systems

MDPI

Article

The AI Annotator: Large Language Models' Potential in Scoring Sustainability Reports

Yue Wu, Peng Hu * and Derek D. Wang

School of Business Administration, Capital University of Economics and Business, 121 Zhangjialukou, Beijing 100070, China
* Correspondence: hu_peng@cueb.edu.cn

Abstract

To explore the potential of Large Language Models (LLMs) as AI Annotators in the domain of sustainability reporting, this study establishes a systematic evaluation methodology. We use the specific case of European football clubs, quantifying their sustainability reports based on the sport Positive matrix as a benchmark to compare the performance of three state-of-the-art models (i.e., GPT-4o, Qwen-2-72b-instruct, and Llama-3-70b-instruct) against human expert scores. The evaluation is benchmarked on dimensions including accuracy, mean absolute error (MAE), and hallucination rates. The results indicate that GPT-4o is the top performer, yet its average accuracy of approximately 56% shows it cannot fully replace human experts at present. The study also reveals significant issues with overconfidence and factual hallucinations in models like Qwen-2-72b-instructon. Critically, we find that by implementing further data processing, specifically a Chain-of-Verification (CoVe) self-correction method, GPT-4o's initial hallucination rate is successfully reduced from 16% to 10%, while accuracy improved to 58%. In conclusion, while LLMs demonstrate immense potential to streamline and democratize sustainability ratings, inherent risks like hallucinations remain a primary obstacle. Adopting verification strategies such as CoVe is a crucial pathway to enhancing model reliability and advancing their effective application in this field.

Keywords: AI; large language model; GPT-4o; sustainability report; sustainability rating

Academic Editor: Wen-Chyuan Chiang

Received: 8 August 2025
Revised: 28 September 2025
Accepted: 9 October 2025
Published: 11 October 2025

Citation: Wu, Y.; Hu, P.; Wang, D.D. The AI Annotator: Large Language Models' Potential in Scoring Sustainability Reports. *Systems* **2025**, *13*, 899. https://doi.org/10.3390/systems13100899

1. Introduction

With the traction of sustainable investing, sustainability disclosure has gradually become an important bridge for interaction between companies and capital markets. The disclosure provides critical information about a company's long-term environmental and social impacts as well as its progress in addressing sustainability goals [1]. Sustainability scores, which can show a company's commitment and reporting to their sustainability goals, are used by investors to evaluate how well a company incorporates sustainability factors into its overall strategy [2]. Professional ESG (Environmental, Social, and Governance) raters, such as Morgan Stanley Capital International (MSCI), Sustainalytics, and London Stock Exchange Group (LSEG), have emerged to provide clear and concise evaluation metrics and scores for investors and other stakeholders.

Currently, sustainability ratings have garnered increasing interest and play a crucial role in the field of socially responsible investing. According to Sustainability (2020), 65% of surveyed investors use ESG ratings at least once a week [3]. Sustainability ratings are

popular mainly because they are among the few tools that allow investors to consider sustainability data in a straightforward and accessible manner. While investors can integrate sustainability information into their sustainable investment portfolios through various strategies, all of these strategies require extensive analysis and specific data [4]. These ESG/sustainability ratings help to overcome these data challenges.

Traditionally, sustainability ratings are created by human research analysts using proprietary methodologies. These analysts scrutinize company disclosures, articles, news and other data to evaluate a company's sustainability performance. Figure 1 illustrates the process underpinning analyst-driven ESG research, highlighting the critical role of analysts" expertise, particularly in the final steps [5].

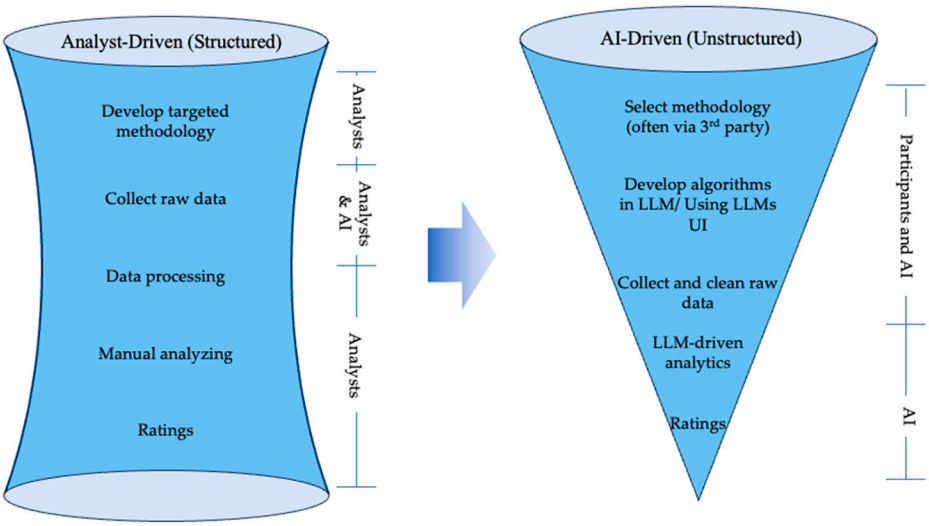

Figure 1. The role of human research analysts vs. AI analysts in the processes of sustainability ratings providers.

However, this traditional approach has its drawbacks. These rating agencies usually charge high fees [6], and the evaluation process is relatively complex and time-consuming. Given these challenges, there is an urgent need for a rating system that is highly accessible, provides fast feedback, and is user-friendly. Large language models (LLMs) are promising because they apply artificial intelligence (AI) to various tasks, such as content generation, text analysis, and trend prediction. Given this context, can AI be applied to sustainability rating? We propose the following process: AI-driven sustainability research can extract information from unstructured data, irrespective of data type or language. In this approach, the role of the analyst is primarily limited to the initial steps, significantly reducing their workload. Additionally, AI-driven methods can be utilized by individuals with no prior experience in sustainability rating, thus broadening access to the rating process. The specific process is illustrated in Figure 1.

To conduct a robust and comprehensive evaluation, this study employs a comparative framework featuring three state-of-the-art LLMs released in close succession in mid-2024, which is OpenAI's proprietary GPT-4o, and two leading open-weight models, Meta's Llama-3-70b-instruct and Alibaba's Qwen-2-72b-Instruct. Then, we select the European professional football sector as our case study. This "non-traditional" industry provides an ideal testbed for several reasons. First, the industry is in the early stages of formalizing its sustainability reporting. While voluntary disclosures have existed, new regulations such as the EU's Corporate Sustainability Reporting Directive (CSRD) are now beginning to mandate more structured reporting for a significant number of clubs, marking a crucial transition phase for the industry [7]. Second, despite its nascent reporting landscape,

the industry possesses clear and emerging evaluation standards that provides a credible basis for assessment. This credibility is anchored at a high level by the United Nations Sports for Climate Action Framework, which provides the strategic guiding principles for the global sports community to combat climate change. These principles are then operationalized by organizations like Sport Positive Leagues. Since 2019, it has developed a detailed, publicly available matrix to rank clubs on their sustainability performance, creating what has become a de facto industry benchmark, widely cited by media and clubs alike [8]. This unique combination of a maturing reporting environment and an established, objective benchmark makes the football industry an ideal, controlled setting for our primary objective: to assess the degree to which the scores generated by these diverse AI models align with the established, human-annotated ratings from the Sport Positive matrix. By comparing the performance of these leading models against this human-expert baseline, this research provides an in-depth exploration of the feasibility, accuracy, and potential biases of the current generation of LLMs in the nuanced field of ESG evaluation.

Our findings indicate that while LLM shows significant potential, it cannot yet fully replace human analysts. In a comprehensive performance assessment of the three models, GPT-4o shows the strongest performance, leading in task completion rate (100%), accuracy (56%), and stability. In contrast, Llama-3 struggled with task completion (82.4%), while Qwen-2-72b-instructon lagged significantly in accuracy (0.34). A key finding relates to the models' confidence calibration: GPT-4o exhibits good calibration, with its confidence level positively correlating with accuracy. Conversely, Llama-3-70b-instruct and Qwen-2-72b-instructon exhibits a significant "overconfidence" bias, with their accuracy being lowest at their highest confidence levels. Furthermore, all models perform better on English than non-English texts, and the accuracy achieved using the API is higher than that from the user interface.

We also investigate the issue of LLM hallucinations. The results show that Qwen-2-72b-instructon's hallucination rate (33%) is substantially higher than that of GPT-4o (16%) and Llama-3-70b-instruct (16%). More critically, the nature of these hallucinations differed profoundly: errors from GPT-4o and Llama-3-70b-instruct are predominantly "faithful hallucinations" (i.e., failing to adhere to rating instructions), whereas Qwen-2-72b-instructon exhibits a high frequency of "factual hallucinations" (i.e., fabricating information does not present in the source material). However, by implementing a self-verification method, "Chain-of-Verification" (CoVe), GPT-4o's overall accuracy increased from 56% to 58%, and its hallucination rate is significantly reduced from 16% to 10%. This improvement not only demonstrates the feasibility of enhancing LLM reliability in sustainability ratings but also highlights its considerable future potential. Through this series of analyses, we provide empirical evidence on the current application status of AI in analyzing sustainability reports and suggest directions for future improvements in AI model performance.

This research makes three primary contributions. First, this study contributes to the existing literature by introducing large language models into the sustainability rating field. It broadens the theoretical boundaries of the intersection between sustainability and AI, and provides a new perspective for understanding the democratization of sustainability disclosure. Second, in terms of methodological innovation, this study employs an exploratory case analysis, integrating the several advanced large language models to develop a feasible methodology for the automation and intelligent analysis of sustainability ratings. This provides future researchers with a new approach and analytical pathway for handling complex sustainability information. Third, in terms of practical significance, this study addresses current challenges in the sustainability rating market, such as high costs

and inefficiencies, through technological means, offering a more economical and reliable sustainability evaluation approach for enterprises and individual participants.

This paper is structured as follows: Section 2 reviews relevant literature on sustainability ratings, large language models (LLMs), and their applications in the sustainability domain. Section 3 outlines our proposed methodology for leveraging the LLMs to conduct sustainability ratings. We present our experimental findings in Section 4, followed by a critical discussion of hallucinations in LLM-generated ratings in Section 5, along with possible ideas for improving LLM performance and proven in Section 6. Finally, Section 7 summarize our key findings and concludes the study.

2. Literature Review

This study examines the convergence of large language models and sustainability ratings. The literature review is organized around three key themes: the current state of sustainability ratings, the application of LLMs, and the application of LLMs in the field of sustainability.

2.1. Current State of Sustainability Ratings

Sustainability ratings capture value that traditional financial reports often overlook and serve as reliable barometers of a company's sustainability performance. ESG is the most widely cited framework in sustainability research; it argues that genuine sustainability can be achieved only through a delicate balance among these three interdependent pillars. As ESG issues grow in global importance, stakeholders, including investors, consumers, and policymakers, are paying ever-closer attention to corporate sustainability performance.

In the corporation, sustainability ratings reflect companies" ESG policy and their intention and ability to practice these policies [9]. When companies are included in ESG evaluations by rating agencies, it serves as an important signal to investors and influences their investment decisions [10]. For instance, after obtaining ESG ratings, companies often experience a decrease in the average cost of capital and an increase in Tobin's Q [11]. Cellier and Chollet (2016) find that there is a strong positive stock market reaction regardless of whether the Vigeo social rating is good or bad [12]. Additionally, ESG ratings can enhance corporate reputation, positively affects risk-adjusted profitability, reduce financial distress risk, and establish firm competitiveness [13]. And Shanaev and Ghimire (2022) find that ESG rating changes significantly impact stock returns, with downgrades leading to notable negative monthly risk-adjusted returns, particularly for ESG leaders [14].

However, current sustainability rating systems face significant challenges. From the perspective of rating providers, these systems are time-consuming, resource-intensively, and require extensive experience. Analysts must carefully scrutinize voluminous reports, and the subjective interpretation of sustainability data can yield divergent ratings for identical sustainability reports [15]. Prominent rating agencies such as MSCI (Morgan Stanley Capital International; New York, NY, United States), employing over 200 analysts; Sustainalytics (Amsterdam, Netherlands) with its team of 200 analysts; and Refinitiv (now is named LSEG, London Stock Exchange Group; London, United Kingdom), harnessing the expertise of 700 research analysts, continue to grapple with the limitations of manual analysis despite leveraging advanced technologies for data collection. This reliance on human judgment often leads to delays and inconsistencies in scoring [16]. The inherent inefficiencies in these labor-intensive processes underscore the urgent need for innovative solutions capable of handling the escalating volume and complexity of sustainability data.

Additionally, from the perspective of rating receivers, the mainstream rating agencies can be divided into two types: those serving institutional investors and those serving large companies. MSCI and Bloomberg (New York, NY, United States) do both, while DJSI Robe-

coSAM (New York, NY, United States (S&P Global)) serves investors and EcoVadis (Paris, France) is focused on the supply chain. This implies that resource-constrained investors and small to medium-sized enterprises (SMEs) may face challenges in accessing comprehensive and affordable sustainability ratings and analysis services. The high costs and complexity associated with the services provided by agencies like MSCI and Bloomberg can be prohibitive. As a result, these smaller entities might not have the same opportunities to leverage detailed ESG data for informed decision-making and performance improvements.

2.2. Applications of LLMs

Large Language Models, such as the GPT series, represent significant advancements in natural language processing (NLP) and deep learning. These models process vast amounts of text data, exhibiting exceptional capabilities in language understanding and generation [17]. The core technology behind LLMs is based on neural network-based deep learning models, which learn and generate natural language through pre-training and fine-tuning [18].

The broad applicability and potential of LLMs have led to its rapid adoption and widespread use across various industries. In healthcare, LLMs are employed for automated medical record analysis, diagnostic support, and medical documentation, improving efficiency and reducing human error [19,20]. In education, LLMs have exhibited its impact on learning, teaching, and assessment, providing integration recommendations and support for diverse learners [18,21,22]. In finance and accounting, LLMs assist with financial-report analysis, market forecasting, risk management, and auditing [23–29]. Bernard et al. (2024) adapted Llama-3 to 10-K footnote data to derive a quantitative gauge of corporate complexity [30]. Evidence from Eulerich and Wood (2023) and Emett et al. (2025) highlights ChatGPT's value in streamlining internal-audit workflows [25,27]; Föhr et al. (2023) employ it to verify EU-taxonomy alignment in sustainability reports [26].

Furthermore, there is significant interest in using LLMs for tasks that traditionally require manual labeling [31]. OpenAI's InstructGPT is a pioneering work in instruction-based prompts in LLMs [32]. The crucial is to train LLMs by providing clear instructions for task execution. These instructions outline the expected response to a given prompt, and the LLM is optimized to generate a response consistent with these instructions. Instruction based prompts are now widely used to solve various information retrieval tasks [33]. However, existing studies heavily rely on metrics of accuracy and inter-rater reliability, limited to binary scores between 0 and 1 [34–36]. Their application in broader quantitative text scoring is minimal.

2.3. Applications of LLMs in the Field of Sustainability

The application of LLMs in the field of sustainability is a promising and valuable topic. AI has the potential to enhance corporate sustainability, particularly by positively influencing environmental governance and social responsibility [37]. Existing studies indicate that LLMs can support ESG data analysis in some extents [34,38,39]. Research in this field can be primarily categorized into two streams: one that examines the use of LLMs for qualitative text analysis in ESG, and another that explores their application in quantitative ESG scoring. Some scholars seek to use LLMs to analyze firms" ESG disclosures and extract insights from these disclosures [40–42]. For instance, Lin et al. (2024) develops GPT4ESG, a BERT and GPT-based system that rapidly analyzes companies" ESG performance [42]. This model outperforms ESG-BERT in classifying ESG data from corporate reports through advanced data processing and fine-tuning techniques. Kim et al. uses ChatGPT to summarize the economic utility disclosed by companies, uncovering the link between information "bloat" and adverse capital market outcomes [43]. Huang

et al. (2023) develops the FinBERT model, which achieved higher accuracy in sentiment categorization of labeled ESG sentences compared to other machine learning models and dictionary approach [23,44]. Ni et al. (2023) develops the CHATREPORT system, which employs the TCFD framework to scrutinize corporate sustainability reports and assess their compliance [45]. Bronzini et al. (2023) leverages LLMs to derive semantically structured ESG-related data from sustainability reports, revealing the web of ESG actions among companies [46]. Moodaley and Telukdarie (2023) bolsters the capability to pinpoint green claims and detect "greenwashing" practices by training LLMs with an extensive collection of sustainability-related texts [47]. Managi et al. (2024) uses GPT-4 to analyze the relationship between the readability of sustainability reports and ESG scores for US companies, finding that context-dependent readability scores positively correlate with ESG scores, particularly among companies with lower social visibility [48]. In the meantime, the LLM is increasingly being employed to identify climate-related risks in corporate disclosures. Luccioni (2020) created ClimateQA, a custom transformer-based model that leverages Natural Language Processing to pinpoint climate-relevant sections in financial reports through a question-answering approach [49]. In a similar vein, Bingler (2022) introduced ClimateBERT, a fine-tuned BERT model combined with text mining algorithms, to analyze climate-risk disclosures across TCFD's main categories [50].

In the realm of ESG rating, scholars have turned to explore the potential of LLMs to argument assessment precision and efficiency. De Villiers (2024) shows AI in the field of non-financial reporting has the potential to improve efficiency, enhance data analysis and information quality, while also increasing the credibility and transparency of reports, making information more comprehensible, and thereby boosting stakeholder engagement [51]. Lee (2024) generates a text-based automated ESG grade assessment framework grounded in pre-trained ensemble models, achieving an accuracy of 80.79% with batch size 20 [52]. Kannan and Seki (2023) constructs a labeling model by fine-tuning a large language model pre-trained on financial documents demonstrating effective extraction of textual evidence for ESG scores, with macro average F1 scores of 0.874 for ESG labeling and 0.797 for ESG sentiment labeling, outperforming models pre-trained on general data [53]. Another study leverages a tree-based machine learning approach to analyze ESG metrics from Refinitiv Asset4 and MSCI, identifying the key metrics for building efficient portfolios and thus addressing the prevalent discrepancies in ESG ratings [54]. However, fine-tuning and inference using large pre-trained language models may require a lot of computing resources, which may limit the application of this method in resource-constrained settings.

In summary, while the literature demonstrates the growing capacity of LLMs for the analysis and scoring of sustainability reports, our review identifies several critical, unaddressed gaps that this study aims to fill. First, where prior research often conducts broad, cross-sectoral analyses using general frameworks like TCFD, this study provides a pioneering methodological benchmark. We test a diverse portfolio of leading LLMs within a novel and specialized domain: professional sports clubs. Crucially, we move beyond generic criteria by evaluating these models directly against a highly structured, industry-specific scoring matrix from a third-party organization (Sport Positive Leagues), enabling a far more controlled and replicable assessment of their capabilities on complex, criteria-driven tasks. Second, and more critically, existing studies predominantly focus on task performance (e.g., accuracy). Our work extends significantly beyond this by conducting a deep, multi-faceted diagnostic analysis of the models" operational reliability and failure modes. We do not simply measure if a model is right or wrong; we investigate how and why it fails. Specifically, this study systematically: (1) evaluates confidence calibration to uncover dangerous overconfidence biases; (2) quantifies linguistic bias by comparing performance on English and non-English reports; and (3) dissects model hallucination, not

just by its frequency, but by classifying its nature into factual (inventing information) versus faithful (violating instructions) errors. This granular analysis provides an unprecedented look into the behavioral characteristics of different LLMs in a high-stakes rating context. Finally, this research completes the cycle from problem identification to potential solution. While other studies often stop at diagnosis, we test a mitigation strategy by implementing and quantifying the impact of the Chain-of-Verification (CoVe) method, demonstrating a concrete path to improving model reliability. By addressing these gaps, our study provides a uniquely comprehensive and critical assessment of the true readiness of off-the-shelf LLMs for specialized ESG evaluation, offering vital insights for their responsible deployment.

3. Method and Data

Rating sustainability reports is indeed a complex task involving both qualitative and quantitative analyses and can be costly. We design specific prompts for the GPT-4o model (API version) to complete this task. Figure 2 illustrates the structure of the entire experimental design.

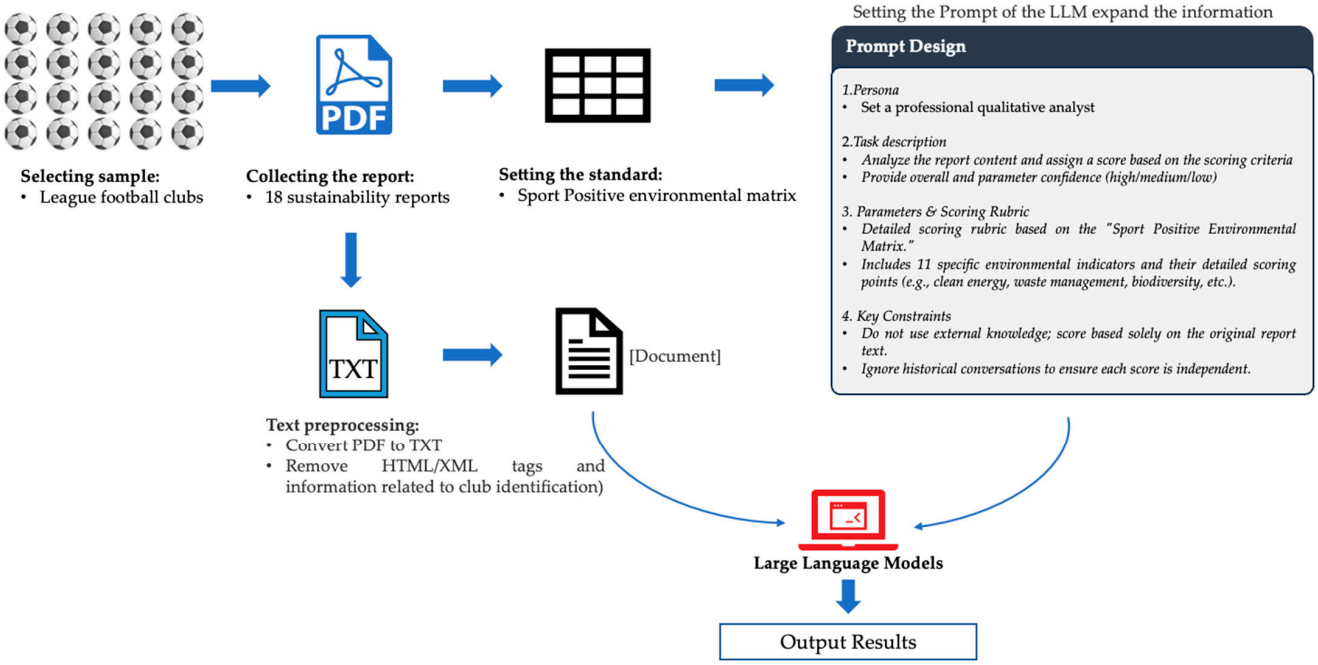

Figure 2. Technical flowchart for this research. Note: This flowchart illustrates the complete steps from sample selection, report collection, text preprocessing, and LLM scoring. The "Prompt Structure Design" section of the figure provides a summary of the prompt structure we designed to guide large-scale language model analysis. Appendix B provides a complete and precise version of the prompts used in API calls.

3.1. Model Selection

To ensure a robust and comprehensive evaluation, this study employs a comparative methodology, analyzing the performance of a carefully selected portfolio of state-of-the-art Large Language Models (LLMs) on the task of ESG report analysis. Rather than focusing on a single model, our approach is designed to benchmark capabilities across different development philosophies, resource origins, and access modalities. To facilitate a fair comparison, our selection includes three models released within the same period. Our selected models are:

GPT-4o: Chosen as the industry-leading benchmark. As a flagship model from OpenAI, GPT-4o represents the high-water mark for commercially available, closed-source

models and serves as a powerful, well-established baseline for performance comparison in complex reasoning tasks [6].

Llama-3-70b-instruct: Developed by Meta, Llama-3-70b-instruct stands at the forefront of the open-source community, offering performance that is competitive with top-tier proprietary models. Its inclusion is crucial for this study as it provides a direct counterpoint to the closed-source paradigm, enabling a nuanced comparison of performance between the two dominant development philosophies [55].

The Qwen-2-72b-instructon: Developed by Alibaba Cloud, these models represent the frontier of LLM research from a non-Western technology leader, allowing us to investigate the consistency of model performance across different training data and cultural contexts [56].

This strategic selection provides a methodologically sound basis for assessing the generalizability of LLMs in the specialized domain of financial sustainability analysis. We acknowledge the existence of other highly capable models, such as Google's Gemini or Anthropic's Claude series, a primary competitor to OpenAI. However, our selection of GPT-4o, Llama-3-70b-instruct, and Qwen-2-72b-instructon was deliberately designed to cover three critical and distinct axes for comparison: (1) the dominant proprietary model, (2) a top-tier open-source alternative, and (3) a leading model from a different geopolitical and data ecosystem. Within this comparative framework, including an additional proprietary model like Claude would be redundant. It would not introduce a new fundamental dimension to our analysis but would unnecessarily increase the study's complexity.

3.2. Sample Selection and Standard Setting

The evaluation of sustainability reports currently relies on a variety of standards, such as the Global Reporting Initiative (GRI) and Bloomberg ESG disclosure score. However, accessing these standards can be prohibitively expensive. For instance, obtaining the underlying criteria behind Bloomberg ESG disclosure score is challenging for the average person. Similarly, while the GRI standards are comprehensive, they are also complex and difficult to implement. In our search for alternative approaches, we discover a compelling area of study: football clubs. This sector exhibits relatively sparse ESG disclosure assessments but has an established, objective standard that has been in use for several years.

In this study, we analyze the 2018–2023 top-tier football markets in Europe, collectively known as the Big 5, which include the Premier League in England, Bundesliga in Germany, La Liga in Spain, Serie A in Italy, and Ligue 1 in France. These leagues represent the largest and most popular domestic football markets in Europe. During this period, we review of the official websites for these clubs reveals that only 18 have published sustainability reports. We subsequently gather the most recent sustainability reports from these 18 clubs for our analysis (refer to Appendix A).

To explore the potential of LLMs in scoring sustainability reports, we adopt the Environmental Sustainability Matrix developed by the professional organization [8] as our benchmark. This matrix provides a comprehensive evaluation framework, assessing clubs based on 11 environmental and social parameters with scores ranging from 0 to 3. Additionally, bonus points are awarded for exceptional policies and commitments, as well as sustainable transportation practices. To align with our study's scope, we exclude the Communications and Engagement indicator, as it relies on diverse sources such as websites that are not reflected in our primary data sources. For a detailed description of all indicators, please refer to Appendix B.

To ensure a robust and fair evaluation process, we assemble a research team consisting of three expert researchers, each without prior knowledge of or preference for any specific football club. Each researcher independently assesses the sustainability reports of the

football clubs, employing the criteria outlined in the Environmental Sustainability Matrix. After completing their individual assessments, the research team convenes to collectively discuss and reconcile their findings, addressing any potential biases that may have arisen. Through this collaborative process, we establish a consensus on an artificial sustainability disclosure score for each club. This consensus score is crucial as it not only ensures the quality and reliability of human annotation but also serves as a benchmark against which the performance of the LLM's can be compared.

It is important to acknowledge that the selection of 18 football clubs constitutes a focused, specific sample. In the broader field of ESG research, acquiring high-quality, standardized ESG reports presents a significant challenge, particularly in niche sectors like the sports industry. Previous studies have highlighted issues of incomplete data disclosure and inconsistent formatting, which inherently limit the feasibility of large-scale sample studies [57,58]. Therefore, this study is positioned as an exploratory investigation. Its primary contribution lies in proposing and testing a novel LLM-based evaluation framework within a well-defined domain, rather than aiming for conclusions with broad, universal applicability.

By using this approach, we aim to provide a comprehensive evaluation of the football clubs' sustainability disclosures and rigorously assess the reliability of the LLM's rating capabilities. This methodology underscores the importance of combining expert human judgment with advanced AI tools to achieve precise and meaningful assessments in the realm of sustainability.

3.3. Pre-Text Processing

Preprocessing is a technique that transforms unstructured data into a comprehensible and logical format [59]. The preprocessing phase is crucial when using LLMs, as it allows us to remove unnecessary information from the data, enabling subsequent processing phases. Initially, we use the Python library PDFplumber (version 0.11.0) to extract raw text from the PDF documents. Recognizing that this conversion method omits structured data from tables and figures and can disrupt the original formatting, our methodology focused specifically on the narrative content of the reports. To address the resulting text fragmentation and to manage inputs for the language model, a multi-step cleaning and restoration phase was performed. The long, extracted text was first systematically divided into smaller, manageable chunks, ensuring each segment was within the model's processing limits. Each chunk was then individually sent to the GPT-4o model with instructions to correct and reconnect disjointed sentences. As the model returned each corrected chunk, they were sequentially reassembled, a method that not only restored the semantic coherence within the text but also preserved the original document's contextual flow. Finally, to prevent model bias based on prior knowledge, we anonymized the data. All club-specific identifiers, such as names and contact information, were manually replaced with generic placeholders ("Club A" to "Club R"). This ensured that the model's ESG rating was based solely on the textual content of the sustainability reports.

3.4. Prompt Engineering

To conduct standardized and reproducible evaluations of the selected large language models (GPT-4o, Llama-3-70b-instruct, and Qwen-2-72b-instructon), we designed and implemented a multi-layered prompt engineering strategy. This strategy centers on building a highly structured and explicit set of system instructions to transform these general-purpose models into focused ESG analysis tools. Our approach places the evaluation task within a zero-shot learning context, where the models are not provided with any completed rating examples before performing the task.

The performance of LLMs largely depends on the prompts and context provided [35]. OpenAI endorses various strategies for effective prompt engineering, including writing clear instructions, dividing complex tasks into simpler sub-tasks, and providing reference texts [6]. Building on these guidelines and after numerous adjustments, we crafted the prompt detailed below.

First, to guide the models' analytical framework, we employed a role-playing paradigm by assigning a specific identity to the model [60,61]. This technique is designed to activate internal reasoning patterns associated with expert-level qualitative analysis, thereby enhancing the objectivity and depth of the output.

> *Persona:* You are a professional researcher named Lilya. You are an expert in qualitative content analysis. You are always focused and rigorous.

Second, we describe the task. According to Wu and Hu, we rephrase all the questions into single sentences and break down complex instructions into multiple prompts [62]. To further ensure the reliability of the analysis, we incorporated a metacognitive requirement instructing the model to provide a "high, medium, or low" confidence level for each assessment [24]. This feature helps identify potential inferences based on insufficient evidence.

> *Task description:* Analyze the content in the uploaded [documents]. Please follow the parameters below to score these [documents]. Ensure that the scoring strictly follows these parameter settings. Please also provide the confidence levels for your analysis, both overall and for each parameter. The confidence levels are categorized into three: high, moderate, and low.

At the core of this prompt is a detailed scoring rubric, which, based on the Environmental Sustainability Matrix, breaks down the complex ESG assessment task into 11 specific parameters and provides clear, graded scoring criteria for each. Here is an example:

> [Waste Management]: it has 2 points available in total: 2 points available if a club has put in place a waste management/ recycling programme that reduces waste, diverts at least 98% of waste from landfill and ensures all waste is recycled/works within the circular economy across all sites - stadium, training facilities, offices and retail stores; 1 point given if clubs have a waste diversion/recycling system in place but it doesn't lead to 98% + diversion from landfill, doesn't operate across all sites OR if zero waste to landfill but no waste management policy or recycling system not in place; 0.5 points if some recycling takes place but no waste management strategy in place; 0 points if a club does not have a waste management programme or attempt to recycle or divert waste from landfill.

To avoid data contamination during the experiment, we set rules in the prompt to ignore previous ChatGPT ratings and run the prompt separately in each round of the rating process.

Additionally, to avoid data contamination, we included rules in the prompt for the model to ignore any previous ratings and executed each evaluation in a separate session.

> Please disregard the previous results and rerate the report according to the following requirements.

We set a strict information source constraint: "Do not search for related information from other sources; only score based on the report." These two instructions ensure a closed evaluation environment, eliminating contamination from external knowledge and enabling our study to purely measure each model's ability to understand, extract, and evaluate a given text.

Taken together, this prompt design, which combines role-playing, guided chain-of-thought reasoning, and reliability constraints, provides a solid methodological foundation for using LLMs in professional financial sustainability analysis. Finally, for each experiment, we uploaded the sustainability report, and the respective model generated the outcomes in a predefined format. (The full version of the prompt used for all models is detailed in Appendix B.)

3.5. Model Parameters

We set the model temperature to 0.5, as it offered the best trade-off by maximizing predictive accuracy without sacrificing the critical output reliability that degraded at higher settings [63]. This approach is supported by studies showing that for qualitative coding tasks, accuracy gains are most reliable at a temperature of 0.5 or lower, whereas higher temperatures can lead to factual inaccuracies and hallucinations [64–66]. To account for the inherent randomness in the model's responses, we conduct five separate ratings for each report and use a carefully designed prompt to eliminate the influence of conversational history.

3.6. Validity

To evaluate the validity of the automated scores generated by the LLM, we conducted a multi-faceted analysis comparing them against the scores from human annotations, which serve as the ground truth. Firstly, we evaluate the validity by comparing the automated scores generated by the LLM with those generated by human annotations. Unlike previous studies that use metrics such as accuracy, precision, recall, and F1-score calculated from a confusion matrix, we rely on a direct comparison approach. We assume that human scores are the ground truth, and we consider the LLM's scores correct if they match the human scores; otherwise, they are considered incorrect.

The rationale for not using a confusion matrix lies in the fundamental nature of our scoring system. Confusion matrices, and the metrics derived from them such as precision, recall, and F1-score, are primarily designed for binary or multiclass classification problems where the objective is to categorize items into distinct, non-ordered classes [67,68]. Our task, however, involves a more nuanced scoring system with five distinct levels spanning the range of 0 to 3 (specifically 0, 0.5, 1, 2, 3), making the application of traditional classification metrics less appropriate. Additionally, using a direct comparison allows us to focus on the practical alignment between human and LLM's assessments, which is crucial for validating the LLM's performance in this specific context.

Recognizing that the degree of difference between scores is also a critical indicator of performance, we calculate the Mean Absolute Error (MAE). The MAE measures the average absolute difference between the LLM's scores and the human scores across all items. This analysis quantifies the average magnitude of error, offering a direct and interpretable view of how much the LLM's scores typically deviate from the ground truth.

By reporting both Strict Accuracy and Mean Absolute Error, our validity assessment offers a dual perspective: it captures the rate of perfect alignment while also providing a transparent measure of the average discrepancy when scores are not identical. This approach ensures a thorough and balanced evaluation of the LLM's performance in this specific task.

4. Results

We first assess the performance of three models (GPT-4o, Qwen-2-72b-instruct, and Llama-3-70b-instruct) across multiple dimensions: task completion rate, accuracy, overall mean absolute error (MAE), and the standard deviation of MAE (as an indicator of stability).

As shown in Figure 3, GPT-4o delivers the strongest overall performance with top scores in task completion, accuracy, and both error metrics. It achieves a 100% task completion rate, ensuring fully generated outputs for each prompt. Its accuracy reaches 0.55, the highest among the models. Moreover, GPT-4o demonstrates the lowest overall MAE at 0.60, meaning its predictions are consistently close to the expected values. Its stability is also top tier, with MAE standard deviation (Stability) being the lowest at 0.05, indicating highly reliable performance.

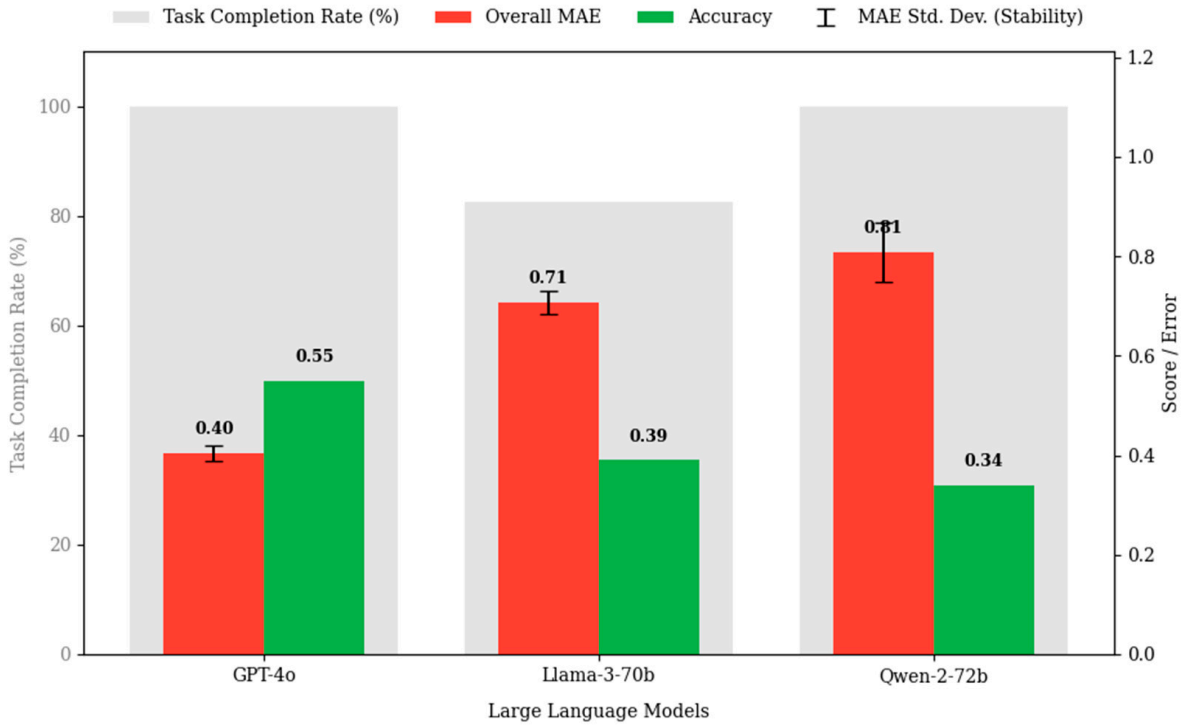

Figure 3. Comprehensive performance comparison of LLMs on sustainability scoring. Note: This chart compares the performance of three LLMs (GPT-4o, Llama-3-70b-instruct, and Qwen-2-72b-instructon) on an ESG scoring task.

While Qwen-2-72b-instructon proves highly reliable in completing tasks, it struggles with prediction precision. The accuracy is lower at 0.34, and it has the highest overall MAE of 0.91, suggesting significant deviation from expected scores. Its stability, with a MAE std. dev of 0.08, while not the worst, still reflects moderate variability in performance; it also achieves a 100% task completion rate.

Llama-3-70b-instruct's biggest drawback lies in output completeness and stability (it experiences a substantial drop-in task completion rate to 82.4%), not accuracy per se, it achieves an intermediate accuracy of 0.39, better than Qwen-2-72b-instructon. Its overall MAE is 0.71, and its stability measure is the worst among the three with an MAE std. dev of 0.11, signaling less consistent outputs even when it does respond.

Figure 4 further displays the three LLM's performance across 11 ESG parameters, and the Mean Absolute Error (MAE), where a lower score indicates higher accuracy. The results reveal distinct capability profiles for each model.

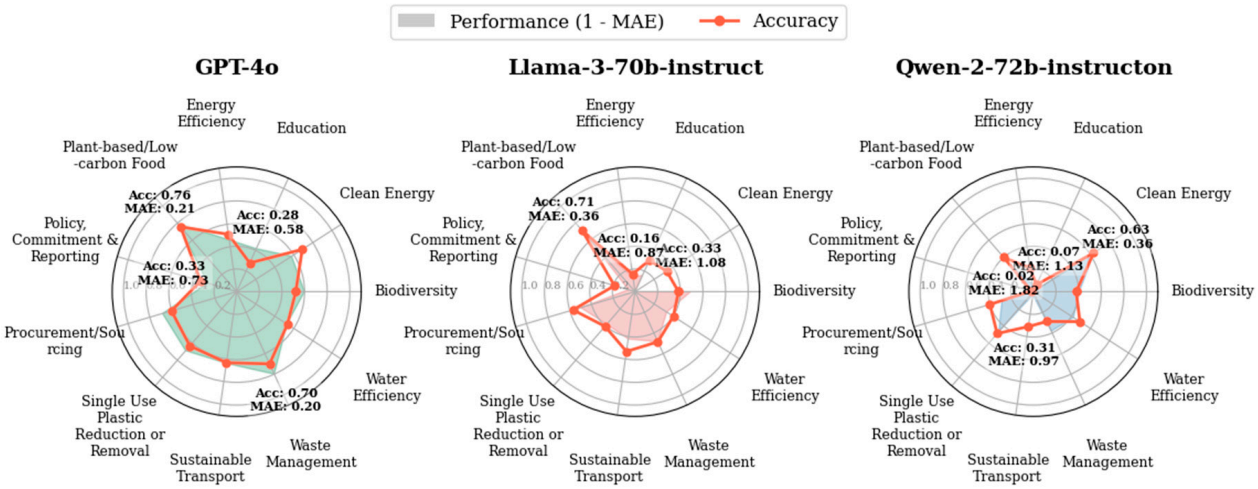

Figure 4. LLMs performance: Each sub-topic accuracy of the three LLMs. Note: This figure compares the performance of three models across 11 sustainability sub-themes. Each model's radar chart displays two metrics: the orange line represents accuracy, and the shaded area represents performance (1-MAE).

The results indicate that GPT-4o consistently outperforms the other models, achieving the highest accuracy in 10 of the 11 assessed parameters. Its efficacy is particularly pronounced in the categories of "Plant-based/Low-carbon Food," where it achieved the highest accuracy score of the entire evaluation (0.756), and "Waste Management" (0.7). In the latter, GPT-4o also recorded the lowest MAE (0.2), signifying a high degree of precision in its assessments. Further evidence of its robust performance is seen in categories such as "Clean Energy" (0.689), "sustainable Transport" (0.633), and "Procurement/Sourcing" (0.6), where it also secured the top accuracy scores.

In contrast, the performance of Llama-3-70b-instruct and Qwen-2-72b-instructon is more variable. While generally lagging behind GPT-4o, Llama-3-70b-instruct demonstrated strong competency in specific areas. It achieved its highest accuracy (0.714) in the "Plant-based/Low-carbon Food" category and was the sole model to outperform GPT-4o in any parameter, securing the highest score in "Education" with an accuracy of 0.304.

The Qwen-2-72b-instructon exhibits the most uneven performance. It achieved a respectable accuracy score in "Clean Energy" (0.633), surpassing Llama-3-70b-instruct in that specific domain. However, its effectiveness was significantly lower in other areas, such as "Policy, Commitment & Reporting" (0.022) and "Education" (0.067), where its accuracy is markedly below that of the other two models.

4.1. LLM's Confidence

Kim et al. indicates that the model performs better when it reports higher confidence [24]. Figure 5 shows the confidence calibration results for three large-scale language models, Qwen-2-72b-instructon, Llama-3-70b-instruct, and GPT-4o, on the ESG scoring task. The results show that GPT-4o exhibits good calibration: its accuracy is positively correlated with confidence (ranging from 68% to 39%), and its MAE remains stable and low across all confidence levels (approximately 0.41–0.43). A supporting t-test confirmed this; for 4 of 11 ESG parameters, including key areas including Clean Energy, Single Use Plastic Reduction or Removal, Biodiversity and Procurement/Sourcing, scores at high confidence were significantly different from those at low confidence ($p < 0.05$).

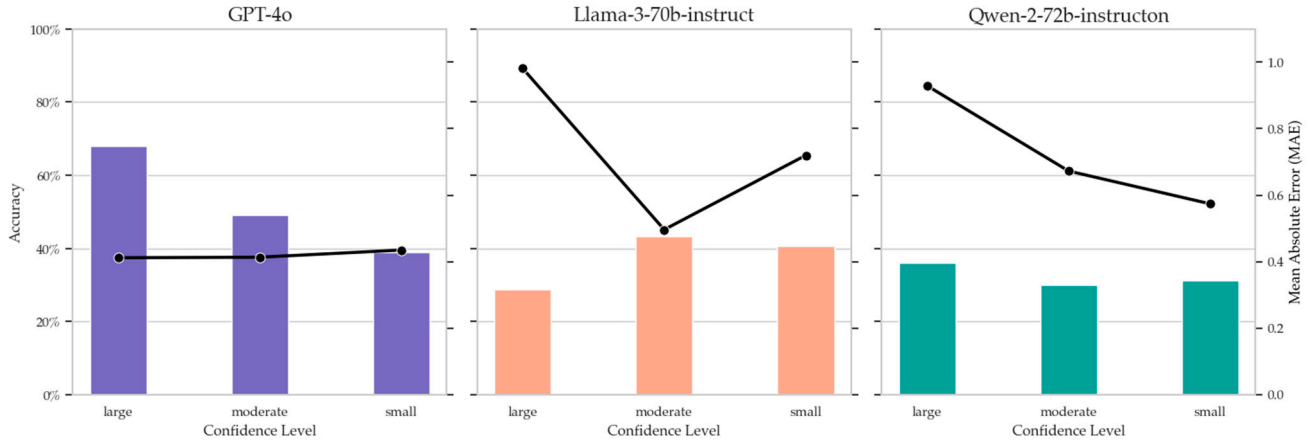

Figure 5. Confidence level of three LLMs. Note: The figure compares the performance of GPT-4o, Llama-3-70b-instruct, and Qwen-2-72b-instructon across three self-reported confidence levels (large, moderate, and small). For each model, the bar chart (left *Y*-axis) displays the Accuracy, while the black line chart (right *Y*-axis) represents the MAE.

In contrast, Qwen-2-72b-instructon and Llama-3-70b-instruct exhibit significant miscalibration. Their accuracy is lowest at high confidence levels (36% and 29%, respectively), while their MAE values peak at 0.93 and 0.98, respectively, indicating a dangerous overconfidence bias. Further t-test analysis underscored this miscalibration. For Qwen-2-72b-instructon, the score differences between confidence groups were not statistically significant for 9 of the 11 parameters (all $p > 0.1$), including Waste Management, Water Efficiency, and Biodiversity, suggesting its confidence levels are largely arbitrary.

Llama-3-70b-instruct showed a more complex disconnect between its confidence levels and its accuracy. For instance, high-confidence scores were paradoxically worse than lower-confidence ones for Waste Management (high vs. low confidence; $p < 0.05$), Education (high vs. low confidence; $p < 0.01$), Energy Efficiency (high or low vs. moderate confidence; $p < 0.01$), and Sustainable Transport (high vs. moderate or low confidence; $p < 0.01$). This inconsistent behavior was also manifested in Procurement/Sourcing, where low-confidence scores were significantly superior to moderate-confidence ones ($p < 0.01$), and in Biodiversity, where confidence had no statistical bearing on performance (all $p > 0.05$).

4.2. LLM's Language Preference

The development of LLMs depends heavily on extensive text corpora, which are often unevenly distributed across different languages [69]. As a result, there is a notable disparity in the inference capabilities of LLMs between English and non-English languages [70]. We further investigate the accuracy of the LLMs in scoring English and non-English reports, as shown in Figure 6. To statistically validate the observed differences, we conducted a series of independent two-sample *t*-tests.

The performance of three large language models on multilingual ESG text scoring accuracy reveals a clear hierarchy (all $p < 0.01$). GPT-4o demonstrates a significant lead across all tested languages. It achieves its highest accuracy on English texts at 0.63, followed by a strong performance in Spanish at 0.59. While German is its relatively weakest language, its score of 0.46 is still substantially higher than the other two models.

In contrast, both Llama-3-70b-instruct and Qwen-2-72b-instruct lag considerably behind GPT-4o. Llama-3-70b-instruct achieved its highest score of 0.47 in English, followed by 0.42 in Spanish and 0.26 in German. Qwen-2-72b-instruct showed its best performance in English at 0.45, while its scores for German and Spanish were both 0.28. While a simple comparison of average scores might suggest Llama-3-70b-instruct is the runner-up, our statistical analysis provides a more nuanced picture. Llama-3's advantage over Qwen-2-72b-

instruct is only statistically significant in Spanish ($p < 0.01$). In both English ($p < 0.01$) and German ($p < 0.01$), their performance differences are not statistically significant, suggesting they are on a comparable level in these languages.

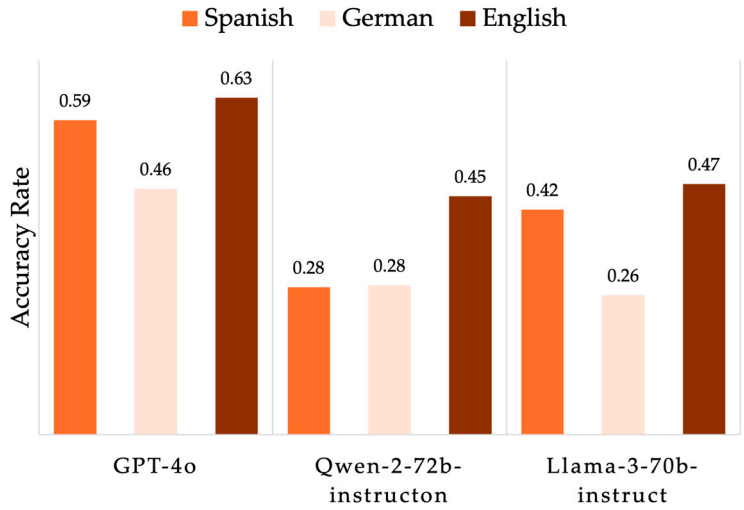

Figure 6. Overall performance comparison of LLMs on multilingual sustainability scoring accuracy. Note: This chart compares the scoring accuracy of three LLMs on sustainability texts across three different languages: Spanish, German, and English.

4.3. Comparison of LLM's Cost and Accuracy: UI vs. API

We further evaluate large language models based on their access method, contrasting the user interface (UI) with the application programming interface (API) in terms of cost and accuracy.

The pricing structures for APIs are predicated on token consumption. For instance, the GPT-4o model's API usage is priced at $5.00 per million input tokens and $15.00 per million output tokens. In comparison, alternative models present different economic considerations. Alibaba's open-source model, Qwen-2-72b-instruct, is priced at ¥0.004 per 1000 input tokens and ¥0.012 per 1000 output tokens, which equates to approximately $0.55 and $1.66 per million tokens, respectively. As our local hardware was insufficient for deployment, we utilized the third-party API provider OpenRouter.ai to access Meta's Llama-3-70b-instruct model. This model is available on the platform at a rate of $0.30 per million input tokens and $0.40 per million output tokens, albeit with a more constrained context window of 8192 tokens. This variation illustrates a clear trade-off among cost, context capacity, and API throughput, as the Qwen-2-72b-instruct model imposes lower rate limits of 60 Requests Per Minute (RPM) and 150,000 Tokens Per Minute (TPM). However, it is typically more expensive and requires technical knowledge to implement effectively. For our specific experimental workload, this pricing translates to a cost of approximately $6.36 for Qwen-2-72b-instructon and $57.60 for GPT-4o, compared to approximately $5.25 for Llama-3-70b-instruct via OpenRouter.ai.

Conversely, the UI typically operates on a fixed-rate subscription model, such as the $20 monthly fee for a ChatGPT Plus subscription. The UI provides greater accessibility, facilitating direct interaction for users without technical expertise [6]. However, this modality presents certain limitations, including a restricted context window (e.g., 128 k tokens for GPT-4o) and the necessity for manual data entry and retrieval. It is also noteworthy that open-source models, while accessible via APIs, generally do not offer a dedicated, first-party user interface.

To empirically evaluate the performance disparity, identical prompts and reports were submitted to both the UI and the GPT-4o API across five scoring rounds. The results indicate

that the accuracy achieved via the UI was 48.0%. This figure is substantially lower than the accuracy observed with the API, reflecting a performance deficit of 8 percentage points.

5. Hallucination in LLMs

Despite the impressive capabilities of LLMs trained on large text corpora, recent studies indicate that LLMs are prone to hallucinations in various applications [71,72]. Hallucination refers to the generation of seemingly reasonable but incorrect or irrelevant information by artificial intelligence, caused by inherent biases, lack of real-world understanding, or limitations of training data [72–74]. These hallucinations result in ratings that either conflict with existing sources or cannot be verified using available knowledge resources, posing potential risks when applying LLMs in real-world rating scenarios.

Current research mainly focuses on understanding the causes of hallucinations in specific tasks and smaller language models [75,76]. For instance, Alkaissi (2023) finds that ChatGPT sometimes creates information, data, and statistics without a reliable basis, even if the required information is not within its sources [77]. This can lead to the fabrication of facts, invention of plots, and even provision erroneous medical explanations. While ChatGPT can assist in writing credible scientific papers, the data it generates may be a mixture of real and fabricated information [77]. Therefore, concerns have been raised about the accuracy and integrity of using LLMs like ChatGPT in academic writing. Alkaissi (2023) also concludes that researchers remain divided on the use of LLMs in scientific writing, as it may mislead individuals who lack of real-life experience and lead to the generation of questionable opinions [77].

According to Huang et al., hallucinations are classified into two main categories: factual hallucinations and faithful hallucinations [78]. Factual hallucinations emphasize the discrepancy between the generated content and verifiable real-world facts, often manifested as inconsistent or fabricated facts. Faithful hallucinations, on the other hand, refer to the differences between the generated content and the context provided by the user's instructions or input, as well as the internal consistency of the generated content.

Therefore, we categorize the corresponding incorrect answers into two types: first, when LLMs provide scores that are not in the parameter settings; second, when unverifiable information is found in the content. Table 1 presents examples of responses from LLMs that exhibit hallucinatory characteristics.

In our study, GPT-4o API (16.6% total) GPT-4o user interface (18%) and Llama-3-70b-instruct (16.3% total) exhibit similar overall hallucination rates, Qwen-2-72b-instructon demonstrates a significantly higher rate at 33.4%. More critically, the nature of these hallucinations differs profoundly between models. For GPT-4o and Llama-3-70b-instruct, errors are overwhelmingly classified as faithful hallucinations (15.6% and 14.1%, respectively), with factual hallucination rates being exceptionally low (1.0% and 2.2%). This suggests that while these state-of-the-art models are highly effective at remaining grounded in the provided source material, they struggle with adhering to the precise, structured requirements of the rating task.

In stark contrast, Qwen-2-72b-instructon exhibits high rates of both factual (13.1%) and faithful (20.3%) hallucinations. Its propensity for factual hallucination is an order of magnitude higher than its counterparts, indicating a greater tendency to invent or fabricate information not present in the reports. This is particularly evident in the "Policy, Commitment & Reporting" category, where Qwen-2-72b-instructon's factual hallucination rate reached a remarkable 68.9%.

Table 1. Examples of the LLMs hallucination.

Example 1 Rating standard on Club C [Water efficiency]
[Water efficiency]: The total score is 2 points, with scoring options of 2 points, 1 point, and 0 points.
Human annotation
Score: 1 point Explanation: Rainwater recycling systems can recover wastewater from irrigation.
LLMs rating
Score: 0.5 points Explanation: The report mentions the use of rainwater for irrigation, indicating some efforts in water efficiency.
Wrong Reason: Incorrect score criteria.
Example 2 Rating standard on Club B [policy and commitment]
[Policy, Commitment & Reporting]: The total score is 3 points, with scoring options of 2 points, 1 point, 0 points, and 1 bonus point.
Human annotation
Score: 0 points Explanation: The report mentioned nothing related.
LLMs rating
Score: 1 point Explanation: The report mentions a commitment to sustainability, but there is no evidence of being a signatory to the UN Sports for Climate Action Framework or public reporting on emissions. No mention of ISO20121 certification (The detailed information is: ISO 20121:2024, *Event sustainability management systems—Requirements with guidance for use*, which was released by the International Organization for Standardization (ISO, Geneva, Switzerland) in 2024.).
Wrong Reason: unverified information.

The occurrence of hallucinations can be attributed to various factors, such as exposure bias and belief misalignment. During training, the model is provided with ground truth sequences; however, during inference, it relies on its own generated tokens for subsequent predictions. This inconsistency can lead to hallucinations [79], especially when an erroneous token generated by the model cascades errors throughout the subsequent sequence, creating a snowball effect [10].

Furthermore, Figure 7 represents a comparative performance analysis of three LLMs, evaluated on their ability to process information across eleven sustainability-related categories. The evaluation is quantified using two critical metrics: Accuracy, which measures the correctness of the model's output, and Hallucination Rate, which measures the frequency of generating factually incorrect or non-verifiable information. The findings reveal significant performance heterogeneity across both models and categories, indicating that no single model achieves universal superiority. Overall, GPT-4o emerges as the leader in aggregate accuracy, while Llama-3-70b-instruct demonstrates remarkable reliability in specific domains by minimizing hallucinations.

A detailed examination of GPT-4o's performance shows its strong capabilities in several areas, achieving the highest accuracy in categories such as "Plant-based/Low-carbon Food" (76%), "Waste Management" (70%), and "Clean Energy" (69%). This suggests a high degree of proficiency in extracting and classifying well-defined factual information. However, the model's performance deteriorates significantly when confronted with more abstract or nuanced topics. Its lowest accuracy scores are observed in the "Education"

(28%) and "Policy, Commitment & Reporting" (33%) categories, which are concurrently associated with its highest hallucination rates (54.4% and 34.4%, respectively). This indicates a challenge in maintaining factual integrity when processing complex, interpretative texts that may lack standardized structure.

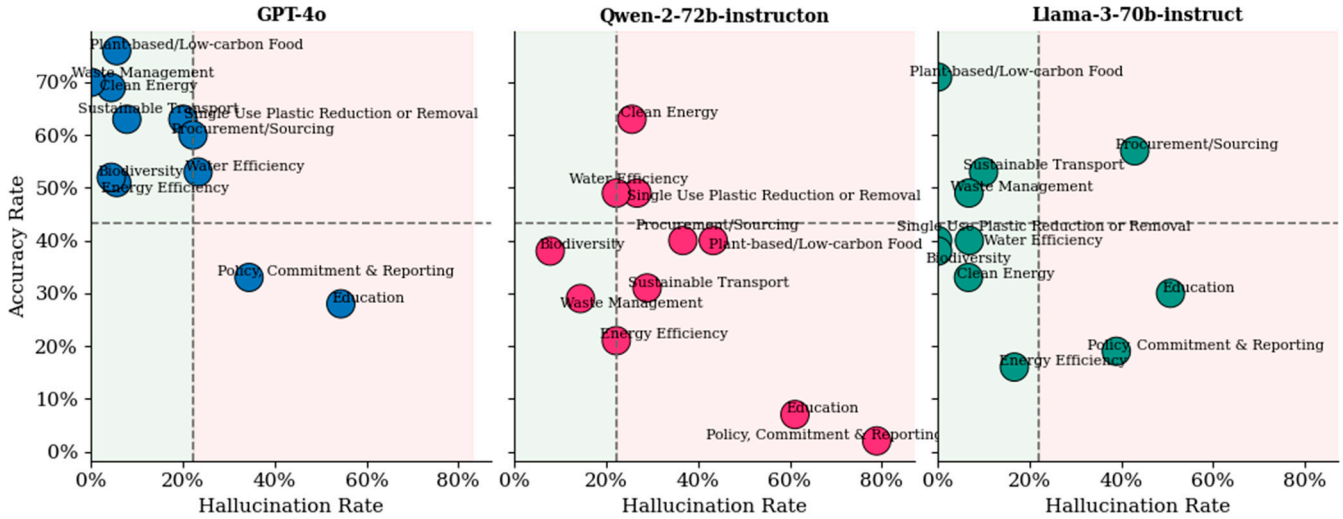

Figure 7. Accuracy rate and hallucination rate for each sub-topic. Notes: This figure shows the average value of accuracy and hallucination ratings in each sub-topic; and the average line of the overall accuracy rate and hallucination rate.

In contrast, Llama-3-70b-instruct and Qwen-2-72b-instructon present different performance profiles. Llama-3-70b-instruct positions itself as a moderately accurate but highly reliable model. While its accuracy in "Plant-based/Low-carbon Food" (71%) is competitive with GPT-4o, its most distinguishing feature is achieving a zero-percent hallucination rate in the "Single Use Plastic Reduction or Removal," "Biodiversity," and "Plant-based/Low-carbon Food" categories. This suggests a robust mechanism for suppressing unfounded assertions in well-scoped domains. Conversely, Qwen-2-72b-instructon consistently underperforms relative to its counterparts, exhibiting the lowest accuracy and the highest hallucination rates in most categories. Its profound difficulty is most evident in the "Policy, Commitment & Reporting" category, where it scored a nominal 2% accuracy with a 78.9% hallucination rate, highlighting significant limitations in its current capacity for this type of semantic task.

Synthesizing the results across all three models reveals critical insights into the current state of LLM capabilities for sustainability analytics. A discernible inverse correlation exists between accuracy and hallucination, where lower performance on a task is often coupled with a higher propensity to generate fallacious content. Furthermore, the "Policy, Commitment & Reporting" and "Education" categories consistently prove to be the most challenging, suggesting that tasks requiring deep contextual understanding, interpretation of ambiguous language, and synthesis of non-standardized information remain a frontier for LLM development. These findings underscore the necessity of a task-specific approach to model selection, where the choice between maximizing accuracy GPT-4o and ensuring factual reliability Llama-3-70b-instruct becomes a critical decision contingent on the application's tolerance for error.

Additionally, several studies have shown that LLMs' activations encapsulate an internal belief related to the truthfulness of their generated statements [80]. However, misalignment can arise between these internal beliefs and the generated outputs. Even when LLMs are refined with human feedback [32], they can sometimes produce outputs that diverge from their internal beliefs. Such behavior, termed sycophancy [81], highlights the model's

tendency to appease human evaluators, often at the expense of truthfulness. Researchers should carefully distinguish the flattery behavior of ChatGPT.

6. Further Research

In this scenario, we are exploring methods to address the issue of hallucinations in LLMs to improve their performance. Our strategy centers around two key approaches. First, we translate the data into English to align with the model's training data to minimizing potential biases in the data. Second, we incorporate more rigorous validation procedures by utilizing the Chain-of-Verification (CoVe) method to ensure the accuracy of feedback, as outlined by [82]. This method comprises four main steps:

(1) Generate an initial response to a given question.
(2) Create a list of verification questions to self-check the original response for errors.
(3) Answer each verification question and compare it to the original response to identify inconsistencies or mistakes.
(4) Produce a final, revised response incorporating the verification results.

Each step involves presenting the same problem in different ways. Figure 8 provides a visual overview of this approach in our experiment, using "Policy, Commitment, and Reporting" as the example.

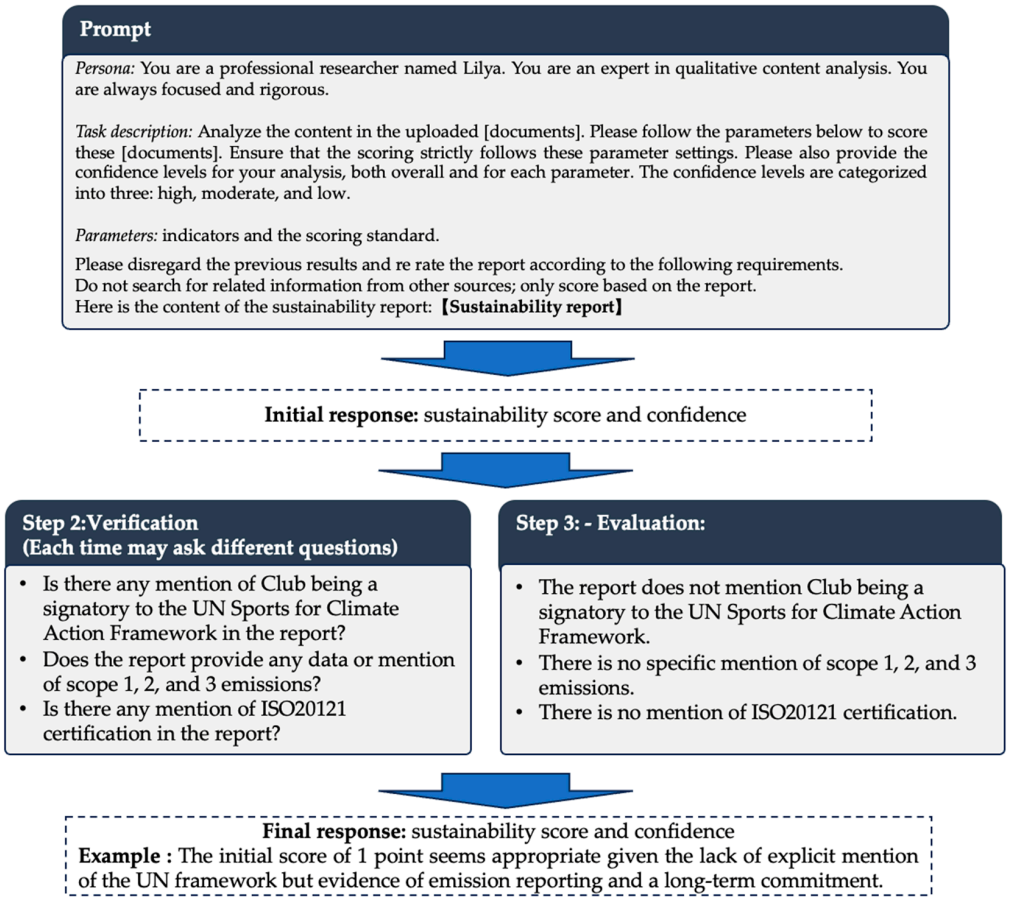

Figure 8. Using CoVe method in the designed experiment. Note: This figure illustrates the Chain-of-Verification (CoVe) method. In this process, a large language model first generates a baseline response to a question, which may contain hallucinations. We then demonstrate a factual example of how CoVe works in our experiment.

After the self-verification of LLM, the accuracy rate increased from 0.56 to 0.58, and the hallucination rate reduced from 16% to 10%. When considering a deviation margin

of 0.5 points in the accuracy assessment, the overall deviation accuracy increased from 0.67 to 0.69. Breaking this down further, the hallucination rates in the English text drop from 0.14 to 0.09, non-English 0.18 to 0.12. This suggests that CoVe is effectively enhancing the model's performance, making it more reliable, with fewer errors and more accurate predictions. It also represents the LLM's potential capabilities in the sustainability rating.

In the future, we will keep looking to improve the performance of LLM in sustainability rating. First, combining human evaluation and automated evaluation methods to identify and correct hallucinations in model outputs, the authenticity and reliability of the model can be more comprehensively detected through multiple evaluation criteria. Then, targeting adjustments to the prompts are then made based on the model's response characteristics to enhance the performance of the LLM in scoring. Finally, strengthen research on model transparency and interpretability to help understand the process of generating text by the model, thereby more effectively identifying and correcting hallucinations.

7. Conclusions

This paper applies large language models in the field of sustainable development, offering a novel approach to analyzing and interpreting sustainability reports. Traditionally, Rating sustainability reports is a cost-intensive and time-consuming task and requires extensive expertise and resources. By automating the assessment of sustainability disclosures, LLMs can significantly reduce the time and effort required for traditional evaluation methods. This approach enhances the ability of diverse stakeholders, including organizations and the public, to understand a company's sustainability performance, enabling them to make more informed decisions and support genuine sustainability initiatives.

Our results clearly demonstrate that among the models tested, GPT-4o demonstrates the strongest overall ability in identifying sustainability topics and assessing relevant disclosures, achieving an overall accuracy of 55% and a mean absolute error (MAE) of 0.60. The model performs particularly well on topics involving quantitative data, achieving its highest accuracy scores in categories like "Plant-based/Low-carbon Food" and "Waste Management." Conversely, its performance degrades significantly on purely qualitative or more abstract topics, with accuracy dropping to its lowest points in "Policy, Commitment & Reporting" and "Education". This stark performance gap suggests that LLMs currently complement, rather than replace, human analysts. GPT-4o's current accuracy is insufficient to independently complete the scoring task and still requires human supervision and verification.

More importantly, the comparative analysis in this study revealed key differences between different models. Compared to GPT-4o, Llama-3-70b-instruct demonstrates moderate accuracy but suffered from severe stability issues. Qwen-2-72b-instructon, while achieving a high task completion rate, had the lowest accuracy and exhibited a dangerous tendency toward overconfidence—its accuracy is lowest at its highest levels of confidence. Furthermore, we find that hallucinations are a common challenge faced by all models. GPT-4o has a hallucination rate of approximately 16.6%, primarily manifesting as faithfulness hallucinations (failure to strictly follow instructions) rather than factual errors. In contrast, Qwen-2-72b-instructon not only had the highest hallucination rate (33.4%) but also contained a significant number of factual hallucinations (fabricated information), which is unacceptable in professional evaluation scenarios. These findings highlight the importance of careful model selection in practical applications.

While these findings establish a novel methodological framework, they must be interpreted within the context of the study's specific scope, which in turn defines clear directions for future research. First, this research is designed to prioritize analytical depth over statistical breadth. Our deep, multi-faceted diagnostic of failure modes, including confidence

calibration, linguistic bias, and hallucination types, necessitated a focused analysis on a limited sample of reports. While this approach provides an unprecedented, granular view into model behavior, the sample size is not intended for broad statistical generalization. Therefore, the performance differences observed should be seen as indicative rather than statistically definitive. The core contribution lies in the pioneering methodological benchmark for how to conduct such a deep evaluation, rather than a large-scale performance ranking. Second, our choice of professional sports clubs is a deliberate strategy to enable a controlled and replicable assessment against a highly structured, industry-specific scoring matrix. This specificity is a core strength of our research design, allowing us to move beyond the generic frameworks common in prior work. However, this focus naturally bounds the direct generalizability of our findings. The behaviors and failure modes identified are deeply contextualized, highlighting the critical challenge of "domain shift" that future cross-sectoral research must address. Finally, the use of proprietary large language models like GPT-4o, while reflecting the current state-of-the-art, introduces fundamental and well-recognized challenges. Their "black box" nature limits the transparency and interpretability of their decision-making processes. Furthermore, the continuous updates to these models pose a threat to long-term reproducibility, and the risk that inherent biases in their training data could skew evaluation results remains a critical concern for the responsible deployment of these technologies.

Acknowledging these limitations provides clear directions for future research. Indeed, this study's initial experiment with the Verification Chain (CoVe) already points to a promising path. Implementing this self-correcting process improved GPT-4o's accuracy to 58% while significantly reducing the hallucination rate from 16.6% to 10%, providing a solid foundation for future exploration. Therefore, future research should be strategically directed to address these challenges. First, to move beyond the limitations of general-purpose models and the domain-specificity highlighted in our study, research should focus on cross-domain comparisons and domain-adaptive fine-tuning. This will help build models that are not only more accurate within a specific sector but also more robust when applied across different industries. Second, using transparent, open-source LLMs is not only about fostering trust in responsible applications; it provides a critical pathway for fundamentally improving the models' evaluation capabilities. Methodologically, since sustainability reports are inherently multimodal, developing architectures that can interpret graphs and tables is a critical next step. Ultimately, these technological improvements should support a more mature "human-in-the-loop" collaborative framework, where AI assists, rather than replaces, human expertise to ensure the highest standards of accuracy and judgment. In summary, while exploratory in nature, this research reveals the significant potential of large language models in promoting the democratization of sustainable development assessments. To fully unleash this potential, future work must systematically address current technical and methodological challenges to forge a new generation of transparent, reliable, and fair analytical tools that truly serve the global sustainable development goals.

Author Contributions: Conceptualization, Y.W., P.H. and D.D.W.; methodology, P.H.; software, Y.W.; validation, Y.W., P.H. and D.D.W.; formal analysis, P.H.; data curation, Y.W.; writing—original draft preparation, Y.W. and P.H.; visualization, Y.W. and P.H.; supervision, D.D.W.; project administration, D.D.W.; funding acquisition, D.D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China (Grant No. 71974201), the Major Project of the National Social Science Foundation of China (Grant No. 22&ZD145), and the Academic Innovation Team of Capital University of Economics and Business (Grant No. XSCXTD202404).

Data Availability Statement: If original data is required, please contact the corresponding author.

Acknowledgments: We are gratefully Sport Positive's support for this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ESG	Environment, Social and Governance
LLM	Large Language Model
NLP	Natural Language Processing

Appendix A

Table A1. 18 football clubs: sustainability report list.

League	Football Club	Report Title	Published Date	Source
Premier League	Wolverhampton Wanderers	Wolves ENVIRONMENTAL SUSTAINABILITY REPORT 2023/24	2024	https://www.wolves.co.uk/media/rqkpsorz/opop_deck-2.pdf (accessed on 11 May 2024)
	Manchester City	Manchester City's Sustainability & Environmental Impact Report	2023	https://www.mancity.com/meta/media/cgpmcj2l/2023-mcfc-sustainability-report.pdf (accessed on 11 May 2024)
Bundesliga	Vfl Wolfsburg	TOMORROW TOGETHER SUSTAINABILITY REPORT OF VfL WOLFSBURG 2022	2022	https://static-typo3.vfl-wolfsburg.de/user_upload/Medien/Dokumente/221209-nachhaltigkeitsbericht-2022-en-vfl-wolfsburg.pdf (accessed on 11 May 2024)
	Vfl Bochum	ZUKUNFT ALS GEMEINSCHAFT GESTALTEN NACHHALTIGKEITSBERICHT 2021/22	2022	https://backend.vfl-bochum.de/site/binaries/content/assets/pdf/csr/csr-nachhaltigkeitsbericht2021-22doppelseiten.pdf?_gl=1*1atg32q*_ga*OTkzNDg3MTE3LjE3NTgxMzQ2NjU.*_ga_J6WZ9208G8*czE3NTgxMzQ2NjUkbzEkZzAkdDE3NTgxMzQ2NjUkajYwJGwwJGgw (accessed on 11 May 2024)
	FC Augsburg	Zusammen wachsen: Brückenbauer für nachhaltige Entwicklung Fortschrittsbericht 2022/23	2023	https://www.fcaugsburg.de/page/fortschrittsbericht-2022-23-453 (accessed on 11 May 2024)
	SV Werder Bremen	WERDER BEWEGT NACHHALTIGKEITS BERICHT 2022/23	2023	https://www.werder.de/fileadmin/Medienservice/Downloads/Werder_Bremen_Nachhaltigkeitsbericht_2022_2023_ungeschuetzt__1_.pdf (accessed on 11 May 2024)
	Borussia Mönchengladbach	sustainability-report 2022	2023	https://www.borussia.de/fileadmin/user_upload/Nachhaltigkeit/Nachhaltigkeitsbericht/Nachhaltigkeitsbericht_2022.pdf (accessed on 11 May 2024)
	Borussia Dortmund	BORUSSIA VERBINDET. BORUSSIA PACKT AN. Nachhaltigkeitsbericht zur Saison 2022/2023	2023	https://www.bvb.de/de/de/aktuelles/news/news.html/News/Uebersicht/Borussia-verbindet.-Borussia-packt-an-Der-BVB-Nachhaltigkeitsbericht-2022-2023.html (accessed on 11 May 2024)
	SC Freiburg	SC Freiburg NACHHALTIGKEITS-BERICHT 2022/23	2023	https://www.scfreiburg.com/fileadmin/01_Content/01_Bilder/12_Nachhaltigkeit/SCF_Nachhaltigkeitsbericht_2022_23__1_.pdf (accessed on 11 May 2024)
	RB Leipzig	Play. Care. Share. 2022	2023	https://storage.googleapis.com/rbl-neos-target-stage-8389658/a7efe95554212ef7c060e42796677734cba8e861/RBL_Nachhaltigkeitsbericht_2022.pdf (accessed on 11 May 2024)

Table A1. *Cont.*

League	Football Club	Report Title	Published Date	Source
La Liga	RC Celta	ESTADO DE INFORMACIÓN NO FINANCIERA	2023	https://rccelta.es/app/uploads/2024/04/RCCelta-Estado-Informaci%C3%B3n-No-Financiera-22-23-1.pdf (accessed on 11 May 2024)
	Real Madrid	Informe de Sostenibilidad y Responsabilidad Social Corporativa 2022-2023	2023	https://www.realmadrid.com/es-ES/el-club/transparencia/informes-anuales-de-sostenibilidad-y-rsc (accessed on 11 May 2024)
	Atlético de Madrid	Atlético de Madrid Memoria de Sostenibilidad 2020/21	2021	https://www.atleticodemadrid.com/files/20211221_2_V_ATM_Memoria_Sostenibilidad_2021_WEB.pdf (accessed on 11 May 2024)
	Valencia CF	ESTADO INFORMACIÓN NO FINANCIERA 2022-2023	2023	https://www.valenciacf.com/public/Attachment/2024/1/estadoinformacionnofinanciera_2022-2023.pdf (accessed on 11 May 2024)
	FC Barcelona	2020/21 2021/22 SUSTAINABILITY REPORT	2021	https://www.fcbarcelona.com/fcbarcelona/document/2024/03/03/5e5c3910-5d80-4bd1-a32f-644cc8ed51c5/Mem-ria-Sostenibilitat_ANG.pdf (accessed on 11 May 2024)
	Athletic Club	Impacto socioeconómico del Athletic Club-Temporada 2018/19	2019	https://cdn.athletic-club.eus/uploads/2020/07/Impacto-econ%C3%B3mico-Athletic-Club_CAST.pdf (accessed on 11 May 2024)
Serie A	AC Milan	SUSTAINABILITY REPORT 20-21	2021	https://www.acmilan.com/en/club/sustainability/reports (accessed on 11 May 2024)
	Juventus	the impact JUVENTUS 10 YEARS IN SUSTAINABILITY 2022/2023	2023	https://www.juventus.com/en/sustainability/reports (accessed on 11 May 2024)

Appendix B. Prompt Instructs for LLMs

Persona: You are a professional researcher named Lilya. You are an expert in qualitative content analysis. You are always focused and rigorous.

Task description: Analyze the parameters related to [environment, social] in the uploaded contents in the Sustainable Report. Please follow the parameters below to score this sustainable report. Please also provide a confidence level for your analysis, including total confidence and confidence for each parameter. The confidence levels are categorized into three: high, moderate, and low.

Parameters include:

[Policy, Commitment & Reporting]: it has 2 points available + 1 bonus point: 2 points available if club has a published sustainability policy/strategy that shows commitment to long term, holistic environmental sustainability efforts AND if club is a signatory to UN Sports for Climate Action Framework on high ambition track with net zero targets & are reporting publicly on their scope 1, 2 and 3 emissions OR have set net zero targets and are reporting publicly their scope 1, 2 and 3 emissions; 1 point available if club has a published sustainability policy/strategy that shows commitment to long term, holistic sustainability efforts, and if club is a signatory to UN Sports for Climate Action Framework AND/OR has made an external net zero or credible emissions reduction commitment; 0.5 point if the club has set an external emissions reduction target but has not yet got a policy/strategy in place towards that OR if a club has a sustainability policy/pledge but has not made any external commitments on emissions reduction targets; 0 points if the club has neither policies on environmental sustainability in place nor an externally published emissions reduction target; 1 bonus point—if in addition to the above, the club is certified to internationally recognized sustainability management system, such as ISO20121.

[Clean Energy]: it has 2 points available in total: 2 points given if 100% of energy at stadium and all other club sites inc. all retail stores is from a renewable source (via utility or mix of utility and onsite generation)—proof required; 1 point for more than 75% of energy being provided from renewable source across all clubs sites, but less than 100%, or for having any onsite generation—proof required; 0.5 points given if club has some energy provided from renewable sources, but not 75% or more (was up to 40% in 2021); 0 points given if club has no energy derived from renewable sources or cannot show that any of their energy is provided via renewable sources.

[Energy Efficiency]: it has 2 points available in total: 2 points given if club has a systemic energy efficiency plan in place across their sites, via building/energy management systems, BREEAM standards, ESOS compliant, etc. 1 point given if multiple energy efficiency efforts have been made across all club sites; 0.5 points given if only one energy efficiency effort in place—i.e., LED lighting; 0 points given if club cannot show that they have any energy efficiency efforts in place.

[Sustainable Transport]: it has 3 points available + 1 bonus point: 3 points available if all criteria for 2 points are met (see below) and if the club can prove that no flights were used for domestic team travel in the last 12 months; 2 points available if club has a sustainable transport policy that extends to staff and team travel, and fans. To include visibly advocating for fans/staff to use sustainable transport options and give incentives to do so—i.e., free travel in fan zones, bike to work scheme, money off public transport, as well as showing a sustainable transport policy for player/team travel to games; 1 point given if clubs actively and visibly advocate for fans and staff to sustainable transport options; public transport, active transport, bike racks, carpooling, etc. 0 points given if clubs don't actively or visibly advocate for fans and staff to travel sustainably; 1 Bonus Point—if club tracks and reports on the percentage of fans taking various modes of transportation to games and reports it/shares the findings publicly.

[Single Use Plastic Reduction or Removal]: it has 2 points available in total: 2 points available if club has entirely removed all single use plastic from across all sites of their organization (inc. retail stores); 1 point available if club has a current policy/systemic effort in place that is actively reducing single use plastic from across all sites of their organisation; 0.5 point available if efforts to remove single use plastic are ad hoc, focused on individual products; 0 points given if clubs have not succeeded in reducing or removing single use plastic from their operations.

[Waste Management]: it has 2 points available in total: 2 points available if a club has put in place a waste management/recycling program that reduces waste, diverts at least 98% of waste from landfill and ensures all waste is recycled/works within the circular economy across all sites—stadium, training facilities, offices and retail stores; 1 point given if clubs have a waste diversion/recycling system in place but it doesn't lead to 98%+ diversion from landfill, doesn't operate across all sites OR if zero waste to landfill but no waste management policy or recycling system not in place; 0.5 points if some recycling takes place but no waste management strategy in place; 0 points if a club does not have a waste management program or attempt to recycle or divert waste from landfill.

[Water Efficiency]: it has 2 points available in total: 2 points available if club has a policy/systemic effort in place that is currently reducing and enabling water reuse from their organization—across stadium, training facilities, offices and retail stores. To include water recycling, reduction and reuse; 1 point available if efforts to conserve/reuse water are isolated across 1 or 2 areas/don't take place across the whole club's operations, 1 point if a strategy is in place but no initiatives started yet; 0 points if a club doesn't currently conserve or recycle water

[Plant-based/Low-carbon Food]: it has 2 points available in total: 2 points available if club offers sustainably sourced, plant based food options across all sites; to fans on the stadium concourse for every game, in hospitality areas and for staff and players across all sites; 1 point given if sustainably sourced plant based food options are available, but not across all sites; 0.5 point given if vegetarian, sustainably sourced food is available at the stadium for fans, or if all foods are sustainably sourced; 0 points given if food is not sourced sustainably, and no plant based food options are available on any sites.

[Biodiversity]: it has 2 points available in total: 2 points available if club has a publicized biodiversity policy/strategy/commitment that reaches across all club sites/in their local community to support nature and local ecosystems through refrain, reduce, restore, renew or similar; 1 point available if club has active efforts to support nature and local ecosystems but doesn't have a policy in place; 0.5 points if club has supported nature and local ecosystems in the past 12 months in an ad-hoc way, but doesn't have current/ongoing activity; 0 points available if club doesn't currently have activity relating to promoting biodiversity or protecting nature.

[Education]: it has 2 points available in total: 2 points available if all criteria for 1 point is met (see below) and club has provided environmental sustainability/climate change education program for ALL players (formal training, not PR support; men and women, academy players, etc.) in the past 12 months; 1 point available if club has provided environmental sustainability/climate change education program or training for staff, fan groups/young people, some players in the past 12 months; 0.5 points available if club has provided environmental sustainability/climate change education program or training in one or more category of staff, players, fan groups or young people at any time; 0 points available if clubs has not provided any environmental sustainability/climate change education programming or training for the stakeholders mentioned.

[Procurement/Sourcing]: it has 2 points available in total: 2 points available if the club has a sustainable sourcing/procurement policy in place for all goods. To include environmental, ethical (human and labor rights, fair/living wages) and social responsibility (diversity, traceability). 1 point available if the club has a sustainable procurement policy in place for goods that includes some but not all of that covered in 2 points (above)—must cover environmental as a minimum;0.5 points if club has taken steps to reduce environmental impact of merchandise in last 12 months but don't have a full procurement policy in place, i.e., rolling kit over from last season, recycled materials in kits, limiting packaging of online merch delivery; 0 points if a club does not have a sustainable procurement policy in place, nor has taken steps to reduce the environmental impact of goods, services or merchandise.

Please disregard the previous results and rerate the report according to the following requirements.

Do not search for related information from other sources; only score based on the report. Here is the content of the sustainability report: [Sustainability report].

References

1. Schaltegger, S.; Bennett, M.; Burritt, R. Sustainability Accounting and Reporting: Development, Linkages and Reflection. An Introduction. In *Sustainability Accounting and Reporting*; Springer: Dordrecht, The Netherlands, 2006; pp. 1–33.
2. Kaplan, A.; Haenlein, M. Rulers of the World, Unite! The Challenges and Opportunities of Artificial Intelligence. *Bus. Horiz.* **2020**, *63*, 37–50. [CrossRef]
3. Sustainability Rate the Raters 2020: Investor Survey and Interview Results. Available online: https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sustainability-ratetheraters2020-report.pdf (accessed on 14 July 2024).
4. Billio, M.; Costola, M.; Hristova, I.; Latino, C.; Pelizzon, L. Inside the ESG Ratings: (Dis)Agreement and Performance. *Corp. Soc. Responsib. Environ. Manag.* **2021**, *28*, 1426–1445. [CrossRef]

5. Hughes, A.; Urban, M.A.; Wójcik, D. Alternative ESG Ratings: How Technological Innovation Is Reshaping Sustainable Investment. *Sustainability* **2021**, *13*, 3551. [CrossRef]
6. OpenAI Hello GPT-4o. 2024. Available online: https://openai.com/index/hello-gpt-4o/ (accessed on 7 July 2024).
7. European Club Association. Sustainability Strategy. 2024. Available online: https://online.flippingbook.com/view/561335824/ (accessed on 7 September 2025).
8. Sport Positive Environmental Sustainability Matrix. Available online: https://www.sportpositiveleagues.com/wp-content/uploads/2023/03/Sport-Positive-Leagues-EPL-Enviro-Sustainability-Matrix-202223-Points-Key.pdf (accessed on 7 July 2024).
9. DesJardine, M.R.; Marti, E.; Durand, R. Why Activist Hedge Funds Target Socially Responsible Firms: The Reaction Costs of Signaling Corporate Social Responsibility. *Acad. Manag. J.* **2021**, *64*, 851–872. [CrossRef]
10. Zhang, M.; Press, O.; Merrill, W.; Liu, A.; Smith, N.A. How Language Model Hallucinations Can Snowball. *arXiv* **2023**, arXiv:2305.13534. [CrossRef]
11. Wong, W.C.; Batten, J.A.; Ahmad, A.H.; Mohamed-Arshad, S.B.; Nordin, S.; Adzis, A.A. Does ESG Certification Add Firm Value? *Financ. Res. Lett.* **2021**, *39*, 101593. [CrossRef]
12. Cellier, A.; Chollet, P. The Effects of Social Ratings on Firm Value. *Res. Int. Bus. Financ.* **2016**, *36*, 656–683. [CrossRef]
13. Gangi, F.; Daniele, L.M.; Varrone, N. How Do Corporate Environmental Policy and Corporate Reputation Affect Risk-adjusted Financial Performance? *Bus. Strategy Environ.* **2020**, *29*, 1975–1991. [CrossRef]
14. Shanaev, S.; Ghimire, B. When ESG Meets AAA: The Effect of ESG Rating Changes on Stock Returns. *Financ. Res. Lett.* **2022**, *46*, 102302. [CrossRef]
15. Berg, F.; Kölbel, J.F.; Rigobon, R. Aggregate Confusion: The Divergence of ESG Ratings. *Rev. Financ.* **2022**, *26*, 1315–1344. [CrossRef]
16. Chatterji, A.K.; Durand, R.; Levine, D.I.; Touboul, S. Do Ratings of Firms Converge? Implications for Managers, Investors and Strategy Researchers. *Strateg. Manag. J.* **2016**, *37*, 1597–1614. [CrossRef]
17. Hou, C.; Zhu, G.; Zheng, J.; Zhang, L.; Huang, X.; Zhong, T.; Li, S.; Du, H.; Ker, C.L. Prompt-Based and Fine-Tuned GPT Models for Context-Dependent and -Independent Deductive Coding in Social Annotation. In Proceedings of the ACM International Conference Proceeding Series, Kyoto, Japan, 18–22 March 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 518–528.
18. Rudolph, J.; Tan, S.; Tan, S. ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education? *J. Appl. Learn. Teach.* **2023**, *6*, 342–363. [CrossRef]
19. Yang, R.; Tan, T.F.; Lu, W.; Thirunavukarasu, A.J.; Ting, D.S.W.; Liu, N. Large Language Models in Health Care: Development, Applications, and Challenges. *Health Care Sci.* **2023**, *2*, 255–263. [CrossRef]
20. Meng, X.; Yan, X.; Zhang, K.; Liu, D.; Cui, X.; Yang, Y.; Zhang, M.; Cao, C.; Wang, J.; Wang, X.; et al. The Application of Large Language Models in Medicine: A Scoping Review. *iScience* **2024**, *27*, 109713. [CrossRef]
21. Yan, L.; Echeverria, V.; Nieto, G.F.; Jin, Y.; Swiecki, Z.; Zhao, L.; Gašević, D.; Martinez-Maldonado, R. Human-AI Collaboration in Thematic Analysis Using ChatGPT: A User Study and Design Recommendations. *arXiv* **2023**, arXiv:2311.03999.
22. Zhai, X. ChatGPT for Next Generation Science Learning. *ACM Mag. Stud.* **2023**, *29*, 42–46. [CrossRef]
23. Huang, A.H.; Wang, H.; Yang, Y. FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemp. Account. Res.* **2023**, *40*, 806–841. [CrossRef]
24. Kim, A.G.; Muhn, M.; Nikolaev, V.V.; Baik, B.; Bradshaw, M.; Dou, Y.; Gassen, J.; Han, S.-Y.; Jain, K.; Koijen, R.; et al. Financial Statement Analysis with Large Language Models. *arXiv* **2024**, arXiv:2407.17866. [CrossRef]
25. Eulerich, M.; Wood, D.A.; Bonrath, A.; Fligge, B.; Krane, R.; Kasper, V.L.; Wagener, M. A Demonstration of How ChatGPT Can Be Used in the Internal Auditing Process. *J. Emerg. Technol. Account.* **2025**, *22*, 47–77. [CrossRef]
26. Föhr, T.L.; Schreyer, M.; Juppe, T.A.; Marten, K.-U. Assuring Sustainable Futures: Auditing Sustainability Reports Using AI Foundation Models. *SSRN Electron. J.* **2023**. [CrossRef]
27. Emett, S.; Eulerich, M.; Lipinski, E.; Prien, N.; Wood, D.A. Leveraging ChatGPT for Enhancing the Internal Audit Process—A Real-World Example from Uniper, a Large Multinational Company. *Account. Horiz.* **2025**, *39*, 125–135. [CrossRef]
28. Cao, Y.; Zhai, J. Bridging the Gap—The Impact of ChatGPT on Financial Research. *J. Chin. Econ. Bus. Stud.* **2023**, *21*, 177–191. [CrossRef]
29. Chen, B.; Wu, Z.; Zhao, R. From Fiction to Fact: The Growing Role of Generative AI in Business and Finance. *J. Chin. Econ. Bus. Stud.* **2023**, *21*, 471–496. [CrossRef]
30. Bernard, D.; Blankespoor, E.; de Kok, T.; Toynbee, S. Confused Readers: A Modular Measure of Business Complexity. *SSRN Electron. J.* **2023**. [CrossRef]
31. Faggioli, G.; Dietz, L.; Clarke, C.L.A.; Demartini, G.; Hagen, M.; Hauff, C.; Kando, N.; Kanoulas, E.; Potthast, M.; Stein, B.; et al. Perspectives on Large Language Models for Relevance Judgment. In Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, Taipei, Taiwan, 23 July 2023; pp. 39–50.

32. Ouyang, L.; Mishkin, P.; Wu, J.; Hilton, J.; Askell, A.; Christiano, P.; Lowe, R. Training Language Models to Follow Instructions with Human Feedback. *Adv. Neural Inf. Process Syst.* **2022**, *35*, 27730–27744.
33. Alizadeh, M.; Kubli, M.; Samei, Z.; Dehghani, S.; Bermeo, J.D.; Korobeynikova, M.; Gilardi, F. Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks. *arXiv* **2023**, arXiv:2307.02179.
34. Huang, F.; Kwak, H.; An, J. Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In Proceedings of the ACM Web Conference 2023—Companion of the World Wide Web Conference, WWW 2023, Austin, TX, USA, 30 April 2023; Association for Computing Machinery, Inc.: New York, NY, USA, 2023; pp. 294–297.
35. Gielens, E.; Sowula, J. Goodbye Human Annotators? Content Analysis of Policy Debates Using ChatGPT. *SocArXiv* **2024**. [CrossRef]
36. Gilardi, F.; Alizadeh, M.I.; Kubli, M.I. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [CrossRef]
37. Chen, P.; Chu, Z.; Zhao, M. The Road to Corporate Sustainability: The Importance of Artificial Intelligence. *Technol. Soc.* **2024**, *76*, 102440. [CrossRef]
38. Zhang, A.Y.; Zhang, J.H. Renovation in Environmental, Social and Governance (ESG) Research: The Application of Machine Learning. *Asian Rev. Account.* **2023**, *32*, 554–572. [CrossRef]
39. Zou, Y.; Shi, M.; Chen, Z.; Deng, Z.; Lei, Z.; Zeng, Z.; Yang, S.; Tong, H.; Xiao, L.; Zhou, W. ESGReveal: An LLM-Based Approach for Extracting Structured Data from ESG Reports. *J. Clean. Prod.* **2023**, *489*, 144572. [CrossRef]
40. El-Haj, M.; Alves, P.; Rayson, P.; Walker, M.; Young, S. Retrieving, Classifying and Analysing Narrative Commentary in Unstructured (Glossy) Annual Reports Published as PDF Files. *Account. Bus. Res.* **2020**, *50*, 6–34. [CrossRef]
41. LI, F. The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. *J. Account. Res.* **2010**, *48*, 1049–1102. [CrossRef]
42. Lin, L.H.-M.; Ting, F.-K.; Chang, T.-J.; Wu, J.-W.; Tsai, R.T.-H. GPT4ESG: Streamlining Environment, Society, and Governance Analysis with Custom AI Models. In Proceedings of the 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), Taipei, Taiwan, 19–21 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 442–446.
43. Kim, A.; Muhn, M.; Nikolaev, V. Bloated Disclosures: Can ChatGPT Help Investors Process Information? *arXiv* **2023**, arXiv:2306.10224. [CrossRef]
44. Loughran, T.; McDonald, B. Textual Analysis in Accounting and Finance: A Survey. *J. Account. Res.* **2016**, *54*, 1187–1230. [CrossRef]
45. Ni, J.; Bingler, J.; Colesanti-Senni, C.; Kraus, M.; Gostlow, G.; Schimanski, T.; Stammbach, D.; Vaghefi, S.A.; Wang, Q.; Webersinke, N.; et al. CHATREPORT: Democratizing Sustainability Disclosure Analysis through LLM-Based Tools. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Singapore, 6–10 December 2023.
46. Bronzini, M.; Nicolini, C.; Lepri, B.; Passerini, A.; Staiano, J. Glitter or Gold? Deriving Structured Insights from Sustainability Reports via Large Language Models. *EPJ Data Sci.* **2023**, *13*, 41. [CrossRef]
47. Moodaley, W.; Telukdarie, A. A Conceptual Framework for Subdomain Specific Pre-Training of Large Language Models for Green Claim Detection. *Eur. J. Sustain. Dev.* **2023**, *12*, 319. [CrossRef]
48. Managi, S.; Shimamura, T.; Tanaka, Y. Evaluating the Impact of Report Readability on ESG Scores: A Generative AI Approach. *Int. Rev. Financ. Anal.* **2024**, *101*, 104027. [CrossRef]
49. Luccioni, A.; Baylor, E.; Duchene, N. Analyzing Sustainability Reports Using Natural Language Processing. *arXiv* **2020**, arXiv:2011.08073. [CrossRef]
50. Bingler, J.A.; Kraus, M.; Leippold, M.; Webersinke, N. Cheap Talk and Cherry-Picking: What ClimateBert Has to Say on Corporate Climate Risk Disclosures. *Financ. Res. Lett.* **2022**, *47*, 102776. [CrossRef]
51. de Villiers, C.; Dimes, R.; Molinari, M. How Will AI Text Generation and Processing Impact Sustainability Reporting? Critical Analysis, a Conceptual Framework and Avenues for Future Research. *Sustain. Account. Manag. Policy J.* **2024**, *15*, 96–118. [CrossRef]
52. Lee, H.; Lee, S.H.; Park, H.; Kim, J.H.; Jung, H.S. ESG2PreEM: Automated ESG Grade Assessment Framework Using Pre-Trained Ensemble Models. *Heliyon* **2024**, *10*, e26404. [CrossRef]
53. Kannan, N.; Seki, Y. Textual Evidence Extraction for ESG Scores. In Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting, Macao, China, 19–25 August 2023; pp. 45–54.
54. Lanza, A.; Bernardini, E.; Faiella, I. Mind the Gap! Machine Learning, ESG Metrics and Sustainable Investment. *SSRN Electron. J.* **2020**. [CrossRef]
55. Meta Llama-3-70B-Instruct. Available online: https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct (accessed on 17 September 2025).
56. Qwen Qwen2-72B-Instruct. Available online: https://huggingface.co/Qwen/Qwen2-72B-Instruct (accessed on 17 September 2025).

57. Drempetic, S.; Klein, C.; Zwergel, B. The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review. *J. Bus. Ethics* **2020**, *167*, 333–360. [CrossRef]
58. Kotsantonis, S.; Serafeim, G. Four Things No One Will Tell You About ESG Data. *J. Appl. Corp. Financ.* **2019**, *31*, 50–58. [CrossRef]
59. Vijayarani, S.; Ilamathi, M.; Nithya, M. Preprocessing Techniques for Text Mining—An Overview. *Int. J. Comput. Sci. Commun. Netw.* **2015**, *5*, 7–16.
60. Clavié, B.; Ciceu, A.; Naylor, F.; Soulié, G.; Brightwell, T. Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification. In Proceedings of the International Conference on Applications of Natural Language to Information Systems, Derby, UK, 21–23 June 2023; Springer Nature: Cham, Switzerland, 2023; pp. 3–17.
61. Ekin, S. Prompt Engineering for ChatGPT: A Quick Guide to Techniques, Tips, and Best Practices. *Authorea Prepr.* **2023**. [CrossRef]
62. Wu, Y.; Hu, G. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In Proceedings of the Eighth Conference on Machine Translation, Singapore, 6–7 December 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 166–169.
63. Windisch, P.; Dennstädt, F.; Koechli, C.; Schröder, C.; Aebersold, D.M.; Förster, R.; Zwahlen, D.R. The Impact of Temperature on Extracting Information from Clinical Trial Publications Using Large Language Models. *Cureus* **2024**, *16*, e75748. [CrossRef]
64. Borchers, C.; Shahrokhian, B.; Balzan, F.; Tajik, E.; Sankaranarayanan, S.; Simon, S. Temperature and Persona Shape LLM Agent Consensus with Minimal Accuracy Gains in Qualitative Coding. *arXiv* **2025**, arXiv:2507.11198. [CrossRef]
65. Davis, J.; Van Bulck, L.; Durieux, B.N.; Lindvall, C. The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research. *JMIR Hum. Factors* **2024**, *11*, e53559. [CrossRef]
66. Liyanage, C.; Gokani, R.; Mago, V. GPT-4 as a Twitter Data Annotator: Unraveling Its Performance on a Stance Classification Task. *Authorea Prepr.* **2023**. [CrossRef]
67. Krstinić, D.; Braović, M.; Šerić, L.; Božić-Štulić, D. Multi-Label Classifier Performance Evaluation with Confusion Matrix. *Comput. Sci. Inf. Technol.* **2020**, *1*, 1–14. [CrossRef]
68. Amin, M.F. Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial. *J. Eng. Res.* **2022**, *6*, 1–12.
69. Li, Z.; Shi, Y.; Liu, Z.; Yang, F.; Liu, N.; Du, M. Quantifying Multilingual Performance of Large Language Models Across Languages. *arXiv* **2024**. [CrossRef]
70. Huang, Z.; Zhu, W.; Cheng, G.; Li, L.; Yuan, F. Mindmerger: Efficiently boosting LLM reasoning in non-english languages. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 34161–34187.
71. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.-L.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination Rates and Reference Accuracy in ChatGPT and Bard for Systematic Reviews: A Comparative Analysis. *J. Med. Internet Res.* **2024**, *26*, e53164. [CrossRef]
72. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
73. Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv* **2023**, arXiv:2302.04023. [CrossRef]
74. Guerreiro, N.M.; Alves, D.M.; Waldendorf, J.; Haddow, B.; Birch, A.; Colombo, P.; Martins, A.F.T. Hallucinations in Large Multilingual Translation Models. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1500–1517. [CrossRef]
75. De Cao, N.; Aziz, W.; Titov, I. Editing Factual Knowledge in Language Models. *arXiv* **2021**, arXiv:2104.08164. [CrossRef]
76. Zheng, S.; Huang, J.; Chang, K.C.-C. Why Does ChatGPT Fall Short in Providing Truthful Answers? *arXiv* **2023**, arXiv:2304.10513. [CrossRef]
77. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, *15*, e35179. [CrossRef]
78. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]
79. Wang, C.; Sennrich, R. On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. *arXiv* **2020**, arXiv:2005.03642. [CrossRef]
80. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It's Lying. *arXiv* **2023**, arXiv:2304.13734.
81. Cotra, A. Why AI Alignment Could Be Hard with Modern Deep Learning. Cold Takes. 2021. Available online: https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/ (accessed on 15 June 2025).
82. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv* **2023**, arXiv:2309.11495.