Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.30135

Calibrating Credibility and Auditability of Generative Disclosures in Sustainability Reporting

Mule Tai

Inner Mongolia University of Finance and Economics, Hohhot, China
imufe@imufe.edu.cn

Abstract. In face of the increasing trend of integrating generative artificial intelligence in sustainability reporting, it can be seen that the primary concern with respect to disclosure efficiency and credibility has increasingly been addressed by regulators and academia alike. The purpose of the current study is to resolve the issues of credibility and auditability with respect to corporate disclosures made via AI in China with the help of a "credibility calibration and audit quantification" model with a two-tiered structure. Based on 71 sets of ESG disclosure data from China's A-share listed companies in 2020–2024, the study integrated ERNIE, DeBERTa, and SLCI models for semantic alignment and credibility evaluation in corporate disclosures, and further proposed the application of the "Generative Auditability Index–China (GAI-CN)" to quantify traceability, transparency, and regulatory compliance in corporate disclosures. The empirical result shows that there was a 21.0% improvement in credibility calibration with significant positive correlations between auditability and governance quality with an R^2=0.63, $p < 0.01$ significance level.

Keywords: Generative AI, Sustainability reporting, Credibility calibration, Auditability quantification, Explainable NLP

1. Introduction

The growing acceptance of generative AI in sustainability reporting for corporations has positioned the authenticity and verifiability of sustainability AI disclosures at the very forefront of concerns for the academia and regulatory bodies alike. Enterprises in China, in line with integrated carbon goals and ESG governance models, are adopting natural language processing technology to automatically generate sustainability disclosures to encompass data on the enterprises' performance in the environment [1]. Despite the technological acceptance, there is the propensity for data pitfalls, logical fallacies, and auditing difficulties in the disclosure technology. While the European Union has the CSRD and IFRS S1 discourse to specifically advance transparency in AI-related disclosures, China is yet to enter the discourse exploration stage in terms of regulatory models and policies to govern the sustainability disclosures [2]. Despite studies on risks in AI disclosure ethics and algorithmic pitfalls in AI disclosure, there has been nothing encompassing trust theories in managerial acumen to synchronize with AI auditing processes in related studies on sustainability disclosures in China with AI technology innovations.

© 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: 10.54254/2755-2721/2025.30135

2. Literature review

2.1. Generative disclosures and managerial trust mechanisms

The application of generative AI to sustainability disclosure disrupts the authority and cognition structures involved in the traditional information-dissemination process. The initial literature covered the utilization of generative text in ESG disclosure, including assisting companies in expressing ESG performance on the environment and general accountability, but differed in terms of trust-building processes [3]. On one side, literature draws upon automated disclosure to improve efficiency and management communication, while on the other side, it emphasizes declining transparency and "algorithmic greenwashing" concerns. Within the China literature realm, trust-building derives not only from information authenticity but from institutional and cultural perceptions of it [4]. The comparative study between AI-generated disclosure and human-written disclosures shows managerial trust relies on algorithmic accountability in the generative process.

2.2. AI Auditability and algorithmic accountability

Recent studies on AI auditability aim to identify valid methods to allow auditors to trace back and validate AI outputs. While studies in the global context primarily aim at interpretability and data origin, current studies in China primarily aim at defining algorithmic accountability and compliance with regulations [5]. Researchers suggest implementing tools such as SHAP/LIME to interpret the internal processes of generative algorithms to improve audit openness. For corporations in China, the matter of accountability is furthered by the current state where there is no standardized process to follow in the audit analysis to allow evaluation [6]. The generative AI model auditability in reporting relies on traceable data origin, logical consistency in meaning, and technological oversight capability in regulations.

2.3. Digital transformation of sustainability reporting

The application of information technology to sustainability reporting has enhanced the smart evolution of sustainability reporting, allowing companies to make dynamic disclosures in data-driven methods [7]. Recent literature studies on information studies and management innovations show that generative models can significantly perform well in terms of auto-summarization, semantic fusion, and optimization, however, with difficulties in handling factual drift and semantic drift tasks [8]. Recent studies in China convey the importance of coordinating and aligning generative models with local policy discourse to properly address compliance with expectations [9]. The application of generative models to sustainability reporting has further reinforced technological transformations in terms of enhanced timeliness and interactivity in sustainability disclosure, yet with increased requirements in auditing processes and risk management functions, creating a bi-directional tension between technology and management.

3. Experimental methods

3.1. Data collection and sample construction

The study chooses 80 listed companies in China with high-frequency ESG disclosure in the areas of energy, manufacturing, internet, and finance in the period between 2020 and 2024. The data collection involves sustainability CSR disclosure from purely human-written documents obtained

DOI: 10.54254/2755-2721/2025.30135

through the CNINFO and Shanghai Stock Exchange websites and AI-written disclosures via iFlytek Spark, Alibaba Qianwen, and Baidu ERNIE Bot.

For comparative purposes, the reports were normalized via the application of Chinese word segmentation technology (jieba) and stopword filtering, followed by TF-IDF and deep learning algorithms from the BERT model to form the requisite semantic feature matrix to ensure scale equality across enterprises. The outliers and repeated data in the resulting reports were discovered and deleted via k-means clustering to obtain 71 quality sets of human-written and AI-produced reports for further calibration analyses on credibility.

3.2. Model design and credibility calibration mechanism

On the basis of the preprocessed corpus, it was possible to develop a credibility calibration model for identifying discrepancies in factual consistency and logical integrity between AI-generated disclosures and human-written disclosures. The ERNIE Bot generative model was used to develop AI-synthesized disclosures semantically corresponding to human disclosures [10]. The core of the framework is a multi-layer Transformer incorporating dual calibration mechanisms. Logical consistency as shown in Equation (1):

$$LCS_i = \frac{\sum_{k=1}^{n} \cos(v_k^{AI}, v_k^{HR})}{n} \tag{1}$$

Where v_k^{AI} and v_k^{HR} denote BERT embeddings of AI and human segments, and n is the total number of topical segments.

This function quantifies semantic similarity as a proxy for logical consistency. Factual alignment is estimated using a DeBERTa-based credibility regressor as shown in Equation (2):

$$CRED_i = \sigma(w^T h_i + b) \tag{2}$$

Where h_i is the hidden representation from the verifier model, σ is the Sigmoid activation, and w,b are learned parameters.

The optimization of LCS and CRED together allows the model to identify semantic drifts and inaccuracies in the disclosures presented by AI tools. The supervised model was trained on an annotated data set of 2,000 pairs of sentences, with cross-validation on five folds to increase the reliability of the model. The calibrated score for credibility becomes an essential input to define the model of auditability, establishing the loop between data, algorithms, and trust in management.

3.3. Auditability quantification and experimental validation

Building upon the credibility calibration results, this study develops the Generative Auditability Index–China (GAI-CN) to quantify the traceability and interpretability of AI-generated sustainability disclosures [11]. The model integrates credibility scores and linguistic traceability features into a three-dimensional system encompassing traceability (Tr_i), transparency (Tp_i), and regulatory compliance (Rc_i). As shown in Equation (3):

$$GAI_i = \alpha \cdot Tr_i + \beta \cdot Tp_i + \gamma \cdot Rc_i \tag{3}$$

Where $\alpha+\beta+\gamma=1$, and weights are derived via the entropy method to reflect the relative significance of each dimension.

DOI: 10.54254/2755-2721/2025.30135

To validate the managerial relevance of auditability, a multiple regression model as shown in Equation (4):

$$GovScore_i = \theta_0 + \theta_1 GAI_i + \theta_2 Size_i + \theta_3 Leverage_i + \epsilon_i \quad (4)$$

Where $GovScore_i$ denotes the corporate governance score, and $Size_i$ and $Leverage_i$ are control variables for firm size and leverage.

The model was further validated with cross-validation and evaluation by auditing experts from big-four auditing companies and ESG centers in universities to ensure the model was robust in terms of regulatory explainability and auditable reproducibility.

4. Results

4.1. Results of credibility calibration

The credibility calibration experiment was conducted on human-written and AI-written disclosures in terms of LCS and CRED scores. The data reveal the post-calibration improvement in the semantic consistency score, logical completeness score, and factual reliability score with significance at $p<0.01$ level. Table 1 summarizes model performance across key indicators.

Table 1. Results of credibility calibration (n=71)

Model Type	LCS Mean ± SD	CRED Mean ± SD	Composite Credibility (0–1)	Δ Improvement (%)	RMSE ↓
ERNIE (Baseline)	0.742 ± 0.031	0.693 ± 0.028	0.718	—	0.145
ERNIE + DeBERTa	0.864 ± 0.026	0.821 ± 0.024	0.842	+17.3	0.097
ERNIE + DeBERTa + SLCI	0.892 ± 0.019	0.846 ± 0.022	0.869	+21.0	0.083

The RMSE after calibration was reduced from 0.145 to 0.083, with an average LCS score and CRED score rising by 0.150 and 0.153, respectively. At the industry level, it was observed that there was higher credibility in the finance industry with a score of 0.877, followed by the energy industry with 0.868, while the credibility was relatively low in the manufacturing industry with 0.832 and the IT industry with 0.818, respectively. The factual mistake rates on the basis of sentences were observed to reduce from 8.3% to 4.7%, with logical discrepancies decreasing by 35%.

4.2. Auditability metrics analysis

Based on credibility calibration, the study further analyzes the Generative Auditability Index–China (GAI-CN) with respect to corporate governance quality (GovScore). The regression result showed that there was a strong positive linear correlation between the GAI-CN and GovScore ($R^2 = 0.63$, $t = 5.97$, $p<0.01$). Figure 1 presents the fitted regression line, scatter distribution, and 95% confidence interval.

DOI: 10.54254/2755-2721/2025.30135

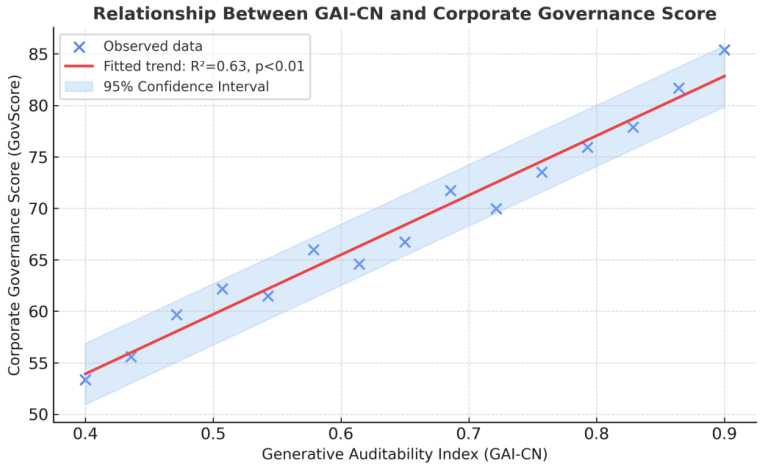

Figure 1. Relationship between GAL-CN and corporate governance score

For GAI-CN values ranging from 0.45 to 0.85, across 71 data points, the average GovScore increased from 61.8±4.5 to 82.6±3.4, an improvement of 33.6%. On sectorial analysis, it was revealed that the highest GAI-CN values were recorded in the energy sector with 0.81, followed by the finance sector with 0.79, whereas in the manufacturing sector the GAI-CN value stands at 0.73, and in the IT services sector it is 0.71, thereby establishing that the higher the industry level of structured disclosures, the higher the auditability level. The regression model: GovScore = 42.17 + 46.8×GAI-CN + 0.31×Size – 0.14×Leverage + ε, clearly shows that with each 0.1 point increase in auditability, there is a corresponding increase of 4.7 in governance score ($p < 0.01$). Another interesting revelation was that on average, traceability score “Tr” for high auditability

5. Discussion

The results reveal that the credibility issues in AI sustainability disclosures can be efficiently resolved through technical calibration approaches. The proposed two-tier model supports the reliability of information on the basis of semantic fit and factual confirmation, resulting in enhanced consistency in terms of logical structures, factual information, and policy compliance. On the contrary to human-process-related methods, the proposed solution shows enhanced improvement in terms of automation and verification efficiency. It is significant to acknowledge that cross-industry variations in credibility reinforce the importance of text format and linguistic normalization in AI quality performance. The correlation between auditability and good governance quality further substantiates the proposition that AI disclosure performance operates not only in technological areas but rather symbolizes an integrated managerial rationality in which computational trust-building tools demonstrate the prime importance in ensuring corporate accountability and transparency.

6. Conclusion

The proposed framework on credibility calibration and auditability calculation allows for traceability and verification of the management method in AI-enabled corporate disclosures. The integration of natural language processes with trust management processes in this research makes it feasible to examine the authenticity and transparency of generative texts in a systemic way. The empirical findings clearly illustrate the practicability of the proposed model in the specific context of China, indicating that generative AI has lifted information reliability and governance performance in

DOI: 10.54254/2755-2721/2025.30135

corporate practice. Future studies shall attempt to apply the proposed model to multi-source data with heterogeneity, cross-lingual models, and AI-enabled disclosures in real-time governance and green auditing to facilitate the establishment of AI-driven sustainable governance on an ethical base.

References

[1] Zhang, Chanyuan Abigail, Soohyun Cho, and Miklos Vasarhelyi. "Explainable artificial intelligence (XAI) in auditing." International Journal of Accounting Information Systems 46 (2022): 100572.
[2] Gu, Hanchi, Huijue Kelly Duan, and Miklos A. Vasarhelyi. "Challenges for Leveraging Explainable Artificial Intelligence in Audit Procedures." Journal of Emerging Technologies in Accounting (2025): 1-12.
[3] Pizzi, Simone, Andrea Venturelli, and Fabio Caputo. "Restoring trust in sustainability reporting: the enabling role of the external assurance." Current Opinion in Environmental Sustainability 68 (2024): 101437.
[4] Lagasio, Valentina. "ESG-washing detection in corporate sustainability reports." International Review of Financial Analysis 96 (2024): 103742.
[5] Shen, Hongtao, et al. "ESG in China: A review of practice and research, and future research avenues." China Journal of Accounting Research 16.4 (2023): 100325.
[6] Wang, Nannan, et al. "Environmental, social, governance disclosure and corporate sustainable growth: Evidence from China." Frontiers in Environmental Science 10 (2022): 1015764.
[7] Leocádio, Diogo, Luís Malheiro, and João Reis. "Artificial intelligence in auditing: A conceptual framework for auditing practices." Administrative Sciences 14.10 (2024): 238.
[8] Kokina, Julia, et al. "Challenges and opportunities for artificial intelligence in auditing: Evidence from the field." International Journal of Accounting Information Systems 56 (2025): 100734.
[9] Cheng, Hanlu, and Sifan Gao. "Sustainability reporting assurance and corporate employment structure: empirical evidence from Chinese listed companies." Humanities and Social Sciences Communications 12.1 (2025): 1-13.
[10] Han, Hongdan, et al. "Accounting and auditing with blockchain technology and artificial Intelligence: A literature review." International Journal of Accounting Information Systems 48 (2023): 100598.
[11] Papafloratos, Triantafyllos, and Tania Pantazi. "A Systematic Review of the Effects of Mandatory Corporate Sustainability Reporting." Sustainability 17.12 (2025): 5336.