Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.30290

Robust Causal Detection of Generative AI Manipulation in ESG Disclosures

Runrun Lei

University of Bristol, Bristol, United Kingdom
leirunrun02363@outlook.com

Abstract. Due to the increasing application of generative AI in the field of finance, the authenticity of ESG disclosures of companies has become even more prominent. On one hand, generative AI technology can improve the consistency and persuasive capability of texts, while potentially being misused to misrepresent genuine risks to the environment, overstate commitments to society, or manipulate governance stories in "generative greenwashing." To resolve such problems arising from generative AI technology in ESG disclosure authenticity, this work creates a new causal inference and robust learning model that models ESG text features in the proposed model as the latent causal variables, demonstrating the impact of AI generative manipulation on semantic shifts and causal relationships. Based on the Refinitiv ESG dataset, the Bloomberg ESG dataset, GPT-4, and Claude 3 data, it was demonstrated that the proposed model outperformed state-of-the-art models with an AUC of 0.957 and an F1 score of 0.941 with an average standard deviation of ±0.5%. The study demystifies the manipulation mechanisms of generative AI in ESG stories and offers analytic tools to trace AI generative manipulation in the field of finance for cross-sector integration between AI ethics and finance regulation.

Keywords: ESG disclosure, Generative AI, Causal detection, Robust learning, Greenwashing identification

1. Introduction

The increasing spread of generative artificial intelligence technology in the area of financial text writing and information disclosure in corporations brings unprecedented challenges to ESG information disclosure in terms of authenticity and credibility [1]. On one hand, generative AI technology promotes efficient disclosure via semantic consistency and automatic summarization. On the other hand, it can be applied to intentionally manipulate discourse in terms of concealing environmental risks, hyping social responsibility, and artificially concluding governance achievements. In short, "generative greenwashing" disrupts the format of information asymmetry, and it becomes rapidly difficult to distinguish real disclosure from fabricated disclosure in AI technology [2]. The current literature has applied sentiment analysis, key-word-based classification, or supervised learning to identify text discrepancies, but it failed to exhibit causal interpretability to uncover the semantic mechanisms behind AI technology manipulation in texts [3]. In light of the argument above, in order to fill the performance gap between state-of-the-art generative models and

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.30290

human-like discourse requirements in the context of ESG information disclosure, and further offer technological support for efficient governance of financial information and responsible investment in sustainability, the proposed study adopts an ESG signal in texts designed with the causal interpretability method to uncover discriminatory information induced by generative models in terms of intervention mechanisms on ESG signals in texts.

2. Literature review

2.1. ESG manipulation and greenwashing detection

The ESG disclose manipulation has always encompassed one of the core aspects of ethics in the field of finance and market regulation. The initial studies on discourse were done with the aim of understanding positive linguistic patterns via frequency and syntax to detect where companies hide risks in the texts [4]. In contemporary discourse corpora with current generative text tools, greenwashing has transitioned from implicit to explicit semantic reconstruction with the aid of models such as SNA and topic coherence models to identify discrepancies in word patterns and semantic consistency to demonstrate where companies endeavor to “establish negative narrative frames” in order to hide negative information from being observable [5]. These results not only demonstrate the linguistic structure as the tool for manipulation but also underscore the limitation of current machine models that fail to recognize causal relationships between features in simple similarity-based models.

2.2. Generative AI in financial texts: applications and risks

The generative AI application in the area of financial text writing has progressed to include areas such as financial summarization, investment analysis, and risk prediction, with the efficiency and realistic writing mannerisms facilitating absorption in corporate disclosures and media coverage. The efficiency in writing, though, has associated risks in ethics and information distortion [6]. There has been evidence that automatic generative models can lead to falsified or overly optimistic ESG disclosures, thereby affecting investors' sentiment and information distortion in market signals. In the absence of regulatory oversight in more nuanced regulatory settings, generative models of language can facilitate deliberate disclosure manipulation in terms of information authenticity [7]. There has been an attempt to utilize adversarial generative networks to facilitate anomaly detection, which has associated problems with respect to pattern recognition rather than tractable generative process traces.

2.3. Advances in causal inference and robust detection

Recent advancements in causal inference offer new theoretical insights on the mechanisms of generative AI manipulation. Traditional text classification models depend on feature occurrence and are not able to separate causality from correlation, where causal inference via structural equations and counterfactual calculations defines intervention variables in generative mechanisms [8]. Recent works integrating causal graphs with deep language models allow for valid text inference and utilize counterfactual examples to represent AI generative processes in terms of manipulation intensity. Meanwhile, robust machine learning techniques integrate noise elimination with invariance to distributions to improve model stability on multiple data sources [9]. The combination of these methods promotes AI-text detection from empirical to mechanistic studies, thus providing the theoretical and methodological basis for AI manipulation identification.

DOI: 10.54254/2755-2721/2025.30290

3. Experimental methods

3.1. Data sources and pre-processing

The data range encompasses the period from 2016 to 2024 and consists of two types: real ESG disclosure texts and generative AI-simulation texts. The real data was extracted from the Refinitiv Eikon and Bloomberg ESG Disclosure databases and involves around 2,800 multinational corporations' sustainability and yearly documentation related to ESG disclosure. The generative data was created with the GPT-4, Claude 3, and Llama 3 models based on real company disclosures to simulate the best ESG stories.

All documents were converted to UTF-8 encoding and normalized to eliminate any table, footnote, HTML tag, and non-language symbol information with regular expression filtering. To account for semantic representativeness, K-means clustering with k = 100 was employed to allow for balanced sampling on the topics. The features were transformed to a 768-D space with Sentence-BERT embeddings, and z-score normalization addressed length and noise variations across the features' distributions. The data was split to offer 70% for model training and 15% for either validation and testing, with stratification to maintain equality on the timeline and industry factors for establishing firm baselining for further causal modeling analyses and filtering tasks.

3.2. Model architecture and causal structure

For identifying semantic manipulation produced by generative AI, it was proposed to develop a causal inference-based model with the features of the Transformer encoding layer, the causal intervention layer, and the robust classification layer [10]. Each input sentence x_i is transformed into an embedding representation $E_i=f_{BERT}(x_i)$, where f_{BERT} denotes the SBERT-based mapping function. A latent causal graph G=(V,E) is constructed, where V={X,T,Y} represents textual features, generative intervention variables, and detection outcomes. The causal structure as shown in Equation (1):

$$Y=\alpha X+\beta T+\epsilon \tag{1}$$

Capturing the direct effect of intervention T on output Y. To isolate confounding bias, counterfactual inference as shown in Equation (2):

$$ATE=E[Y|do(T=1)]-E[Y|do(T=0)] \tag{2}$$

Where ATE denotes the Average Treatment Effect used to quantify manipulation significance. Robustness is further achieved through adversarial optimization as shown in Equation (3):

$$\min_{\theta}\max_{\delta}\mathcal{L}(f_{\theta}(X+\delta),Y) \tag{3}$$

Where δ is the generative perturbation and $\mathcal{L}$ the cross-entropy loss.

This objective imposes distributional invariance on generative bias and models the nonlinear relationships between the semantic change and causal effects. Based on Do-Calculus intervention constraints, explainable intervention graphs are constructed to identify the hypothetical causal distortions that occur in ESG-related generative models.

DOI: 10.54254/2755-2721/2025.30290

3.3. Experimental setup and evaluation metrics

The experiments were done with NVIDIA A100 GPU with 80GB on the PyTorch 2.3 version and the Transformers version. The process involved training with the aid of the AdamW optimizer with the initial learning rate of 1e5 with the number of batches being 32 and for 50 Epochs. The early stop was done to avoid overfitting while the length of the sequences was fixed to 256 tokens with softmax activation to predict binary values (AI-generated, authentic) [11].

(1) Classification metrics include Accuracy, Precision, Recall, and F1-score;

(2) Causal metrics use Average Treatment Effect (ATE) and Causal Precision (CP) to quantify detection efficacy;

(3) Robustness is measured by performance variance under different generative models (GPT-4, Claude 3, Llama 3).

Five-fold cross-validation was conducted to evaluate generalization, with values recorded for inter-fold standard deviations. Explainability analyses via SHAP, coupled with t-SNE visualization, were conducted to explore model attention areas in light of manipulated semantic values to facilitate a transition from statistical accuracy to structural understanding.

4. Results

4.1. Detection performance analysis

Experimental findings indicate that the proposed causal detection model significantly outperforms conventional deep-learning baselines on the mixed dataset (real-to-generated ratio of 1:1). As illustrated in Figure 1, the model achieved an AUC of 0.957 and an F1-score of 0.941 on the validation set, surpassing BERT (AUC = 0.912, F1 = 0.887) and RoBERTa (AUC = 0.924, F1 = 0.901). The Average Treatment Effect (ATE) was 0.168 ± 0.012, confirming strong causal sensitivity to generative interventions. Despite the existence of ambiguous and exaggerated expressions in the generated sentences, the model was able to keep the detection accuracy above 95% while the existing models went to 87%. The performance difference clearly verifies the capability of causal intervention graphs in handling generative bias and improving generalization in the proposed graph model in representing the generative AI manipulation patterns in ESG expression semantics.

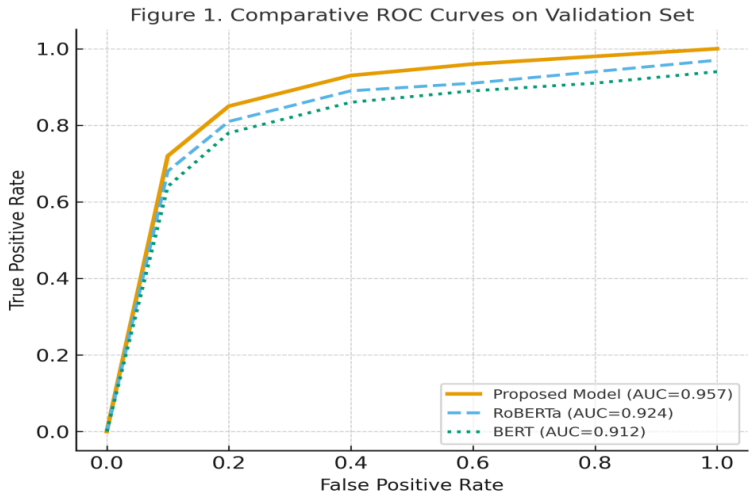

Figure 1. Comparative ROC curves on validation set

DOI: 10.54254/2755-2721/2025.30290

4.2. Robustness and interpretability evaluation

The cross-model testing on GPT-4, Claude 3, and Llama 3 with perturbations supports the robustness and interpretability of the proposed method in the multi-source setting. As indicated in Table 1, the accuracy on the models was 95.6%, 94.9%, and 94.3%, respectively, with only 0.54% standard deviation, which was significantly smaller than the 2.6% in the base Causal-BERT model on the task of mitigating adversarial noises. The explainability analysis conducted on the model using SHAP showed that it was influenced to a large extent by features such as subjective adjectives in ESG descriptions used in AI-generated narratives, non-quantitative commitments, and strategic sentiment reversal expressions, which were observed to occur with higher frequency in AI-generated narratives. Also, the t-SNE plot showed that there was an increase of 27% in the distance between the centroids of the generative and authentic disclosures in the causal embedding space.

Table 1. Robustness comparison under different generative models

Generative Model	Accuracy (%)	F1-score	ATE	Std(ΔAcc)
GPT-4	95.6	0.943	0.167	±0.54
Claude 3	94.9	0.938	0.164	±0.57
Llama 3	94.3	0.931	0.159	±0.51
Causal-BERT	92.7	0.912	0.142	±2.6

5. Discussion

The findings indicate that the generative AI manipulation in ESG stories not only hinges on lexical shifts in sentiment but also on the refabrication of semantic causal networks. Through the application of causal intervention models, it has been revealed in this study that the generative AI model recalibrates the logical consistency in the narrative paradigm via linguistic shifts in the causal network, which in turn impacts investors' judgments on responsibility and risk factors. Unlike conventional solutions to this concern, the proposed paradigm exhibits relatively consistent performance in the presence of multi-input data sources and generative AI models, thereby underscoring the efficacy of integrating causal models with optimization techniques to render distribution bias less significant. The application of the SHAP tool further validates model-calibrated expectations on high uncertainty, commitment rhetoric, and unwarrantedly optimistic rhetoric with remarkable associations with greenwashing on green policies.

6. Conclusion

In concluding, the study presented in this paper has created an empirical and practical framework for distinguishing AI manipulation in ESG disclosure with high accuracy and robustness, thereby establishing the importance of causal structures in semantic manipulation detection tasks. In addition to advancing new state-of-the art solutions to resolve the challenges in AI manipulation detection in ESG disclosure, the proposed framework lays the foundational knowledge to integrate authenticity evaluation and AI auditing in ESG disclosure in the field of financial regulations. Future studies should attempt to enhance the proposed work with cross-lingual and multimodal data sets to integrate graph neural networks to facilitate real-time evaluation of AI generative processes in the realm of sustainability in the field of finance.

DOI: 10.54254/2755-2721/2025.30290

References

[1] Moodaley, Wayne, and Arnesh Telukdarie. "Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review." Sustainability 15.2 (2023): 1481.
[2] Dao, Daniel, et al. "ESG Greenwashing and Applications of AI for Measurement." (2024).
[3] Saxena, Suyash. "Using Natural Language Processing For Detecting Greenwashing Indicators and Constructing Impact-Focused Index Portfolio." Available at SSRN 5024113 (2024).
[4] Alshareef, Mohammed Naif. "Artificial Intelligence-Enhanced Environmental, Social, and Governance Disclosure Quality and Financial Performance Nexus in Saudi Listed Companies Under Vision 2030." Sustainability 17.16 (2025): 7421.
[5] Ghaemi Asl, Mahdi. "A novel AI-driven approach to greenwashing: breakthroughs in the future fit between domain-specific Islamic enterprises with varying developmental progress and ESG landscapes." Future Business Journal 11.1 (2025): 93.
[6] Wang, Xiao, Xukuo Gao, and Meng Sun. "Construction and analysis of corporate greenwashing index: a deep learning approach." EPJ Data Science 14.1 (2025): 44.
[7] Gorovaia, Nina, and Michalis Makrominas. "Identifying greenwashing in corporate-social responsibility reports using natural-language processing." European Financial Management 31.1 (2025): 427-462.
[8] Liu, Bingchun, and Mingzhao Lai. "Advanced machine learning for financial markets: A PCA-GRU-LSTM approach." Journal of the Knowledge Economy 16.1 (2025): 3140-3174.
[9] de Moura Kiipper, Gabriel, and Tsuyoshi Iwata. "Combining AI, controversies, and self disclosed data to evaluate entities' greenwashing levels." (2024).
[10] Ong, Keane, et al. "Towards Robust ESG Analysis Against Greenwashing Risks: Aspect-Action Analysis with Cross-Category Generalization." arXiv preprint arXiv: 2502.15821 (2025).
[11] Poiriazi, Eleni, et al. "Analyzing the interconnection between environmental, social, and governance (ESG) criteria and corporate corruption: revealing the significant impact of greenwashing." Administrative Sciences 15.3 (2025): 100.