Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25-29 May 2026,Vancouver,BC

TASR: A Trustworthy LLM-based Framework for TCFD-Aligned Sustainability Report Analysis

Bo Huang†, Sheng Yang†, Wenjun Lin‡, Jing Lu†, Yan Yan †,*
† University of Guelph
‡ Algoma University

Abstract

Reliable and transparent assessment of environmental, social, and governance (ESG) disclosures is critical for sustainable finance, regulatory oversight, and risk-aware decision-making. However, existing sustainability reporting evaluations rely on costly manual reviews or third-party ratings, which limit reproducibility. This work proposes a trustworthy large language model (LLM)-based framework for automated sustainability report analysis aligned with the Task Force on Climate-related Financial Disclosures (TCFD). We propose TASR (Trustworthy Analysis for Sustainability Report), a three-stage framework for TCFD-aligned sustainability report analysis that integrates LLM-based scoring, benchmarking against third-party ESG ratings, and downstream predictive modeling. Experiments on 100 sustainability reports from U.S. oil, gas, and mining companies demonstrate strong alignment with Bloomberg Environmental Disclosure scores and high score stability across repeated evaluations. Furthermore, predictive models trained on the LLM-generated TCFD scores achieve meaningful predictive performance in forecasting disclosure benchmarks, highlighting their practical utility for sustainability rating. The results suggest that LLM-based TCFD scoring offers a potentially scalable and transparent alternative for sustainability disclosure assessment.

Keywords: ESG, Sustainability Reports, TCFD, Large Language Models, Machine Learning, Automated Scoring

1. Introduction

As climate change accelerates and environmental accountability becomes central to corporate governance, sustainability reports have emerged as essential tools for communicating non-financial performance. These reports offer insights into how firms manage environmental, social, and governance (ESG) risks and opportunities. Increasingly, investors, regulators, and stakeholders rely on them to assess climate-related strategies, measure progress toward net-zero goals, and inform responsible investment decisions [1–3]. To ensure comparability and transparency, various frameworks have been developed to structure sustainability disclosures. Among them, the Task Force on Climate-related Financial Disclosures (TCFD) has achieved broad international adoption. Established in 2015 by the Financial Stability Board (FSB), the TCFD provides consistent disclosure standards that enable investors, lenders, and insurers to assess climate-related risks. It recommends that firms disclose along four dimensions: Governance, Strategy, Risk Management, and Metrics & Targets. These pillars address board-level oversight, strategic climate impacts, risk identification and mitigation processes, and measurable indicators and targets respectively. Although the TCFD task force was formally disbanded in 2023, with oversight transferred to the IFRS Foundation, its principles continue to shape major global reporting initiatives, including CDP, SASB, GRI, and IFRS standards.

Despite these standards, evaluating the quality and completeness of sustainability reports remains a significant challenge. Traditional approaches to scoring disclosures rely heavily on human analysts or proprietary third-party ESG rating providers, such as Bloomberg, MSCI, Sustainalytics, or LSEG. These scores play a critical role in shaping investor perceptions

* Corresponding author: Yan Yan (yyan15@uoguelph.ca).

and capital allocation decisions, as they are widely used in portfolio screening, risk assessment, and ESG-focused investment strategies. Firms with higher ESG disclosure scores may benefit from improved access to capital, enhanced reputation, and inclusion in sustainability indices, whereas lower scores may signal greater risk exposure or weaker transparency. However, these providers often apply opaque scoring rubrics and inconsistent methodologies, resulting in considerable variation across ratings [4]. Such a lack of transparency undermines confidence in ESG evaluations and hampers reproducibility in academic and policy-oriented research.

More recently, artificial intelligence, particularly LLMs, has shown promise for automating disclosure analysis at scale. LLMs such as GPT-4 and its compact variants are capable of interpreting natural language, retrieving relevant content, and providing structured assessments with minimal supervision [5]. However, despite these capabilities, concerns remain regarding the reliability, trustworthiness, and consistency of AI-generated scores, especially when applied to high-stakes domains like ESG disclosure [6].

While prior studies have explored using LLMs to interpret ESG-related text [7], few have systematically benchmarked LLM-generated scores against third-party ratings, evaluated their stability across repeated model runs, or assessed their utility for downstream predictive modeling. Moreover, existing approaches rarely incorporate domain expertise into AI system, limiting the interpretability and credibility of LLM-based ESG assessments. As a result, the potential of LLMs to provide scalable, interpretable, and trustworthy sustainability disclosure evaluation remains underexplored.

To address these gaps, this work proposes TASR, a three-stage framework for trustworthy and TCFD-aligned sustainability disclosure analysis. The framework includes: (1) generating eleven TCFD-aligned scores using an LLM-based scoring pipeline; (2) benchmarking these scores against third-party ratings through correlation analysis; (3) evaluating the predictive power of LLM-generated scores using machine learning models.

Key Contributions: (1) We propose TASR, a three-stage framework for trustworthy and TCFD-aligned sustainability disclosure analysis, integrating LLM-based scoring, benchmarking, and downstream predictive modeling. (2) We develop a LLM-based scoring pipeline to extract and evaluate TCFD-aligned information from unstructured sustainability reports, and empirically demonstrate its reliability through repeated evaluations. (3) We benchmark LLM-generated scores against authoritative third-party ESG ratings, demonstrating strong alignment. (4) We demonstrate that LLM-generated TCFD scores can effectively predict ESG disclosure benchmarks using supervised learning models, with tree-based and regularized linear models achieving the strongest performance.

Overall, our work demonstrates that LLM-based tools offer a potentially scalable, interpretable, and replicable alternative to manual ESG evaluations, supporting greater transparency and automation in sustainability assessment.

2. Related Work

The assessment of sustainability reports has traditionally been conducted through manual analysis or third-party ESG rating agencies such as Bloomberg, MSCI, LSEG, and Sustainalytics. These evaluations typically rely on expert judgment to assess the presence, completeness, and quality of disclosed ESG information. However, the underlying scoring criteria and weighting schemes are rarely disclosed, resulting in limited transparency and substantial disagreement across rating providers [4]. Such inconsistencies have raised concerns about the reliability and reproducibility of ESG assessments and motivated calls for more transparent and interpretable evaluation methods [8, 9].

Beyond commercial ratings, academic research has explored computational approaches to analyze sustainability reports at scale. Early methods relied on rule-based or dictionary-driven techniques, such as adapting financial lexicons to identify ESG-related tone, risk signals, or thematic coverage in corporate disclosures [10]. While these approaches improved scalability compared to manual review, they are limited in their ability to capture contextual meaning, nuanced disclosures, or implicit relationships within long-form narrative reports.

To bridge the gap between manual assessment and full automation, some ESG rating agencies and research systems adopted semi-automated pipelines that combine rule-based text parsing, entity recognition, and structured data extraction. These systems can identify predefined indicators or compliance-related disclosures more efficiently than human analysts. Nevertheless, they remain constrained by predefined rules, struggle to generalize across industries or reporting standards, and offer limited explainability regarding how individual disclosure elements contribute to final scores [11]. Overall, these traditional and computational approaches improve efficiency but remain limited by opaque scoring logic, constrained interpretability, and restricted adaptability across heterogeneous reporting formats.

In recent years, artificial intelligence and natural language processing techniques have been increasingly applied to sustainability disclosure analysis. A range of methods has been proposed to automate ESG-related information extraction. For example, Billert and Conrad [12] introduced *Nano-ESG*, which mines sustainability signals from financial news, demonstrating the potential of external data sources to complement corporate disclosures. Similarly, Birti et al. [13] emphasized the importance of prompt tuning and domain adaptation for ESG activity detection in financial texts.

Directly targeting sustainability reports, Bronzini et al. [14] proposed an LLM-based pipeline to derive structured ESG insights from unstructured disclosures, highlighting the ability of large language models to capture nuanced sustainability narratives. However, their approach did not benchmark model outputs against authoritative third-party ESG ratings or assess consistency across repeated evaluations.

In parallel, machine learning and deep learning models have been applied to ESG analysis and sustainability-related prediction tasks. Prior studies explored neural network architectures such as convolutional neural networks and recurrent models for ESG score prediction and index forecasting [15–17]. More recently, transformer-based models, including BERT and RoBERTa variants, have been fine-tuned for ESG classification and disclosure labeling tasks [18–20]. While these models demonstrate improved predictive performance, they typically rely on domain-specific labeled datasets and offer limited interpretability, constraining their suitability for transparent sustainability disclosure assessment. Moreover, prediction-oriented models are less effective for directly evaluating disclosure quality from long-form, unstructured reports.

Advances in autoregressive LLMs, such as GPT-family models, have further expanded the scope of automated sustainability analysis. Recent studies have proposed GPT-based systems to assess sustainability reports or climate transition plans, including ChatReport [7] and related GPT-3.5/4-based pipelines [9, 21, 22]. These works demonstrate the potential of LLMs to reason over long, unstructured disclosures. However, challenges remain in consistency and reproducibility, as LLM outputs may vary across prompts and often lack explicit grounding and systematic benchmarking against authoritative ESG evaluation frameworks.

Other studies have examined broader applications of AI in ESG contexts. Chen [23] and De Villiers et al. [24] highlighted both the opportunities and risks of integrating AI into sustainability reporting, emphasizing concerns related to trust, explainability, and misalignment between AI-generated outputs and real-world ESG performance. These critiques further underscore the need for validated and interpretable AI pipelines.

Despite these promising developments, most existing approaches lack systematic benchmarking against authoritative ESG ratings (e.g., Bloomberg or LSEG), rigorous evaluation of score stability, and explicit incorporation of domain expertise into model design. Consequently, there remains a need for integrated frameworks that combine expert-guided prompting, explainable scoring, and quantitative benchmarking to ensure the reliability and practical utility of AI-driven sustainability disclosure analysis.

3. Methodology

Figure 1 illustrates the overall architecture of TASR and its three sequential stages. In the first stage, the sustainability reports are segmented and analyzed using a GPT-based model to generate eleven scores aligned with the TCFD's four categories. In the second stage, these scores are benchmarked against third-party ESG ratings from Bloomberg and other providers to evaluate alignment and validity. In the final stage, the TCFD-aligned scores are used as input features to train predictive models that estimate ESG scores, demonstrating the utility of LLM-generated disclosures for downstream ESG analytics. Details of each stage are provided in the following subsections.

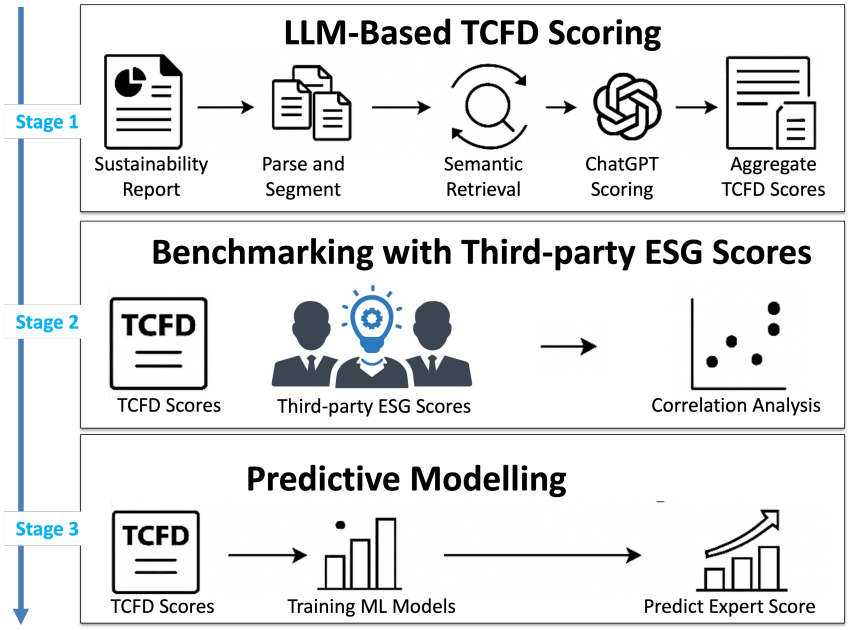


Figure 1. Overview of TASR, a three-stage framework for automated sustainability report analysis. Sustainability reports are evaluated using LLM-based TCFD-aligned scoring, benchmarked against third-party ESG ratings, and used for downstream predictive modeling.

3.1. Data Collection

Our dataset consists of 100 sustainability reports published by 28 U.S. companies in the oil, gas, and mining sectors, covering reporting years from 2009 to 2018. These reports were collected from publicly available sustainability repositories, company investor relations websites, and third-party ESG disclosure platforms.

We focus on the oil, gas, and mining sectors because these industries face heightened climate-related risks and are subject to more stringent disclosure expectations under the

TCFD framework. This sectoral focus enables a controlled evaluation of disclosure quality in contexts where climate-related reporting is both highly material and extensively documented.

The reports span multiple firms and reporting years, capturing temporal variation in disclosure practices. Across the dataset, the LLM-generated disclosure scores exhibit a broad range of values, indicating substantial heterogeneity in reporting completeness and depth. This diversity supports meaningful correlation, stability, and predictive analyses without the results being driven by a narrow subset of firms or reporting periods.

For benchmarking, we collect ESG disclosure and performance scores from three third-party rating agencies. Bloomberg ESG Disclosure Scores, including overall and pillar-level (Environmental, Social, Governance) indicators, measure the extent and completeness of ESG disclosures rather than underlying performance. In contrast, ESG performance scores from LSEG Eikon and Sustainalytics assess firms exposure to and management of financially material ESG risks. These complementary benchmarks enable validation of LLM-generated disclosure scores against both disclosure-focused and performance-oriented ESG indicators.

3.2. **Stage 1: LLM-based TCFD Scoring**

Figure 2 provides an overview of the proposed LLM-based TCFD scoring pipeline. The process consists of document segmentation, semantic retrieval, and question-level scoring using a large language model, designed to extract structured and interpretable disclosure indicators from unstructured sustainability reports.

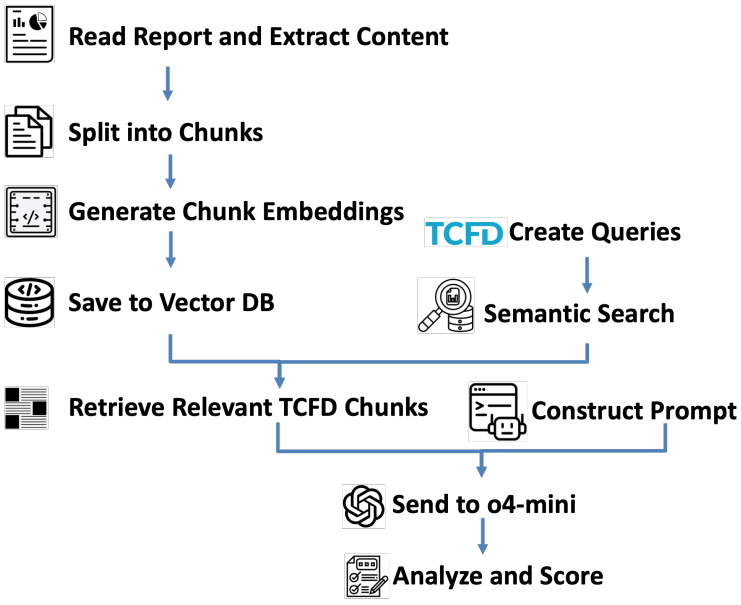


Figure 2. Stage 1: Workflow of the LLM-based TCFD scoring pipeline. Sustainability reports are segmented, semantically indexed, and evaluated against eleven TCFD-aligned queries to produce structured disclosure scores.

In the first stage, each sustainability report is segmented into overlapping text chunks and indexed using semantic embeddings. Relevant passages are retrieved in response to eleven predefined questions aligned with the four TCFD pillars: Governance, Strategy, Risk Management, and Metrics & Targets. Semantic retrieval is implemented using FAISS with OpenAI's `text-embedding-3-large` model.

For each TCFD question, the retrieved text segments are incorporated into a structured prompt and submitted to a compact GPT-based model (o4-mini), and the resulting averaged disclosure score is referred to as the o4-mini score. The model returns a structured JSON output containing a numeric score and a brief analysis explaining the rationale for the assigned score. The model outputs a numeric score ranging from 0 to 100, where higher values indicate a greater degree of fulfillment of the TCFD disclosure requirements. A score close to 0 reflects missing or insufficient disclosure, whereas a score near 100 indicates that most requirements are met with specific and detailed evidence. Intermediate values capture varying levels of partial or incomplete disclosure. The eleven question-level scores are averaged to produce an overall disclosure score for each report.

To improve robustness and mitigate stochastic variability, each report is evaluated three times using identical prompts, and final scores are obtained by averaging across runs. This design yields stable indicators of disclosure quality that serve as inputs for subsequent benchmarking and predictive modeling stages.

3.3. **Stage 2: Benchmarking with Third-Party ESG Ratings**

To evaluate the validity of the LLM-generated TCFD scores, we benchmark them against established third-party ESG ratings, including Bloomberg ESG Disclosure Scores, LSEG Eikon ESG Performance Scores, and Sustainalytics ESG Scores.

We conduct both Pearson correlation (r) and Spearman rank correlation (ρ) analyses. Pearson correlation assesses linear agreement between LLM-generated scores and external benchmarks, while Spearman correlation evaluates consistency in relative rankings. Using both metrics allows us to examine alignment in both magnitude and ordering.

This benchmarking serves two complementary purposes. First, correlation with overall ESG scores evaluates whether LLM-generated TCFD scores approximate the holistic disclosure assessments commonly used by investors and rating agencies. Second, correlation with pillar-level ESG dimensions (Environmental, Social, Governance) provides diagnostic insight into which aspects of sustainability disclosure are most reliably captured by the automated scoring framework.

3.4. **Stage 3: Predictive Modeling**

In the final stage, the eleven TCFD question-level scores are used as input features to train supervised regression models, including Linear Regression, Lasso, Ridge, Random Forest, and Gradient Boosting, to predict expert-driven ESG scores. This step evaluates whether structured, interpretable disclosure scores generated by the LLM can approximate expert-driven ESG assessments and support automated sustainability analytics.

3.5. **Evaluation Metrics**

Predictive performance is evaluated using three standard regression metrics: coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). Together, these metrics capture explained variance, average error magnitude, and sensitivity to large prediction errors, enabling robust comparison across models and target variables.

4. **Results**

4.1. **Stage 1 Results: LLM-based TCFD Scoring, Distribution and Stability**

We evaluate the proposed LLM-based TCFD scoring framework on 100 sustainability reports and analyze its validity, stability, and predictive utility. Figure 3 summarizes the distribution of third-party expert ESG benchmarks (a-c) and the LLM-generated (o4-mini)

scores (d) across the dataset. The observed spread across disclosure- and performance-oriented scores provides necessary context for the subsequent correlation, stability, and predictive modeling analyses.

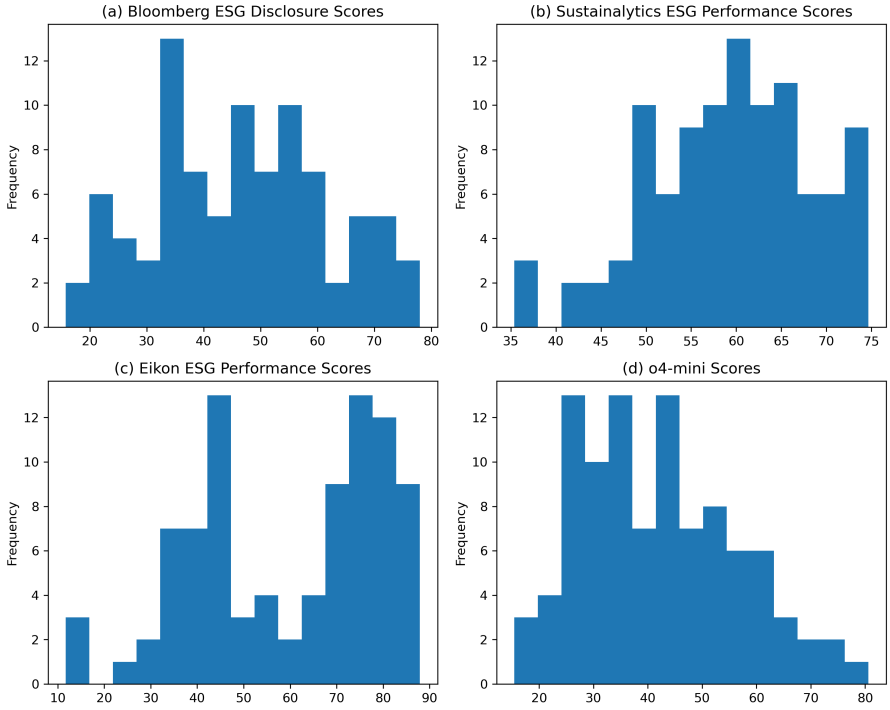


Figure 3. Distribution of ESG-related scores in the dataset. Panels show (a) Bloomberg ESG disclosure scores, (b) Sustainalytics ESG performance scores, (c) Eikon ESG performance scores, and (d) LLM-generated (o4-mini) disclosure scores. The distributions exhibit substantial variability across benchmarks and the LLM-generated scores, indicating that subsequent correlation and predictive analyses are not driven by narrow or degenerate score ranges.

Figure 4 summarizes the distribution of standard deviation across reports. Overall, the LLM-based scores exhibit high stability. The mean standard deviation across reports is 1.17, with a median of 0.95, and 90% of reports showing variability below 2.24. These results indicate that stochastic effects inherent to LLM generation have minimal impact on the final scores, supporting the robustness of the proposed scoring pipeline.

4.2. **Stage 2 Results: Benchmarking with Third-Party ESG Ratings and Correlation Analysis**

We evaluate the alignment between LLM-based TCFD scores and third-party expert-driven ESG benchmarks (overall and by pillar) using Pearson and Spearman correlation analyses. Figure 5 visualizes the pairwise Pearson correlations. The y-axis represents LLM-generated TCFD scores and their four components: Governance, Strategy, Risk Management, and Metrics and Targets. The x-axis represents expert-driven ESG scores, including the overall score and its three dimensions: Environmental, Social, and Governance. Strong positive correlations are observed between the overall o4-mini score and Bloomberg ESG disclosure metrics (a), particularly for the Environmental dimension (up to 0.71), indicating that the LLM-based framework effectively captures disclosure-related information and

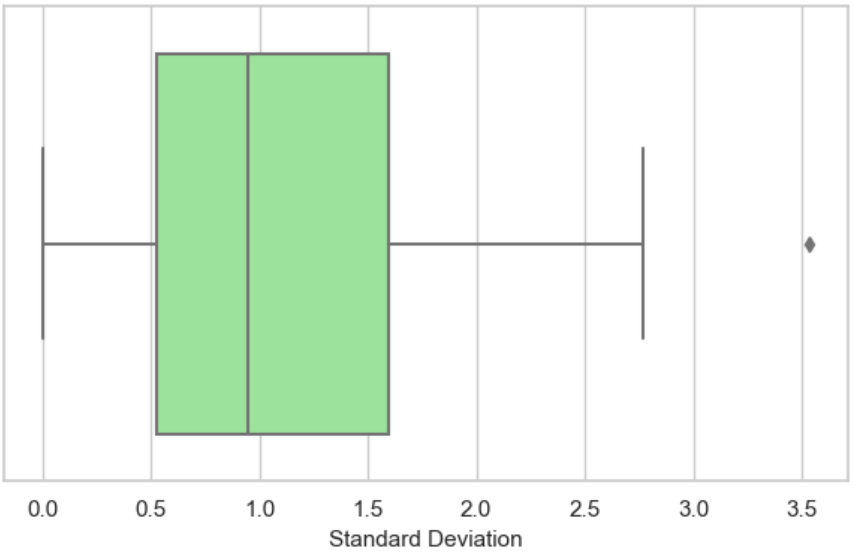
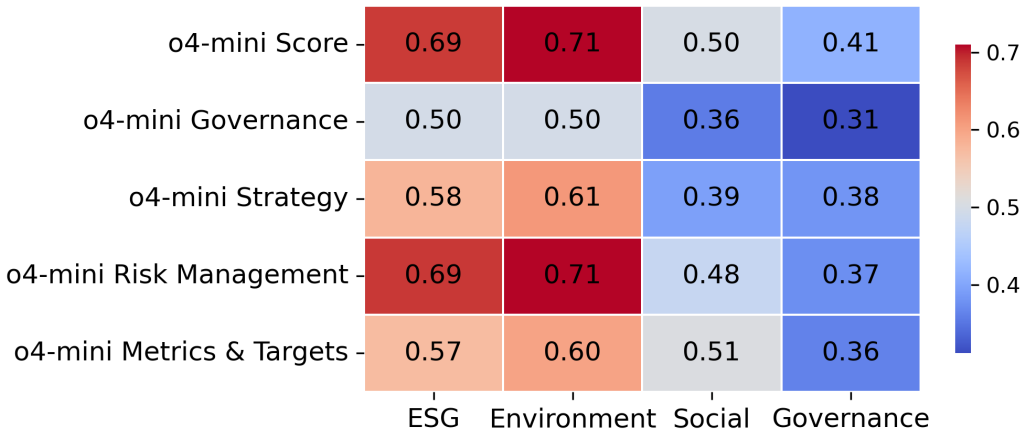


Figure 4. Boxplot showing the dispersion of standard deviation values from the o4-mini model. The majority of scores lie within a tight range, with very few outliers.

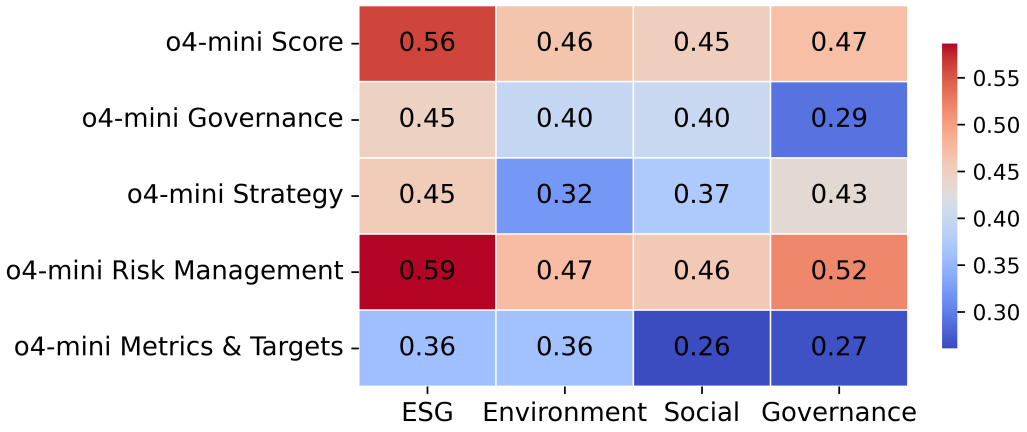
aligns well with external benchmarks. In this context, a strong positive correlation means that reports receiving higher scores from the LLM also tend to receive higher scores from Bloomberg, reflecting consistent agreement in their relative rankings. Among the TCFD pillars, the Risk Management dimension consistently exhibits the highest correlations with Bloomberg indicators, suggesting that risk-related disclosures are most reliably identified by the model. In contrast, LLM-generated TCFD scores exhibit weaker correlations with Sustainalytics and Eikon scores (b-c), reflecting the distinction between disclosure-focused metrics and performance-oriented ESG evaluations. Overall, the heatmaps indicate that the proposed TASR approach aligns more closely with disclosure-based benchmarks than with ESG performance measures.

Spearman rank correlation analysis further confirms these trends, as summarized in Table 1. The LLM-based scores show the strongest alignment with Bloomberg disclosure benchmarks. For example, a Spearman rank correlation of $= 0.70$ indicates a strong positive association between the LLM-generated TCFD scores and expert-driven Bloomberg ESG disclosure scores. In practical terms, this means that reports ranked highly by the LLM-based scoring framework also tend to be ranked highly by Bloomberg, and vice versa. While the rankings are not identical, a correlation of this magnitude suggests substantial agreement in the relative ordering of reports. That is, a large proportion of report pairs are ranked in a consistent order across both scoring methods, although some discrepancies remain. This level of correlation is generally considered strong in empirical studies involving subjective or noisy evaluation criteria, such as ESG disclosure assessments.

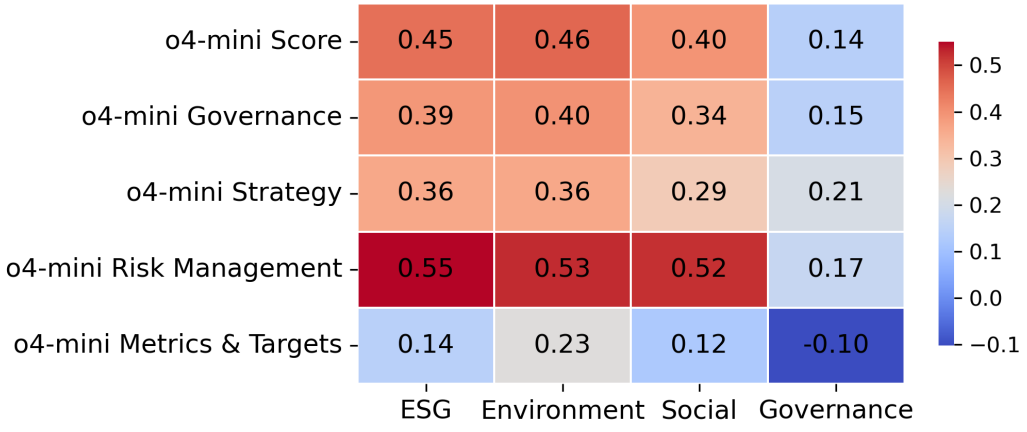
Together, these results indicate that the proposed framework reliably captures disclosure quality, while exhibiting weaker alignment with performance-oriented ESG metrics. This distinction clarifies the intended scope of the method and motivates the use of Bloomberg disclosure scores as predictive targets in subsequent modeling.



(a) o4-mini vs Bloomberg ESG Disclosure Scores



(b) o4-mini vs Sustainalytics ESG Performance Scores



(c) o4-mini vs Eikon ESG Performance Scores

Figure 5. Correlation heatmaps between o4-mini TCFD scores and expert-driven ESG benchmarks. Larger values denote stronger positive correlations.

Benchmark	**Spearman ρ**	***p*-value**
Bloomberg Environmental Disclosure	0.70	$< 10^{-12}$
Bloomberg ESG Disclosure	0.69	$< 10^{-11}$
Sustainalytics ESG Score	0.43	$< 10^{-5}$
Eikon ESG Performance	0.45	$< 10^{-5}$

Table 1. Spearman Rank Correlation Between LLM-based TCFD Scores and expert-driven ESG Benchmarks

4.3. **Stage 3 Results: Predictive Model Evaluation and Comparison**

Overall, the LLM-generated TCFD score (o4-mini score) exhibits strong alignment with third-party ESG disclosure benchmarks and supports downstream predictive analysis. We assess the predictive utility of the LLM-generated TCFD scores using eight supervised regression models with five-fold cross-validation. The eleven question-level TCFD scores serve as input features, while BloombergEnvDisclosure and BloombergESGDisclosure are used as target variables.

Table 2 summarizes model performance across R^2, RMSE, and MAE. Random Forest regression achieves the strongest overall performance, attaining the highest R^2 and lowest error metrics for both targets. Lasso regression performs comparably well, particularly for environmental disclosure prediction, indicating that regularized linear models effectively capture relationships among TCFD-derived features.

In contrast, SVR, Gradient Boosting, and MLP models underperform, with MLP exhibiting unstable behavior and negative R^2 for ESG disclosure prediction. Overall, these results suggest that regularized linear models and tree-based ensembles provide the best balance between accuracy, robustness, and interpretability for ESG disclosure prediction tasks.

Model	**Bloomberg Env Disclosure**			**Bloomberg ESG Disclosure**		
	R^2	**RMSE**	**MAE**	R^2	**RMSE**	**MAE**
Linear	0.514	13.62	11.36	0.432	10.44	8.58
Ridge	0.514	13.62	11.36	0.432	10.44	8.58
Lasso	**0.519**	**13.53**	11.28	0.446	**10.30**	8.45
ElasticNet	0.517	13.57	11.31	0.440	10.36	8.50
SVR	0.316	16.14	13.49	0.326	11.37	9.59
GradientBoosting	0.399	15.13	12.26	0.336	11.29	8.78
Random Forest	**0.519**	13.54	**10.83**	**0.494**	**9.85**	**7.75**
MLP	0.308	16.25	12.82	-0.324	15.93	12.48

Table 2. Comparison of Regression Models on Bloomberg Environmental Disclosure and Bloomberg ESG Disclosure

5. **Conclusion**

This work presents an automated and trustworthy framework for evaluating climate-related disclosures in corporate sustainability reports using large language models (LLMs). By integrating semantic retrieval, expert-informed prompting, and TCFD-aligned LLM-based scoring, we derive structured and interpretable indicators that capture the quality of ESG disclosures from unstructured reports.

Experimental results on 100 sustainability reports demonstrate strong alignment between the LLM-derived scores and established third-party disclosure benchmarks, particularly Bloomberg's Environmental Disclosure score (Spearman's $\rho = 0.70$). Repeated evaluations further show high score stability, indicating that the proposed pipeline is robust to stochastic variability in LLM generation. In addition, supervised learning models trained on the TCFD-aligned scores achieve meaningful predictive performance, with Random Forest and Lasso regression attaining R^2 values of up to 0.519 for environmental disclosure prediction.

Overall, these findings suggest that LLM-based TCFD scoring can serve as a scalable and cost-effective alternative to manual ESG disclosure assessment based on sustainability reports. The proposed framework supports transparent sustainability analytics and has practical implications for investors, regulators, and researchers seeking reliable automated tools for climate disclosure evaluation.

Several limitations warrant consideration. First, while benchmarking against Bloomberg ESG disclosure scores provides external validation, these third-party ratings are themselves proprietary and may introduce measurement noise. Second, the empirical analysis is limited to 100 reports from the oil, gas, and mining sectors, which may constrain generalizability to other industries or regulatory contexts. Finally, although predictive performance is competitive, the observed R^2 values indicate room for improvement, suggesting opportunities for future work on richer modeling strategies and broader data integration.

Future work will extend the proposed TASR framework along several directions. First, we plan to scale the analysis to larger and more diverse datasets spanning additional industries and geographic regions, enabling cross-sector and cross-country comparisons of sustainability disclosure practices. Second, future research will incorporate more advanced retrieval and grounding mechanisms to further improve the faithfulness and explainability of LLM-generated scores. Third, we aim to integrate external signals, such as financial market data and climate-related news, to study how disclosure quality relates to real-world outcomes and market reactions. Finally, the framework can be adapted to emerging disclosure standards beyond TCFD, such as ISSB and IFRS sustainability requirements, reinforcing its role as a flexible and trustworthy foundation for automated sustainability disclosure analysis.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) Insight Development Grant (IDG) and by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant.

Declaration of AI Use

AI-assisted tools (e.g., ChatGPT) were used for language editing and polishing. All scientific content, including methodology, analysis, and conclusions, was developed and verified by the authors.

References

[1] A. Amel-Zadeh and G. Serafeim. "Why and how investors use ESG information: Evidence from a global survey". In: *Financial analysts journal* 74.3 (2018), pp. 87–103. DOI: 10.2469/faj.v74.n3.2.

[2] P. Krueger, Z. Sautner, and L. T. Starks. "The Importance of Climate Risks for Institutional Investors". In: *The Review of Financial Studies* 33.3 (2020), pp. 1067–1111. DOI: 10.1093/rfs/hhz137.

[3] TCFD. *Final Report: Recommendations of the Task Force on Climate-related Financial Disclosures*. Tech. rep. Financial Stability Board, 2017. URL: https://www.fsb-tcfd.org/publications/final-recommendations-report/.

[4] F. Berg, J. F. Kölbel, and R. Rigobon. "Aggregate Confusion: The Divergence of ESG Ratings". In: *Review of Finance* 26.6 (2022), pp. 1315–1344. DOI: 10.1093/rof/rfac033.

[5] J. Achiam, S. Adler, S. Agarwal, et al. "GPT-4 Technical Report". In: *arXiv preprint https://arxiv.org/abs/2303.08774* (2023).

[6] R. Bommasani, D. A. Hudson, et al. "On the Opportunities and Risks of Foundation Models". In: *arXiv preprint arXiv:2108.07258* (2021).

[7] J. Ni, J. Bingler, et al. "CHATREPORT: Democratizing Sustainability Disclosure Analysis Through LLM-based Tools". In: *arXiv preprint arXiv:2307.15770* (2023).

[8] U. Ferjančič, R. Ichev, et al. "Textual Analysis of Corporate Sustainability Reporting and Corporate ESG Scores". In: *International Review of Financial Analysis* 96 (2024), p. 103669. DOI: 10.1016/j.irfa.2024.103669.

[9] T. Shimamura, Y. Tanaka, and S. Managi. "Evaluating the Impact of Report Readability on ESG Scores: A Generative AI Approach". In: *International Review of Financial Analysis* 101 (2025), p. 104027. DOI: 10.1016/j.irfa.2025.104027.

[10] T. Loughran and B. McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks". In: *The Journal of Finance* 66.1 (2011), pp. 35–65. DOI: 10.1111/j.1540-6261.2010.01625.x.

[11] R. G. Eccles, M. P. Krzus, and L. A. Watson. "The Need for Sector-specific Materiality and Sustainability Reporting Standards". In: *Journal of Applied Corporate Finance* 32.2 (2020), pp. 117–125. DOI: 10.1111/j.1745-6622.2012.00380.x.

[12] F. Billert and S. Conrad. "Nano-ESG: Extracting Corporate Sustainability Information from News Articles". In: *arXiv preprint arXiv:2412.15093* (2024).

[13] M. Birti, F. Osborne, and A. Maurino. "Optimizing Large Language Models for ESG Activity Detection in Financial Texts". In: *arXiv preprint arXiv:2502.21112* (2025).

[14] M. Bronzini, C. Nicolini, B. Lepri, et al. "Glitter or Gold? Deriving Structured Insights from Sustainability Reports via Large Language Models". In: *EPJ Data Science* 13.1 (2024), p. 1. DOI: 10.1140/epjds/s13688-024-00481-2.

[15] T. Aue, A. Jatowt, and M. Färber. "Predicting Company ESG Ratings from News Articles Using Multivariate Timeseries Analysis". In: (2025), pp. 1774–1780. DOI: 10.1145/3701716.3717509.

[16] D. Aggarwal and S. Banerjee. "Forecasting of S&P 500 ESG Index by Using CEEMDAN and LSTM Approach". In: *Journal of Forecasting* 44.2 (2025), pp. 339–355. DOI: 10.1002/for.3201.

[17] J. Zhao. "Forecasting ESG Index Based on Machine Learning Methods". In: (2025), pp. 241–246. DOI: 10.1145/3745133.3745174.

[18] E. L. Pontes, M. Benjannet, et al. "Using Contextual Sentence Analysis Models to Recognize ESG Concepts". In: *arXiv preprint arXiv:2207.01402* (2022).

[19] H. Lee, J. H. Kim, and H. S. Jung. "ESG-KIBERT: A New Paradigm in ESG Evaluation Using NLP and Industry-specific Customization". In: *Decision Support Systems* 193 (2025), p. 114440. DOI: 10.1016/j.dss.2025.114440.

[20] M. Zhang, Q. Shen, Z. Zhao, S. Wang, and G. Q. Huang. "Optimizing ESG Reporting: Innovating with E-BERT Models in Nature Language Processing". In: *Expert Systems with Applications* 265 (2025), p. 125931. DOI: 10.1016/j.eswa.2024.125931.

[21] C. Colesanti Senni, T. Schimanski, et al. "Combining AI and Domain Expertise to Assess Corporate Climate Transition Disclosures". In: *SSRN Electronic Journal* (2024). DOI: 10.2139/ssrn.4826207.

[22] M. Chuang, G. Chuang, et al. "Judging It, Washing It: Scoring and Greenwashing Corporate Climate Disclosures using Large Language Models". In: *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*. 2025, pp. 17–31.

[23] S. Chen. "The Influence of Artificial Intelligence and Digital Technology on ESG Reporting Quality". In: *International Journal of Global Economics and Management* 3.1 (2024), pp. 301–310. DOI: 10.62051/IJGEM.v3n1.36.

[24] C. De Villiers, R. Dimes, and M. Molinari. "How will AI Text Generation and Processing Impact Sustainability Reporting?" In: *Sustainability Accounting, Management and Policy Journal* 15.1 (2024), pp. 96–118. DOI: 10.1108/SAMPJ-02-2023-0097.