this domain limits the extent to which we can make binary judgments about right or wrong predictions. As in many real-world data scenarios, especially those involving human interpretation or subjective criteria, the emphasis is less on absolute correctness and more on reasoned and interpretable assessments. In this context, the LLM performs as a useful tool that offers reliable and explainable annotations, albeit with some degree of deviation in more ambiguous cases.

Moreover, as the discourse around greenwashing continues to evolve, the LLM's performance in identifying and scoring potentially misleading corporate climate claims provides meaningful insights. The model's understanding of greenwashing, aligned with academic definitions from prior research (Bernini *et al.*, 2023; de Freitas Netto *et al.*, 2020) suggests a solid conceptual grounding. When tasked with rating 1,000 real-world statements on a five-point greenwashing scale, the LLM produced scores that, while not always identical to human ratings, revealed patterns that can be interpreted and used constructively.

These findings support the use of LLMs in augmenting human judgment in ESG analysis, particularly in high volume tasks where manual annotation is costly or impractical. Nonetheless, caution is warranted. The observed discrepancies, especially in the correctness dimension, indicate areas where human oversight remains essential. Furthermore, as corporate sustainability communication becomes increasingly sophisticated, the potential for subtle or well-disguised greenwashing grows. This places renewed importance on developing tools that not only detect overt misrepresentation but also recognize more nuanced forms of deception.

While the findings are encouraging, several limitations must be acknowledged. First, the evaluation relies on human annotations that are themselves inherently subjective, especially

in areas such as greenwashing, where definitions and interpretations may vary, introducing a level of uncertainty in ratings. Second, the analysis focuses on a specific domain, i.e. corporate climate-related claims, limiting generalizability to other forms of ESG disclosures or to broader corporate communication contexts. Third, although the study uses robust performance metrics, it does not capture the full nuance of language, such as tone, implied meaning, or rhetorical strategies, which can be critical in the detection of subtle greenwashing. Finally, the model's performance may be sensitive to prompt design, fine-tuning settings, or updates to the underlying LLM architecture, all of which are subject to change.
Future research can extend this work in several directions. One avenue is to explore cross-domain generalizability, assessing whether the LLM's ability to evaluate sustainability claims transfers to other ESG domains (e.g., diversity, governance, or supply chain ethics). Additionally, incorporating multilingual capabilities would expand the applicability of such models to global corporate reporting. Another important direction is the integration of external verification tools or databases, such as emissions registries or financial disclosures, to provide factual grounding and enhance the model's accuracy. Finally, deeper analysis of disagreement cases, especially in correctness, may uncover systematic biases or reveal areas where model interpretability and training could be improved.

7 Conclusion

This study demonstrates the potential of large language models (LLMs) to support the evaluation of corporate sustainability claims, particularly in detecting and interpreting instances of greenwashing. Through a detailed performance assessment across key criteria – Correctness, Faithfulness, and Conciseness – the model shows strong alignment with human judgments, especially in areas where factual grounding is essential. While some variability remains, particularly in the evaluation of correctness, the model consistently delivers interpretable and actionable results.
Importantly, this work highlights the value of LLMs as tools that can mimic human judgment surprisingly well, particularly in complex and subjective domains where no definitive ground truth exists. In such settings – like greenwashing assessment – the emphasis should be less on rigid accuracy metrics and more on the model's ability to generate informed, consistent, and interpretable judgments.
As sustainability communication becomes more sophisticated, so too must our tools for scrutinizing it. By combining LLM capabilities with transparent methodologies and expert oversight, we can build more scalable and reliable approaches to ESG analysis. Future research should expand the scope and factual grounding of these models, ensuring they remain robust, adaptable, and well-calibrated to evolving standards of corporate accountability aligned with the evolving landscape of corporate sustainability.

References

Alan, A. Y., Karaarslan, E., & Aydin, Ö. (2025). *Improving LLM Reliability with RAG in Religious Question-Answering: MufassirQAS. Turkish Journal of Engineering*, *9*(3), 544–559.

Balaguer, A., Benara, V., Cunha, R. L. d. F., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., et al. (2024). *RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. arXiv preprint arXiv:2401.08406*.

Bernini, F., Giuliani, M., & La Rosa, F. (2023). *Measuring greenwashing: A systematic methodological literature review. Business Ethics, the Environment & Responsibility*.

Calamai, T., Balalau, O., Guenedal, T. L., & Suchanek, F. M. (2025). *Corporate Greenwashing Detection in Text–a Survey. arXiv preprint arXiv:2502.07541*.

Cartellier, F., Tankov, P., & Zerbib, O. D. (2023). *Can investors curb greenwashing? Available at SSRN 4644741*.

CFA Institute. (2024). *Climate Data in the Investment Process: Challenges, Resources, and Considerations*. https://blogs.cfainstitute.org/marketintegrity/2024/05/31/climate-data-in-the-investment-process-challenges-resources-and-considerations/

Cherief, A., Coz, O., Hicham, L., Sekine, T., Semet, R., & Svystunova, L. (2025). *Human Capital Assessment via Large Language Models alignment. SSRN*.

Chua, J. Z. (2025). *How I Won Singapore's GPT-4 Prompt Engineering Competition*. https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41/?ref=portkey.ai

Church, K. W., Chen, Z., & Ma, Y. (2021). *Emerging trends: A gentle introduction to fine-tuning. Natural Language Engineering*, *27*(6), 763–778.

Cohen, J. (1968). *Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin*, *70*(4), 213.

Crestani, F., Lalmas, M., Van Rijsbergen, C. J., & Campbell, I. (1998). *"Is this document relevant? ... probably" a survey of probabilistic models in information retrieval. ACM Computing Surveys (CSUR)*, *30*(4), 528–552.

de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. d. L. (2020). *Concepts and forms of greenwashing: A systematic review. Environmental Sciences Europe*, *32*, 1–12.

Delmas, M. A., & Burbano, V. C. (2011). *The Drivers of Greenwashing. California management review*, *54*(1), pp. 64–87.

Dumitrescu, A., Gil-Bazo, J., & Zhou, F. (2022). *Defining greenwashing. Available at SSRN 4098411*.

Durmus, E., He, H., & Diab, M. (2020). *FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. arXiv preprint arXiv:2005.03754.*

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). *The impact of corporate sustainability on organizational processes and performance. Management science*, *60*(11), 2835–2857.

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). *Ragas: Automated evaluation of retrieval augmented generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.

Feng, S., Balachandran, V., Bai, Y., & Tsvetkov, Y. (2023). *Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. arXiv preprint arXiv:2305.08281.*

Folke, O., & Ivan Erik Troedsson, A. (2025). *How effectively can AI be applied to extract ESG-related KPIs from annual reports?*

Ghitti, M., Gianfrate, G., & Palma, L. (2024). *The agency of greenwashing. Journal of Management and Governance*, *28*(3), 905–941.

Grewal, J., Hauptmann, C., & Serafeim, G. (2021). *Material sustainability information and stock price informativeness. Journal of Business Ethics*, *171*(3), 513–544.

Gutman, S. S., Egorova, M. M., Ibragimov, B. B., Nosirova, S. S., & Seytenova, V. (2024). *The Relationship Between ESG Metrics and Financial Performance of an Enterprise in the Oil and Gas Sector. E3S Web of Conferences*, *574*, 03004.

He, Q., Qiu, M., Buhui, Ben, J., Nguyen, Nguyen, N., & Visaltanachoti, N. (2024). *Greenwashing: Measurement and Implications. GRASFI 2024 Annual Conference, Annual Conference, Greenwashing.*

Islam, S. B., Rahman, M. A., Hossain, K., Hoque, E., Joty, S., & Parvez, M. R. (2024). *Open-rag: Enhanced retrieval-augmented reasoning with open-source large language models. arXiv preprint arXiv:2410.01782.*

Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2024). *Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. arXiv preprint arXiv:2403.14403.*

Jin, B., Yoon, J., Han, J., & Arik, S. O. (2024). *Long-context LLMs meet RAG: overcoming challenges for long inputs in RAG (2024). arXiv preprint arXiv:2410.05983.*

Joshi, S. (2025). *Mitigating LLM Hallucinations: A Comprehensive Review of Techniques and Architectures. Available at SSRN 5267540.*

Kaikaus, J., Li, H., & Brunner, R. J. (2023). *Humans vs. chatgpt: Evaluating annotation methods for financial corpora. 2023 IEEE International Conference on Big Data (BigData)*, 2831–2838.

Kim, D., Kim, B., Han, D., & Eibich, M. (2024). *Autorag: Automated framework for optimization of retrieval augmented generation pipeline. arXiv preprint arXiv:2410.20878.*

Krishna, S., Krishna, K., Mohananey, A., Schwarcz, S., Stambler, A., Upadhyay, S., & Faruqui, M. (2024). *Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. arXiv preprint arXiv:2409.12941.*

LangChain. (2025). *LangChain: Language Models in Application.*

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). *Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems*, *33*, 9459–9474.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Lyon, T. P., & Montgomery, A. W. (2015). *The means and end of greenwash. Organization & environment*, *28*(2), 223–249.

Mao, K., Liu, Z., Qian, H., Mo, F., Deng, C., & Dou, Z. (2024). *RAG-Studio: Towards In-Domain Adaptation of Retrieval Augmented Generation Through Self-Alignment. Findings of the Association for Computational Linguistics: EMNLP 2024*, 725–735.

Merth, T., Fu, Q., Rastegari, M., & Najibi, M. (2024). *Superposition prompting: Improving and accelerating retrieval-augmented generation. arXiv preprint arXiv:2404.06910.*

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2023). *Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. arXiv preprint arXiv:2401.00396.*

Ontiveros, A., Nikishina, I., Gomm, M., Schmitt, C., & Biemann, C. (2025). *ESG-Consultant: Developing of an ESG Compliance Consulting Tool for Companies Using RAG. International Conference on Applications of Natural Language to Information Systems*, 223–229.

OpenAI. (2024, July). *ChatGPT: Chat with GPT-4 Turbo.*

Rajaraman, S., Ganesan, P., & Antani, S. (2022). *Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PloS one*, *17*(1), e0262838.

Ran, K., Sun, S., Anh, K. N. D., Spina, D., & Zendel, O. (2025). *RMIT-ADM+ S at the SIGIR 2025 LiveRAG Challenge. arXiv preprint arXiv:2506.14516.*

Robeco. (2025). *Global Climate Survey 2025: A deepening divide in climate investing priorities.* https://www.robeco.com/en/insights/2025/06/global-climate-survey-2025.html

Rosa, G. M., Rodrigues, R. C., Lotufo, R., & Nogueira, R. (2021). *Yes, bm25 is a strong baseline for legal case retrieval. arXiv preprint arXiv:2105.05686*.

Roychowdhury, S., Soman, S., Ranjani, H., Gunda, N., Chhabra, V., & Bala, S. K. (2024). *Evaluation of rag metrics for question answering in the telecom domain. arXiv preprint arXiv:2407.12873*.

Ruiz-Blanco, S., Romero, S., & Fernandez-Feijoo, B. (2022). *Green, blue or black, but washing–What company characteristics determine greenwashing? Environment, Development and Sustainability*, *24*(3), 4024–4045.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108*.

Song, J., Wang, X., Zhu, J., Wu, Y., Cheng, X., Zhong, R., & Niu, C. (2024). *RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1548–1558.

Stagnol, L., Cherief, A., Farah, Z., Le Guenedal, T., Sakout, S., & Sekine, T. (2023). *Answering Clean Tech Questions with Large Language Models. Available at SSRN 4663447*.

Stammbach, D., Webersinke, N., Bingler, J., Kraus, M., & Leippold, M. (2022). *Environmental claim detection. Available at SSRN 4207369*.

Taguchi, C., Maekawa, S., & Bhutani, N. (2025). *Efficient Context Selection for Long-Context QA: No Tuning, No Iteration, Just Adaptive-k. arXiv preprint arXiv:2506.08479*.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). *Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663*.

Veturi, S., Vaichal, S., Jagadheesh, R. L., Tripto, N. I., & Yan, N. (2024). *Rag based question-answering for contextual response prediction system. arXiv preprint arXiv:2409.03708*.

Week, L. C. A. (2025). *Investors shift capital from US to Europe and Asia amid climate policy uncertainty*. https://www.londonclimateactionweek.org/news/investors-shift-climate-funds-2025

Whelan, T., Atz, U., Van Holt, T., & Clark, C. (2021). *ESG and financial performance. Uncovering the Relationship by Aggregating Evidence from*, *1*(2015-2020), 10.

Yang, Z., Nguyen, T. T. H., Nguyen, H. N., Nguyen, T. T. N., & Cao, T. T. (2020). *Greenwashing behaviours: Causes, taxonomy and consequences based on a systematic literature review. Journal of business economics and management*, *21*(5), 1486–1507.

A Appendix

A.1 Fine-Tuning $RoBERTa_{base}$

Table 3: Selected model's classification report

	precision	recall	f1-score	support
FALSE	0.98	0.93	0.95	398
TRUE	0.82	0.94	0.87	132
accuracy			0.93	530
macro avg	0.90	0.93	0.91	530

Source: Authors' calculations, Amundi Investment Institute.

A.2 BM25

The BM25 score between a query q and a document d is calculated by summing the contributions of each query term $t \in q$ that appears in the document d:

$$\mathrm{BM25}(q,d) = \sum_{t \in q \cap d} \log\left(\frac{N - \mathrm{df}(t) + 0.5}{\mathrm{df}(t) + 0.5}\right) \cdot \frac{\mathrm{tf}(t,d) \cdot (k_1 + 1)}{\mathrm{tf}(t,d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\mathrm{avgdl}}\right)} \tag{5}$$

where:

- N is the total number of documents in the corpus,
- $\mathrm{df}(t)$ is the number of documents containing term t (document frequency),
- $\mathrm{tf}(t,d)$ is the term frequency of t in document d,
- $|d|$ is the length of document d (e.g., number of words),
- avgdl is the average document length in the corpus,
- k_1 and b are free parameters, set to $k_1 = 1.2$ and $b = 0.75$.

BM25 ranks by term relevance, which supports high precision top-k retrieval. The criteria from the analyst framework are short clauses containing specific keywords. BM25 continues to serve as a robust baseline for text ranking, often delivering performance comparable to that of more complex retrieval models (Kim *et al.*, 2024; Rosa *et al.*, 2021; Thakur *et al.*, 2021).

A.3 Example: Criteria (Climate Engagement Framework)

The below tables contains a subset of criteria and guidance from the climate engagement framework that were used to generate the evaluation metrics set out in section 3.5.

Criterion	Guidance
Executive remuneration linked to the climate strategy?	Ok if executive remuneration is linked to relevant climate criteria, with thresholds in line with the strategy (assess thresholds against the company's targets), and representing at least 15% of total executive remuneration.
Incentive to develop low carbon solutions in executive remuneration?	Ok if inclusion of low carbon solution KPI in executive remuneration.
Ban of all incentives to grow hydrocarbon production in executive remuneration?	Ok if constraint so the executive remuneration policy does not incentivize to grow hydrocarbon production.

A.4 Example: Verbose Explanations of Scores

All explanations in the below table are metrics relating to the criterion "Ban of all incentives to grow hydrocarbon production in executive remuneration?".

Metric	Rating	Verbose Explanation
Faithfulness	5	(fully faithful) The answer accurately reflects the context provided. It clearly states that the context does not directly address a specific ban on all incentives tied to hydrocarbon production in executive remuneration. Instead, it summarizes the details from the context, such as S.'s remuneration policies being aligned with energy transition metrics, without introducing unsupported claims or speculation.
Conciseness	4	The answer is clear and focused, effectively addressing the question and staying on topic. However, it could be slightly more concise by trimming some phrases, such as "While it mentions S.'s remuneration policies and their integration..." to make it even more direct. Removing redundant details would make it cleaner without sacrificing clarity.
Correctness	3	Justification: The answer was factually correct in stating that the provided context does not directly mention a ban on all incentives tied to hydrocarbon production in executive remuneration. However, it missed clarity and completeness about S.'s current practices and goals concerning executive remuneration linked to climate goals and energy transition metrics. This could have been better contextualized within the broader shift in aligning executive pay structures with sustainability objectives. It lacks depth in explaining S.'s apparent stance on the matter, despite referencing relevant content.

Table 4: LLM-Generated Ratings Examples

A.5 Prompt structure for LLM-based evaluation metrics

Correctness

Answer: {response}
Context: {context}
Rate from 1 (poor) to 5 (perfect): How well does the answer address the question? Provide a short justification.

Faithfulness

Answer: {response}
Context: {context}
Rate from 1 (not faithful) to 5 (fully faithful): Is the answer fully supported by the context, or does it introduce unsupported claims? Explain briefly.

Conciseness

Answer: {response}
Rate from 1 (verbose/unclear) to 5 (concise and clear): Is the answer short, focused, and easy to understand? Explain briefly.

The ordinal scores rate how well the LLM is able to address a criterion based on a given context. A higher score implies documentation with richer information content whereas a lower score indicates a sparser informational content.

A.6 Does ChatGPT know what greenwashing is?

Prompt: Explain greenwashing in a few sentences.

Response: Greenwashing is a deceptive marketing practice in which companies exaggerate or falsely claim that their products or practices are environmentally friendly to attract eco-conscious consumers. This can involve misleading claims, irrelevant assertions, or the use of fake certifications, often distracting from the company's actual negative environmental impact. As a result, greenwashing undermines genuine sustainability efforts and can lead to consumer skepticism about authentic eco-friendly products.

A.7 Guidelines for Human Annotators

Table 5 contains the annotation guidelines for rating correctness, faithfulness and conciseness. Table 6 contains the annotation guidelines for greenwashing.

Table 5: Annotation Guidelines for Correctness, Faithfulness, and Conciseness

Score	Correctness	Faithfulness	Conciseness
1	Factually incorrect; major errors or contradictions.	Does not reflect source; introduces unrelated or false content.	Overly verbose or repetitive; much irrelevant information.
2	Contains some factual inaccuracies or unsupported claims.	Loosely based on source but omits key points or adds misleading elements.	Some unnecessary detail; lack of focus.
3	Mostly correct with minor factual issues.	Generally faithful, but some misinterpretation or omission.	Adequately concise but could be more focused.
4	Accurate with minor phrasing issues.	Faithfully represents most of the key content from the source.	Clear and concise with little redundancy.
5	Fully accurate; all facts correct.	Completely faithful to the source; no omissions or additions.	Highly concise; includes only essential and relevant information.

Table 6: Greenwashing Annotation Guidelines (1 - 5 Rating Scale)

Rating	Guideline
1 None	Clear, factual, and specific claim with supporting evidence; no misleading language.
2 Low	Mostly factual but includes vague or unqualified language; minor risk of misinterpretation.
3 Moderate	Mix of accurate and misleading elements; may omit context or selectively present facts.
4 High	Exaggerated or misleading claim with little evidence; likely to create a false impression.
5 Definite	Clearly deceptive or unsubstantiated; contradicts known facts or aims to mislead.

A.8 Example: Greenwashing Assessments

We apply our methodology introduced in section 3 to identify climate claims that are potential greenwashing candidates.

Claim:

C. is a member of the Greater Houston Partnership and its One Houston Together buyer's cohort, dedicated to advancing racial equity and increasing spending on minority business enterprises. Our Chief Procurement Ofcer (CPO), Steve Freeman, participates in the cohort's supplier diversity workstream, CPO Convening, aimed at expanding collaboration and increasing minority business growth. Racial equity audit In 2022, C. commissioned an external review of our racial equity and diversity and inclusion initiatives. The resulting report recognizes our progress and offers recommendations for how we might continue to improve. investing in our people We believe in fostering leadership at every level, whether supervising a team or informally influencing peers. Our development programs are designed

to help people achieve their full potential, accelerate leadership capabilities, and cultivate a diverse and robust talent pipeline in an everchanging work environment. Because C. believes in developing leaders at all levels, we introduced a new coaching program to reach deeper into the organization, including front-line supervisors, managers and, more recently, individual contributors. The BetterUp platform offers live, one-on-one coaching sessions and Coaching Circles with small groups of peers. Since its launch in 2020, C. has provided coaching to over 2,900 supervisors, managers and individual contributors in 45 countries and 18 languages. C. also offers multiple leadership development programs with a focus on self-refection, continuous learning and strategic networking. Specifically for women, the Transformational Leadership for Multicultural Women Program addresses headwinds faced by multi-cultural women in their leadership journeys. The Global Women's Leadership Development Program, now in its third year, promotes acceleration of womenâs leadership development and experience. Almost 100 women have participated in these two programs.
Score: 2
Explanation:
Score: **2 (Low Greenwashing)**

Explanation: C.'s claims in the provided context primarily focus on racial equity, diversity, sustainability initiatives, and workplace culture rather than ambitious or overstated environmental promises. However, given C.'s historical operations as a major oil and gas company, there are areas that could raise questions of inconsistency between messaging and broader corporate activities. Here's the breakdown of this assessment:

Why the score is low (2): 1. **Concrete Actions and Specific Initiatives**: - The document outlines clear steps taken by C., such as partnerships with minority organizations (e.g., the Thurgood Marshall College Fund), the use of measurable diversity-focused programs (e.g., Global Women's Leadership Development and Transformational Leadership for Multicultural Women programs), and targeted hiring initiatives to attract underrepresented groups (e.g., partnerships with HBCUs and historically African American organizations). - Programs such as their supplier diversity initiatives and investments to support underrepresented communities are verifiable and focus on material outcomes.

2. **Efforts Beyond Marketing Rhetoric**: - C. commissioned a racial equity audit and has publicized its findings and next steps, which allows for increased transparency. - It also provides specific participation metrics (e.g., 90% full-time offers extended to interns were accepted, coaching programs reaching over 2,900 individuals in 45 countries) and names the third-party organizations they support or collaborate with, like Opportunity@Work.

3. **Proactive Steps for Employee and Community Engagement**: - C. appears to be actively growing its diversity and inclusion efforts, including "skills-based hiring" initiatives and investments in programs tailored to multicultural women and underrepresented talent.

Where skepticism arises: 1. **Broader Context of Corporate Activities**: - As a fossil fuel company, most of C.'s environmental and social activities can be seen through a lens of balancing corporate reputation against its primary business model, which inherently has significant environmental and public health impacts. - While the claims focus on social issues like equity and inclusion, they may draw attention away from C.âs broader environmental practices or potential shortcomings in climate action compared to their commitments.

2. **Mention of Climate or Carbon Goals (Tangentially Related)**: - Although not the focus of this particular excerpt, C. is known to present initiatives like offset-paired LNG as part of their sustainability pitch. Critics might argue that these types of environmental claims could undermine trust due to overemphasis on lower-impact carbon solutions while broader fossil fuel extraction operations continue. - Claims like "ever-cleaner energy" and investments in renewable initiatives are somewhat tangentially mentioned, but the document does little to acknowledge systemic challenges tied to the role of oil and gas in maintaining

racial inequities perpetuated by environmental harms (e.g., refineries and pipelines often impacting marginalized communities).

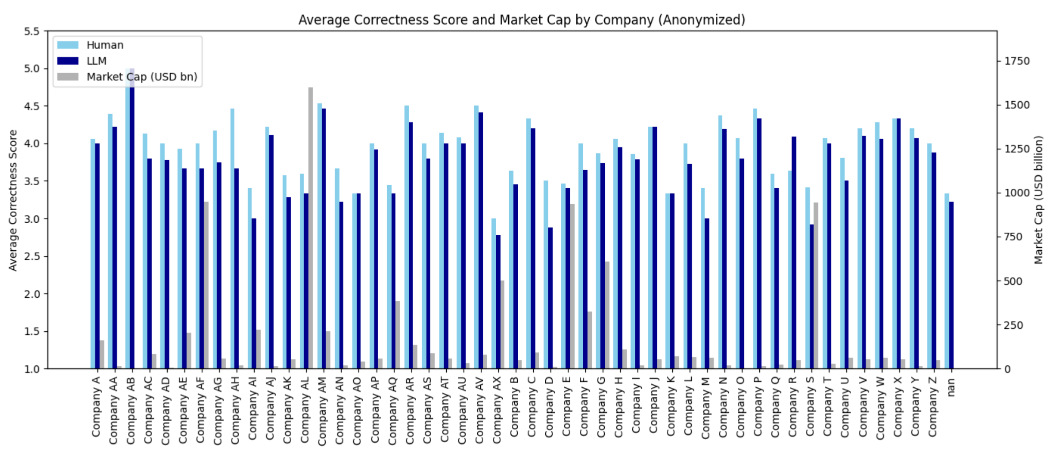
3. **Potential Lack of Scale or Impact Quantification**: - While individual initiatives outlined are compelling, they may not necessarily reflect systemic change at a broad scale. For example, while C.'s $15 million pledge to address racial equity is noted, this amount could be viewed as minimal in comparison to the company's overall earnings and environmental harm caused in vulnerable communities.

Summary: In the context of diversity and equity, C. seems to genuinely invest resources and implement measurable strategies, which sets it apart from companies merely touting good intentions without delivering results. However, there is a lingering perception that these initiatives may function, at least in part, as reputation buffers against the more systemic critiques commonly levied against the oil and gas industry. While there is no direct evidence of greenwashing in the given claims, caution is warranted when interpreting the motives behind their social and environmental commitments. Therefore, a score of 2 (not free of suspicion, but mostly credible) is justified.

A.9 Correctness and Market Capitalization

No clear or consistent relationship is observed between company market capitalization and average correctness scores.

Figure 5: Correctness and Market Capitalization



Source: Amundi Investment Institute.

Investment Institute

WORKING PAPER 178 | **OCTOBER 2025**