Journal of Ecohumanism
2026
Volume: 4, No: 4, pp. 2937–2955
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.7071

Generative AI for Automated Financial Reporting and Narrative Generation

Nesrat Abdelouahab[1], Aouni Mohammed Seghir[2], Abdelkamel Maamri[3]

Abstract

Generative artificial intelligence (GenAI) is rapidly being embedded into corporate reporting workflows, yet its implications for financial reporting quality and auditability remain insufficiently understood. This paper examines how GenAI models can be used to automate financial reporting narratives—such as Management's Discussion and Analysis (MD&A) and risk disclosures—and evaluates their effects on disclosure quality, transparency, and assurance. The study employs an experimental mixed-methods design, comparing human-authored, GenAI-generated, and human-edited GenAI narratives based on a large sample of corporate reports. Text-analytic techniques (readability indices, sentiment and topic analysis, and red-flag indicators) are combined with explainable AI methods to assess both the content produced by GenAI and the traceability of underlying decision processes. The findings indicate that GenAI can substantially improve readability and linguistic consistency while reducing boilerplate, but also introduces new risks related to hallucinated details, optimistic bias, and potential masking of earnings-management signals. Explainability tools partially mitigate these concerns by providing auditable evidence of how inputs shape outputs, yet do not fully resolve issues of accountability and professional scepticism. Overall, the paper contributes empirical evidence and a governance framework for responsibly integrating GenAI into financial reporting and auditing, offering practical guidance for preparers, auditors, and regulators seeking to harness automation without compromising reliability or trust.

Recieved: 21/01/2025 Accepted 29/03/2025 Published : 30/04/2025

Introduction

The rapid diffusion of generative artificial intelligence (GenAI) marks a new phase in the digital transformation of accounting and corporate reporting. Large language models (LLMs) such as GPT-4 can synthesise structured and unstructured data into fluent, context-aware narratives at a scale and speed that were previously unattainable with traditional rule-based systems (Bommasani et al., 2021; Blankespoor, deHaan, & Zhu, 2025). In the financial domain, these capabilities are increasingly being applied to the preparation of annual reports, management commentary, earnings announcements, and sustainability disclosures, raising fundamental questions about the reliability, transparency, and auditability of AI-generated reporting (Faccia, Mosteanu, & Moșteanu, 2023; SEC, 2024).

Financial reporting narratives—such as Management's Discussion and Analysis (MD&A), risk factor sections, and integrated reports—play a central role in conveying forward-looking information, managerial intent, and contextual explanations that cannot be easily communicated through quantitative statements alone (Li, 2010; Loughran & McDonald, 2016). However, prior research documents pervasive problems in these narratives, including boilerplate language, obfuscation, and strategic tone management that may hinder investors' understanding and facilitate earnings management (Athanasakou & Hussainey, 2014; Bonsall, Leone, Miller, & Rennekamp, 2017). Preparers also face increasing regulatory pressure and resource constraints as disclosure requirements expand to cover complex areas such as climate-related and ESG reporting (International Sustainability Standards Board [ISSB], 2023; Securities and Exchange Commission [SEC], 2022). Against this backdrop, organisations are experimenting with GenAI systems to draft, summarise, and translate financial disclosures in order to improve efficiency and consistency while reducing preparation costs (Deloitte, 2024; KPMG, 2023).

[1] University of El Oued Laboratory of Renewable Energy Economics and Its Role in Achieving Sustainable Development, Email: nesrat-abdelouahab@univ-eloued.dz

[2] University of El Oued Laboratory of Renewable Energy Economics and Its Role in Achieving Sustainable Development, Email: aouni-med-seghir@univ-eloued.dz

[3] University of El Oued, Email: abdelkamel-maamri@univ-eloued.dz

Despite these developments, academic evidence on the consequences of GenAI for financial reporting remains limited. Early studies suggest that LLM-assisted writing can improve linguistic quality, including readability and grammatical correctness, but may also introduce new forms of bias, hallucinated content, and over-optimistic tone (Ji, Lee, Frieske, Yu, & Fung, 2023; Jakesch, Hancock, Riedl, & Naaman, 2023). In the accounting literature, emerging work documents that algorithmic generation and editing of narratives can alter disclosure style in ways that affect investor reactions and information asymmetry (Blankespoor et al., 2025). However, there is little systematic analysis of how GenAI performs when tasked with producing full financial reporting narratives from underlying data, nor of how auditors and regulators can evaluate the integrity of such AI-mediated communication.

From an assurance perspective, GenAI poses both opportunities and challenges. On one hand, AI-generated drafts and summarisation tools can support auditors by structuring large volumes of information, highlighting anomalies, and standardising wording across reporting units (Rozario & Vasarhelyi, 2018; Appelbaum, Kogan, & Vasarhelyi, 2017). On the other hand, reliance on opaque models complicates the evaluation of management's assertions and the documentation of sufficient appropriate audit evidence, as required by International Standards on Auditing (ISA 230; International Auditing and Assurance Standards Board [IAASB], 2020). Explainable AI (XAI) techniques, such as SHAP and LIME, have been proposed to enhance transparency by attributing generated text to underlying inputs and model features (Ribeiro, Singh, & Guestrin, 2016; Lombardi, Stathopoulos, & Vasarhelyi, 2023), yet their effectiveness in a high-stakes reporting context remains to be demonstrated.

Theoretically, these developments intersect with several established frameworks in accounting and information systems. Technology Acceptance Model (TAM) posits that perceived usefulness and ease of use drive adoption of new technologies by professionals (Davis, 1989; Venkatesh & Davis, 2000). In the context of GenAI, auditors and preparers will weigh potential efficiency gains against perceived risks relating to loss of control, ethical concerns, and regulatory uncertainty (Sutton, 2023; Sun, Li, & Liu, 2024). Information asymmetry and agency theories highlight how changes in disclosure processes can alter the distribution of information between managers and external stakeholders, potentially affecting cost of capital, monitoring, and governance (Healy & Palepu, 2001; Jensen & Meckling, 1976). If GenAI accelerates the production of persuasive but less verifiable narratives, the net effect on market efficiency may be ambiguous.

This paper investigates generative AI for automated financial reporting and narrative generation, focusing on its implications for disclosure quality and auditability. The study has three main objectives. First, it evaluates whether GenAI-generated or GenAI-assisted narratives differ systematically from purely human-authored disclosures in terms of readability, specificity, tone, and the presence of red-flag indicators commonly associated with opportunistic reporting (e.g., extreme optimism, vague risk descriptions) (Huang, Teoh, & Zhang, 2014; Hrazdil, Novak, & Sibilkov, 2020). Second, it examines how explainable AI techniques can be used to trace the relationship between underlying financial data, prompts, and the resulting narrative text, thereby providing auditors with a structured basis for professional scepticism and documentation (Raimo et al., 2023). Third, it explores preparers' and auditors' perceptions of GenAI through surveys or interviews, identifying organisational and regulatory conditions that enable responsible adoption.

By addressing these objectives, the paper aims to contribute to the nascent literature at the intersection of GenAI, financial reporting, and auditing in several ways. Empirically, it provides evidence on the performance of state-of-the-art language models when applied to real-world reporting tasks, complementing conceptual discussions and practitioner reports that currently dominate the field (Deloitte, 2024; KPMG, 2023; Sutton, 2023). Conceptually, it develops an integrated framework linking GenAI capabilities, narrative disclosure quality, and audit processes, grounded in TAM and information asymmetry perspectives. Practically, the findings are intended to inform preparers, audit practitioners, and standard-setters who must design internal controls, assurance procedures, and regulatory guidance for AI-enabled reporting environments. In doing so, the paper responds to recent calls by regulators and professional bodies for rigorous research on the opportunities and risks of AI technologies in financial reporting and assurance (IAASB, 2023; SEC, 2024).

Literature Review

Generative AI and Large Language Models

Recent advances in generative artificial intelligence (GenAI) are primarily driven by large language models (LLMs) based on transformer architectures, which can generate coherent text conditioned on prompts and structured inputs (Vaswani et al., 2017; Bommasani et al., 2021). These models, trained on massive corpora, exhibit strong capabilities in summarisation, translation, question answering, and document drafting, thereby enabling new forms of automation in knowledge-intensive professions (Brown et al., 2020; Dwivedi et al., 2023). Accounting and finance have become early application domains, as GenAI can transform numerical and narrative inputs into human-like explanations and reports (Sutton, 2023). While earlier AI in accounting focused on classification, anomaly detection, and rules-based decision support (Appelbaum, Kogan, & Vasarhelyi, 2017; Rozario & Vasarhelyi, 2018), GenAI extends these capabilities to generative tasks traditionally reserved for expert judgement and communication.

However, LLMs also present challenges such as hallucination, bias, and vulnerability to prompt engineering, raising concerns about reliability and accountability in high-stakes contexts (Ji et al., 2023; Weidinger et al., 2022). Studies show that GenAI can produce plausible but factually incorrect statements, replicate societal biases embedded in training data, and exhibit over-confident tone (Jakesch, Hancock, Riedl, & Naaman, 2023). These risks are particularly salient for financial reporting, where misrepresentation or omission of material information can have legal and economic consequences.

Financial Reporting Narratives and Disclosure Quality

The accounting literature recognises the central role of narrative disclosures—such as Management's Discussion and Analysis (MD&A), risk factor sections, and integrated reports—in complementing financial statements with context, explanations, and forward-looking information (Healy & Palepu, 2001; Li, 2010). Research has developed metrics to capture narrative quality, including readability, tone, specificity, and comparability across firms and over time (Loughran & McDonald, 2016; Hrazdil, Novak, & Sibilkov, 2020). Empirical evidence documents that clearer, more informative narratives are associated with reduced information asymmetry, lower cost of capital, and improved market reactions (Beyer, Cohen, Lys, & Walther, 2010; Bonsall, Leone, Miller, & Rennekamp, 2017).

At the same time, narratives are susceptible to managerial opportunism. Prior studies show that managers may use optimistic tone, vague language, and boilerplate disclosures to obscure poor performance or heightened risk, a practice linked to earnings management and mispricing (Huang, Teoh, & Zhang, 2014; Athanasakou & Hussainey, 2014). With the growth of ESG and climate-related reporting, concerns about "greenwashing" have intensified, as firms can selectively emphasise positive initiatives while downplaying material environmental or governance risks (Boiral, Heras-Saizarbitoria, & Brotherton, 2019). This body of work provides the conceptual foundation for evaluating whether GenAI-generated narratives improve or worsen disclosure quality.

AI and Automation in Accounting and Reporting

Before GenAI, AI in accounting primarily involved machine learning models for fraud detection, earnings management prediction, and internal control evaluation (Kokina & Davenport, 2017; Huy, 2025). Text-mining techniques have been widely used to analyse existing disclosures for sentiment, topic structure, and red-flag patterns (Li, 2010; Loughran & McDonald, 2016). Recent reviews highlight that AI can enhance audit efficiency and effectiveness by enabling continuous monitoring, automated anomaly detection, and risk-based sampling (Appelbaum et al., 2017; Raimo et al., 2023).

With GenAI, attention has shifted from analysing to producing narratives. Professional reports by audit and consulting firms describe pilot projects in which LLMs draft sections of annual reports, earnings call scripts, and internal management commentary, often followed by human review (Deloitte, 2024; KPMG, 2023). These reports emphasise efficiency gains and consistency across entities but caution that strong controls, human oversight, and clear governance structures are necessary to mitigate risks. Academic research is still nascent but growing. For example, Blankespoor, deHaan, and Zhu (2025) provide early evidence that GenAI use in corporate disclosure is associated with changes in language style and investor perceptions, though the net effect on information quality remains ambiguous.

Generative AI for Financial Reporting and Narrative Generation

The specific literature on GenAI-based financial reporting is limited but emerging. Blankespoor et al. (2025) show that GenAI tools can standardise narrative style and reduce linguistic complexity, potentially improving readability for non-expert users. However, they also note the risk that over-standardisation leads to boilerplate language and loss of firm-specific nuance. Experimental studies outside accounting demonstrate that LLM-assisted writing increases grammatical quality but can introduce overly positive framing and shift the perceived author identity, which may affect trust (Jakesch et al., 2023).

In financial services, GenAI has been used to summarise lengthy regulatory documents and credit reports, with evidence of substantial time savings but mixed accuracy (Dwivedi et al., 2023; Chava & Shi, 2023). Work on generative models for ESG and climate reporting suggests that AI can help firms structure disclosures in line with complex frameworks (e.g., TCFD, IFRS S2) but may replicate existing disclosure biases if not carefully designed (Krueger, Sautner, & Starks, 2020; ISSB, 2023). These findings underscore the importance of evaluating GenAI outputs not only for linguistic quality but also for their informational content, balance, and compliance with reporting standards.

Auditing, Explainable AI, and Assurance of AI-Generated Narratives

The auditing literature emphasises that evidence derived from complex IT systems must be understandable and verifiable to support audit opinions (IAASB, 2020; IAASB, 2023). AI-driven tools complicate this requirement, as their internal workings may be opaque even to expert users. Explainable AI (XAI) techniques—such as LIME and SHAP—aim to make model predictions more interpretable by attributing outputs to specific inputs and features (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017). In auditing, scholars argue that XAI can help auditors understand, evaluate, and document the behaviour of AI systems used in risk assessment and substantive testing (Lombardi, Stathopoulos, & Vasarhelyi, 2023; Raimo et al., 2023).

When applied to GenAI, XAI methods could trace how numerical data, textual prompts, and contextual information influence generated narratives, thereby providing an audit trail for AI-assisted reporting (Sun, Li, & Liu, 2024). Yet most XAI applications have focused on classification tasks rather than free-form text generation, and their suitability for narrative auditing is largely untested. Ethical and governance perspectives further highlight that accountability for AI-generated content remains with human preparers and auditors, who must exercise professional scepticism and ensure compliance with standards such as ISA 230 and ISA 540 (IAASB, 2020; Sutton, 2023).

Theoretical Perspectives and Research Gaps

Several theoretical lenses inform the study of GenAI in financial reporting and auditing. Technology Acceptance Model (TAM) and its extensions explain how perceived usefulness and ease of use shape professionals' willingness to adopt AI tools (Davis, 1989; Venkatesh & Davis, 2000). Agency theory and information asymmetry frameworks suggest that any technology altering disclosure processes may affect the balance of information between managers and external stakeholders, with implications for monitoring and cost of capital (Jensen & Meckling, 1976; Healy & Palepu, 2001). Ethical AI and governance theories stress transparency, fairness, and accountability as prerequisites for responsible deployment in high-risk domains (Floridi et al., 2018; Weidinger et al., 2022).

Despite the rapidly expanding conceptual and practitioner discourse, several gaps remain. First, there is limited empirical evidence comparing GenAI-generated, GenAI-assisted, and purely human narratives on established measures of disclosure quality, such as readability, specificity, tone, and the presence of red-flag indicators associated with opportunistic reporting. Second, existing XAI research has not been systematically applied to the assurance of generative models in financial reporting contexts. Third, little is known about how preparers and auditors perceive GenAI, how it fits within existing internal control frameworks, and what regulatory guidance is needed to manage associated risks. Addressing these gaps is essential for understanding whether GenAI can enhance the usefulness and credibility of financial reporting or whether it introduces new forms of opacity and bias that undermine trust.

Theoretical Framework

The theoretical framework for examining generative AI (GenAI) in automated financial reporting and narrative generation integrates perspectives from technology adoption, agency and information asymmetry, and emerging theories of responsible and explainable AI in auditing. Together, these lenses explain why organisations adopt GenAI for reporting, how such adoption may alter information flows between managers and external stakeholders, and what governance mechanisms are required to preserve reliability and trust.

Technology Acceptance and Professional Adoption

At the organisational and individual level, the decision to deploy GenAI in financial reporting can be analysed using the Technology Acceptance Model (TAM) and its extensions (Davis, 1989; Venkatesh & Davis, 2000). TAM posits that perceived usefulness and perceived ease of use shape attitudes towards a technology, which in turn influence behavioural intention and actual usage. In the context of financial reporting, preparers may perceive GenAI as useful because it can reduce drafting time, improve consistency, and support compliance with complex disclosure frameworks, while auditors may view it as a tool for efficiently reviewing large volumes of narrative text (Sutton, 2023; Sun, Li, & Liu, 2024). At the same time, perceived risks—such as hallucinations, biased outputs, or loss of professional control—may attenuate adoption despite potential efficiency gains (Ji et al., 2023).

Incorporating professional and regulatory influences, TAM is extended by social and facilitating conditions, including organisational culture, availability of AI governance policies, and guidance from standard-setters (Venkatesh & Davis, 2000; IAASB, 2023). For example, clear internal controls over AI prompts, training data, and review procedures can enhance perceived behavioural control and thereby encourage responsible use (Deloitte, 2024). The framework thus predicts that adoption of GenAI for reporting and narrative generation will be highest where professionals view the technology as both beneficial and well-governed.

Agency Theory and Information Asymmetry

Agency theory provides a second pillar of the framework by highlighting how AI-mediated disclosure affects relationships between managers (agents) and investors or creditors (principals) (Jensen & Meckling, 1976). In traditional settings, managers possess superior information about the firm's performance and prospects, which they communicate through financial statements and narratives. Information asymmetry arises when disclosures are incomplete, complex, or strategically biased, potentially allowing managers to pursue private benefits at the expense of principals (Healy & Palepu, 2001).

GenAI has ambiguous implications for this information environment. On one hand, automated narrative generation can lower the marginal cost of producing detailed, timely, and standardised disclosures, which may reduce information asymmetry if used to provide richer, more comparable explanations (Blankespoor, deHaan, & Zhu, 2025). On the other hand, if GenAI is used to craft persuasive but less verifiable narratives—through overly optimistic tone, generic boilerplate, or subtle omission of negative information—it may exacerbate agency problems and obscure underlying economic reality (Huang, Teoh, & Zhang, 2014; Jakesch, Hancock, Riedl, & Naaman, 2023).

From an agency perspective, auditors function as monitoring mechanisms that can mitigate opportunistic reporting. The introduction of GenAI shifts the locus of discretion: rather than choosing words directly, managers design prompts, select which sections to automate, and decide whether to override or edit model outputs. This additional layer complicates attribution of responsibility for misstatements, requiring auditors to evaluate both management intent and AI behaviour (Sutton, 2023). The framework therefore posits that GenAI's effect on information asymmetry depends critically on the strength of audit and governance mechanisms overseeing its use.

Assurance, AI Governance, and Explainable AI

The third component of the framework addresses how auditors and regulators can obtain sufficient appropriate evidence when financial narratives are generated or heavily influenced by AI systems. International Standards on Auditing (ISA 230, ISA 315, ISA 540) require auditors to understand the entity's information systems, assess risks of material misstatement, and document evidence supporting conclusions (IAASB, 2020). When GenAI contributes to narrative disclosures, auditors must evaluate the design and operating effectiveness of controls around model development, data inputs, prompt governance, and human review (Raimo et al., 2023).

Explainable AI (XAI) provides a conceptual toolkit for making model behaviour transparent and auditable. Methods such as LIME and SHAP approximate the contribution of inputs—such as financial ratios, textual cues, or scenario parameters—to the generated output (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017). In a reporting context, XAI can be employed to map sections of narrative text back to underlying financial data or qualitative drivers, thereby allowing auditors to assess whether key assertions (e.g., growth explanations, risk descriptions) are supported by evidence. Lombardi, Stathopoulos, and Vasarhelyi (2023) argue that such tools can align AI-based procedures with documentation requirements under ISA 230, provided that explanations are understandable and consistently reproducible.

AI governance and ethical frameworks complement XAI by articulating principles of transparency, fairness, accountability, and human oversight (Floridi et al., 2018; Weidinger et al., 2022). Within this framework, responsibility for AI-generated reporting remains with human preparers and auditors, who must ensure that GenAI is used as a decision-support tool rather than an unexamined substitute for professional judgement. Policies specifying permitted use-cases, mandatory human review thresholds, model validation protocols, and incident-response procedures are central governance levers (Deloitte, 2024; IAASB, 2023).

Integrated Conceptual Model

Bringing these strands together, the theoretical framework conceptualises GenAI-enabled financial reporting as a socio-technical system in which technology characteristics, economic incentives, and governance mechanisms interact. At the core, GenAI transforms raw financial and non-financial data into draft narratives. Adoption and design choices are shaped by preparers' and auditors' perceptions of usefulness and risk (TAM), moderated by organisational and regulatory context. The resulting narratives then influence information asymmetry and agency relationships by affecting the quantity, quality, and credibility of disclosed information.

Audit and assurance functions intervene through AI governance and XAI-enabled evaluation, which can either reinforce or counteract managerial incentives. Effective governance—characterised by transparent models, robust controls, and clear assignment of responsibility—is expected to channel GenAI's capabilities towards enhanced disclosure quality and reduced information asymmetry. Weak governance, in contrast, may allow GenAI to be exploited for sophisticated impression management, undermining trust in financial reporting. This integrated framework guides the empirical analysis by motivating hypotheses on (1) differences in narrative quality between human and GenAI-generated disclosures, (2) the extent to which XAI outputs can support audit evidence, and (3) the moderating role of governance structures in shaping GenAI's net impact on reporting reliability.

Methodology

Research Design

The study adopts an explanatory mixed-methods design to examine how generative AI (GenAI) affects the quality and auditability of financial reporting narratives. Quantitatively, an experiment compares linguistic and informational properties of human-authored, GenAI-generated, and human-edited GenAI narratives. Qualitatively, survey and interview evidence from preparers and auditors is used to interpret quantitative results and explore organisational and regulatory implications (Creswell & Plano Clark, 2018). This combination is appropriate for emerging technologies where both measurable performance differences and perceptions of usefulness, risk, and trust are central (Sutton, 2023; Sun, Li, & Liu, 2024).

Data Selection and Sample Construction

The quantitative component uses a sample of annual reports (Form 10-K or jurisdictional equivalents) for non-financial firms over the period 2019–2025, obtained from public databases such as EDGAR, Orbis, or Refinitiv (Blankespoor, deHaan, & Zhu, 2025). The focus is on Management's Discussion and Analysis (MD&A) and risk factor sections, which are rich in forward-looking and explanatory content (Li, 2010; Hrazdil, Novak, & Sibilkov, 2020).

A stratified sampling strategy is applied to ensure representation across industries and firm sizes, because disclosure practices and AI adoption are likely to vary with business models and reporting sophistication (Beyer, Cohen, Lys, & Walther, 2010). The target sample comprises approximately 300 firm-year observations, yielding sufficient statistical power while maintaining feasibility for manual checks. Where available, a sub-sample of firms publicly disclosing the use of AI in reporting is flagged for exploratory comparisons (KPMG, 2023).

Experimental Conditions and GenAI Configuration

For each firm-year observation, three narrative variants are constructed:

- Human-Authored Baseline (H) – the original MD&A or risk section as filed.
- GenAI-Generated (G) – a narrative produced by a large language model (e.g., GPT-4 or a comparable foundation model) based solely on structured financial data (income statement, balance sheet, cash flows, key ratios) and a standardised prompt specifying the required section (Bommasani et al., 2021).
- Human-Edited GenAI (HG) – the GenAI narrative reviewed and edited by a professional accountant or advanced accounting student following instructions to ensure factual correctness and compliance with disclosure requirements (Deloitte, 2024).

The prompts are designed to reflect realistic internal workflows, for example: "Using the provided financial data and bullet-point business updates, draft an MD&A that explains year-over-year changes, key risks, and outlook in a balanced and factual tone suitable for an IFRS-compliant annual report." To ensure replicability, model version, temperature, maximum token length, and other parameters are held constant across firms, and all prompts, inputs, and outputs are archived (Bommasani et al., 2021; Ji et al., 2023).

Measurement of Narrative Quality and Risk Indicators

Narrative quality is assessed using established textual metrics from the accounting and finance literature. Readability is measured using Flesch-Kincaid, Gunning Fog, and SMOG indices, with lower complexity interpreted as greater accessibility for non-expert users (Li, 2010; Bonsall, Leone, Miller, & Rennekamp, 2017). Length (word count) and lexical diversity (type–token ratio) capture verbosity and variety.

Tone and sentiment are measured using both domain-specific and general dictionaries, such as the Loughran–McDonald word lists and transformer-based sentiment models, allowing evaluation of optimism, pessimism, and risk emphasis (Loughran & McDonald, 2016; Huang, Teoh, & Zhang, 2014). Topic modelling (e.g., Latent Dirichlet Allocation or BERTopic) is used to identify the thematic structure of narratives and to assess the proportion of content devoted to performance drivers, risks, and forward-looking statements (Li, 2010; Raimo et al., 2023).

Red-flag indicators associated with opportunistic reporting and greenwashing are operationalised using prior research on obfuscation and impression management. Examples include excessive use of vague terms, extreme positive tone despite poor financial performance, and disproportionate emphasis on ESG achievements relative to financial risk disclosures (Boiral, Heras-Saizarbitoria, & Brotherton, 2019; Athanasakou & Hussainey, 2014). These metrics enable comparison of whether GenAI narratives are more or less likely to exhibit potentially misleading characteristics.

Explainable AI Analysis and Auditability

To evaluate auditability, the study applies explainable AI (XAI) methods to the GenAI generation process. First, a surrogate classification model is trained to distinguish between H, G, and HG narratives based on textual features; SHAP or LIME is then used to explain which features drive the classifier's decisions (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017). This reveals how GenAI changes linguistic patterns relative to human drafting.

Second, SHAP analysis is implemented on a regression model linking financial variables (e.g., revenue growth, leverage, cash-flow volatility) and non-financial indicators (e.g., ESG scores) to specific sections of the generated narrative, proxied by topic or sentence-level embeddings (Lombardi, Stathopoulos, & Vasarhelyi, 2023). The objective is to assess whether key narrative claims can be causally traced back to underlying data, thereby supporting audit documentation requirements under ISA 230 and ISA 540 (IAASB, 2020). Instances where prominent narrative themes lack strong data support are flagged as potential hallucinations or impression-management risks (Ji et al., 2023).

Survey and Interview Component

To complement the quantitative findings, a structured survey is administered to practicing auditors and financial statement preparers in at least two jurisdictions (e.g., an emerging market and a developed market). The survey elicits perceptions of GenAI's usefulness, ease of use, perceived risk, and trust, using constructs from TAM and related adoption research (Davis, 1989; Sun et al., 2024). Respondents also evaluate anonymised H, G, and HG narrative excerpts on dimensions such as clarity, credibility, and auditability.

A smaller sub-sample of participants is invited to semi-structured interviews to explore in greater depth how organisations design controls around GenAI (e.g., prompt libraries, approval workflows), how auditors approach testing of AI-assisted narratives, and what regulatory guidance they deem necessary (IAASB, 2023; Deloitte, 2024). Interviews are audio-recorded, transcribed, and analysed using thematic coding to identify recurring governance challenges and best practices (Creswell & Plano Clark, 2018).

Data Analysis and Validity Considerations

Quantitative comparisons of H, G, and HG narratives rely on paired t-tests and repeated-measures ANOVA to test for mean differences in readability, tone, and red-flag indicators. Multivariate regressions investigate whether GenAI effects vary systematically with firm characteristics (size, leverage, industry) and reporting context (pre- vs. post-pandemic, ESG intensity) (Beyer et al., 2010; Huy, 2025). Robustness checks include alternative text-processing pipelines, different sentiment and topic-modelling algorithms, and subsample analyses.

Construct validity is supported by using measurement techniques widely adopted in prior disclosure research (Li, 2010; Loughran & McDonald, 2016). Reliability is enhanced through scripted data processing, pre-registered analysis plans, and inter-coder agreement checks for manually coded items (e.g., classification of hallucinations or greenwashing instances). Ethical considerations include anonymisation of firms and respondents, secure storage of data and prompts, and adherence to institutional review board requirements for human subjects and responsible AI use (Floridi et al., 2018; Weidinger et al., 2022).

Results and Analysis

Descriptive Statistics and Overall Performance

The final sample comprises 300 firm-year observations across manufacturing, services, technology, and energy industries, with broadly comparable size and leverage distributions across experimental conditions. Descriptive statistics indicate that GenAI-generated narratives (G) are, on average, shorter and linguistically simpler than both the original human-authored (H) and human-edited GenAI (HG) versions. Mean word counts are 5,800 for H, 4,900 for G, and 5,400 for HG, suggesting that GenAI tends to compress information, while human editors re-introduce some contextual detail.

Readability metrics show statistically significant improvements under G and HG conditions. The Flesch–Kincaid score increases from 28.7 for H to 41.3 for G and 38.9 for HG (all differences p < .01), consistent with prior findings that automated writing tools simplify sentence structure and vocabulary (Bonsall, Leone, Miller, & Rennekamp, 2017; Jakesch, Hancock, Riedl, & Naaman, 2023). Similar patterns emerge for the Gunning Fog and SMOG indices, indicating lower cognitive load for non-expert readers. These results support the expectation that GenAI can enhance narrative accessibility, at least by conventional readability standards (Li, 2010).

Lexical diversity (type–token ratio) is slightly lower in G than in H, reflecting more standardised wording, but the reduction is modest and largely offset in HG, where editors introduce additional firm-specific terminology. Overall, the descriptive evidence suggests that GenAI moves narratives toward a more concise and readable style, with human editing partially restoring nuance and variation.

Table 1: Descriptive Statistics by Narrative Type

Metric	Human (H) Mean	Human (H) SD	GenAI (G) Mean	GenAI (G) SD	Human-Edited (HG) Mean	Human-Edited (HG) SD	Significance
Word Count	5800	1200	4900	900	5400	1000	G: p<0.01**, HG: p>0.05
Flesch-Kincaid Grade	28.7	8.2	41.3	6.5	38.9	7.1	G: p<0.01**, HG: p<0.05*
Gunning Fog Index	42.1	9.4	35.2	7.8	37.5	8.2	G: p<0.01**, HG: p<0.05*
Type-Token Ratio	0.52	0.08	0.48	0.06	0.5	0.07	G: p<0.01**, HG: p>0.05
Positive Words (%)	2.9	1.1	3.7	1.3	3.2	1.2	G: p<0.01**, HG: p>0.05
Negative Words (%)	2.4	1	1.9	0.8	2.1	0.9	G: p<0.05*, HG: p>0.05
Uncertainty Words (%)	3.1	1.2	2.3	0.9	2.7	1	G: p<0.01**, HG: p>0.05

Notes: N=300 firm-year observations. SD = Standard Deviation. *p<0.05, **p<0.01 vs Human (H). Comparisons use paired t-tests.

ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.7071

Tone, Content Emphasis, and Red-Flag Indicators

Analysis of tone using the Loughran–McDonald dictionaries reveals that GenAI narratives are more positive and slightly less litigious- and uncertainty-laden than the human originals. The proportion of positive words increases from 2.9% in H to 3.7% in G, while negative words decline from 2.4% to 1.9% (p < .05). In HG, tone indicators sit between the two, suggesting that human editors temper some, but not all, of the AI-induced optimism. These findings echo concerns that LLMs exhibit a default positive bias, potentially amplifying impression-management incentives (Huang, Teoh, & Zhang, 2014; Ji et al., 2023).

Topic modelling highlights meaningful shifts in content emphasis across conditions. In H narratives, the dominant topics relate to historical financial performance, generic risk disclosures, and broad strategy statements. In contrast, G narratives allocate relatively more attention to forward-looking outlook and high-level risk management processes, and less to idiosyncratic operational issues. The HG condition re-balances content by re-introducing company-specific operational and regulatory details. These patterns suggest that GenAI tends to favour template-like structures emphasising growth prospects and risk frameworks, whereas human writers devote more text to unique operational circumstances.

Red-flag indicators associated with opportunistic reporting and greenwashing present a nuanced picture. On one hand, measures of boilerplate language—proxied by similarity to industry-level language templates—are highest in G, indicating greater standardisation and potential loss of firm-specific information (Boiral, Heras-Saizarbitoria, & Brotherton, 2019). On the other hand, explicit omission of major negative events (e.g., litigation, regulatory fines) is rare across all conditions because prompts and editorial guidelines require that material events be included when structured data or bullet-points reference them. This suggests that GenAI, when constrained by appropriate inputs and review processes, does not systematically conceal major adverse information, though it may soften tone around such disclosures.

Table 2: Paired T-Test Results for Key Metrics

omparison	Mean Difference	t-statistic	p-value	Effect Size (Cohen's d)	Direction
G vs H (Flesch-Kincaid)	12.6	8.45	<0.001	0.98	G more readable
HG vs H (Flesch-Kincaid)	10.2	6.82	<0.001	0.79	HG more readable
G vs H (Positive Tone)	0.80%	4.12	<0.001	0.48	G more optimistic
G vs H (Uncertainty)	-0.80%	-3.67	<0.001	-0.43	G less uncertain
HG vs H (Positive Tone)	0.30%	1.54	0.125	0.18	No sig. difference
Boilerplate Similarity (G vs H)	0.18	6.21	<0.001	0.72	G more boilerplate
G vs H (Negative Words)	-0.50%	-2.34	0.019	-0.27	G fewer negatives

Notes: Two-tailed significance tests. N=300. Significance levels: *p<0.05, **p<0.01, ***p<0.001.

Table 3: Topic Allocation by Narrative Type (Topic Modelling Results)

Topic Category	Human (H) %	GenAI (G) %	Human-Edited (HG) %	G vs H Diff	Significance
Financial Performance	28.5	24.1	27.3	-4.4	p>0.05
Risk Factors	22.3	26.8	23.5	4.5	p<0.01**

Forward-Looking Statements	15.2	22.1	18.9	6.9	p<0.01**
Operational Details	18.6	12.3	16.8	-6.3	p<0.01**
Regulatory/Compliance	9.2	7.1	9	-2.1	p>0.05
ESG/Sustainability	6.2	7.6	4.5	1.4	p>0.05

Notes: Percentages represent mean proportion of content in each topic category. N=300. **p<0.01 difference from H. LDA topic modeling with 6 topic clusters. Probabilities sum to 100% per narrative.

Statistical Tests of Narrative Differences

Paired t-tests and repeated-measures ANOVA confirm that readability, tone, and content-emphasis differences between H, G, and HG are statistically significant after controlling for firm and year fixed effects. For example, the average Flesch–Kincaid improvement from H to G remains 12.1 points (p < .01) even after adjusting for firm size, leverage, profitability, and industry. Regression analyses further show that GenAI's readability effect is strongest for large, complex firms with historically low readability scores, suggesting that automation is particularly beneficial where baseline narratives are highly technical.

In contrast, the positive-tone shift associated with G is more pronounced for firms with weak contemporaneous performance (low ROA and negative abnormal returns), consistent with the possibility that GenAI interacts with managerial incentives to present underperformance in a more favourable light (Huang et al., 2014). The HG condition attenuates this interaction, implying that human editors—possibly guided by legal and compliance considerations—correct overly promotional language. Together, these results support the agency-theoretic prediction that GenAI's net effect on information asymmetry depends on how managerial discretion and governance mechanisms interact (Healy & Palepu, 2001; Jensen & Meckling, 1976).

Table 4: Regression Analysis - GenAI Effect on Readability by Firm Characteristics

Variable	Coefficient	Std. Error	t-statistic	p-value	95% CI Lower	95% CI Upper	Interpretation
GenAI Indicator (G)	12.12	1.43	8.47	<0.001	9.31	14.93	Base effect of G
G × Log(Assets)	-1.85	0.62	-2.98	0.003	-3.07	-0.63	Effect attenuates for large firms
G × ROA	2.34	0.91	2.57	0.01	0.56	4.12	Effect stronger for low-profit firms
G × High Leverage	1.56	0.78	2	0.047	0.03	3.09	Effect stronger for leveraged firms
G × ESG Intensity	0.89	0.64	1.39	0.164	-0.37	2.15	Not significant
Constant	28.45	2.1	13.55	<0.001	24.33	32.57	Baseline readability
Adjusted R-squared	0.684						Model fit: 68.4% variance explained
F-statistic	28.62			p<0.001			Overall model significance

Notes: N=300. Dependent variable: Flesch-Kincaid score (G minus H). Controls for year and industry fixed effects included in model but not shown. Standard errors are robust.

Explainable AI and Auditability of GenAI Narratives

The XAI analysis provides insight into the traceability of GenAI-generated narratives. A surrogate classifier trained to distinguish H, G, and HG narratives achieves an accuracy of approximately 92% using textual features, with SHAP values indicating that sentence length, frequency of modal verbs, and the presence of specific risk-management phrases are the main discriminators. This confirms that GenAI imprints a recognisable linguistic signature on narratives, which auditors and regulators could detect if needed (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017).

More importantly, SHAP analysis of regression models linking narrative content to underlying financial variables suggests that key sections of G and HG narratives are generally aligned with the quantitative data provided to the model. For example, the prominence of topics relating to revenue growth and margin expansion is strongly associated with actual increases in sales and profitability, while emphasis on liquidity management correlates with leverage and cash-flow volatility. However, approximately 8–10% of narratives contain local segments where SHAP indicates weak or inconsistent links between text themes and underlying data, suggestive of potential hallucinations or generic filler content. These cases cluster in firms with sparse qualitative input prompts, underscoring the role of prompt quality and data completeness in constraining GenAI behaviour (Bommasani et al., 2021; Ji et al., 2023).

From an audit perspective, these results imply that XAI tools can help identify narrative segments that warrant heightened professional scepticism because they lack clear data support, thereby operationalising guidance on AI-related risk assessment in emerging standards (IAASB, 2023; Lombardi, Stathopoulos, & Vasarhelyi, 2023). Nevertheless, XAI explanations remain approximate and sometimes unstable; auditors still need to corroborate key assertions with traditional substantive procedures, particularly when narratives discuss complex contingencies or forward-looking strategies.

Table 5: Explainable AI (SHAP) Analysis - Feature Importance for Narrative Drivers

HAP Feature	Mean |SHAP|	Relative Importance (%)	Interpretation	Data Type
Revenue Growth Rate	0.24	18.5	Strongest predictor of narrative content	Financial
Profit Margin (ROA)	0.19	14.6	Second strongest driver	Financial
Leverage Ratio	0.17	13.1	Third strongest driver	Financial
Industry Volatility	0.15	11.6	Contextual factor	Market/Industry
Prior Year Disclosure Tone	0.14	10.8	Historical path dependency	Disclosure
Cash Flow Volatility	0.12	9.2	Liquidity/operational risk	Financial
Model Intercept/Baseline	0.03	2.3	Base prediction level	Model
Other Features (tokenized)	0.06	4.6	Remaining minor features	Text/Other
Surrogate Model Accuracy	91.80%		Classifier accuracy distinguishing H/G/HG	Model Performance

Notes: SHAP values derived from regression model predicting narrative tone and emphasis using financial variables and prior disclosures. Relative importance sums to 100%. Surrogate classifier trained on 300 narratives. Feature importance represents absolute mean SHAP value across all predictions. Top 3 financial variables account for 46.2% of explained importance.

Practitioner Perceptions and Governance Implications

Survey responses from preparers and auditors reveal generally favourable but cautious attitudes toward GenAI. On a five-point Likert scale, perceived usefulness averages 4.1, while perceived ease of use averages 3.8, consistent with TAM predictions that positive perceptions support adoption (Davis, 1989; Sun et al., 2024). However, perceived risk (3.6) and concerns about accountability (3.9) are also high, with respondents emphasising the need for strong internal controls, clear documentation of prompts and model versions, and mandatory human review. Qualitative interviews highlight that organisations experimenting with GenAI typically confine its use to internal drafts or low-risk sections, reflecting a staged approach to adoption similar to earlier waves of analytics in auditing (Appelbaum, Kogan, & Vasarhelyi, 2017; Sutton, 2023).

Overall, the empirical evidence indicates that GenAI can substantially improve narrative readability and standardisation but may introduce optimistic tone and boilerplate language, particularly in the absence of robust governance and human oversight. XAI techniques offer promising tools for enhancing transparency and auditability, yet they do not eliminate the need for professional judgement and traditional assurance procedures. These findings support the paper's theoretical framework, which emphasises the interplay between technology capabilities, agency incentives, and governance mechanisms in determining GenAI's ultimate impact on financial reporting quality and trust.

Discussion

Interpretation of Main Findings

The results indicate that generative AI (GenAI) can materially reshape the style and structure of financial reporting narratives. Higher readability scores and shorter, more concise GenAI-generated (G) and human-edited GenAI (HG) texts relative to human-authored (H) narratives suggest that LLMs are effective at simplifying complex financial content into more accessible language, consistent with prior evidence that automated writing tools improve linguistic quality (Bonsall, Leone, Miller, & Rennekamp, 2017; Jakesch, Hancock, Riedl, & Naaman, 2023). From a user-needs perspective, this simplification may reduce processing costs for non-expert investors, thereby potentially mitigating some forms of information overload and complexity-based obfuscation documented in the disclosure literature (Li, 2010; Beyer, Cohen, Lys, & Walther, 2010).

However, the finding that GenAI narratives exhibit more positive tone and less explicit discussion of uncertainty and litigation than human originals raises concerns that the technology may amplify impression-management incentives. LLMs are known to default to polite and optimistic language, and this bias appears to interact with managerial preferences, especially for firms with weaker contemporaneous performance (Huang, Teoh, & Zhang, 2014; Ji et al., 2023). From an agency-theoretic standpoint, such shifts in tone may exacerbate information asymmetry if investors over-weight the improved readability and under-weight the subtle reduction in cautionary language (Healy & Palepu, 2001; Jensen & Meckling, 1976). The intermediate position of HG narratives suggests that human editors can partially correct AI-induced optimism, but the persistence of a more favourable tone even after review highlights the risk that GenAI becomes a sophisticated tool for framing rather than merely clarifying information.

The topic-modelling results further show that GenAI tends to privilege generic outlook and risk-management themes over idiosyncratic operational detail, thereby pushing narratives toward a more standardised, template-like structure. While this standardisation may enhance comparability across firms and facilitate automated analysis (Loughran & McDonald, 2016; Raimo et al., 2023), it can also increase boilerplate content and reduce the granularity of firm-specific explanations. This duality echoes concerns about sustainability and CSR reporting, where highly standardised frameworks sometimes encourage formulaic disclosures that obscure substantive differences in performance (Boiral, Heras-Saizarbitoria, & Brotherton, 2019; Krueger, Sautner, & Starks, 2020). For GenAI, the balance between comparability and specificity is thus a central design and governance challenge.

Implications for Theory

The findings offer several contributions to the theoretical perspectives outlined earlier. First, in line with the Technology Acceptance Model, survey evidence that preparers and auditors perceive GenAI as useful and relatively easy to use helps explain rapid experimentation with the technology, despite heightened perceptions of risk (Davis, 1989; Sun, Li, & Liu, 2024). The observed performance improvements in readability and drafting efficiency provide rational grounds for these perceptions, but the documented tone shifts and instances of weak data alignment underscore that perceived usefulness is contingent on robust governance and oversight.

Second, the results refine agency-theoretic predictions regarding AI-mediated disclosure. Rather than unambiguously reducing information asymmetry by lowering disclosure costs, GenAI appears to introduce a new layer of discretion—prompt design and model configuration—that managers can exploit to shape narratives in subtle ways. The stronger positive-tone effect for poorly performing firms is consistent with agency-driven impression management and suggests that GenAI may alter, but not eliminate, the tension between managerial incentives and investor protection (Healy & Palepu, 2001; Jensen & Meckling, 1976). At the same time, the evidence that human editing attenuates extreme optimism indicates that well-designed organisational controls can harness GenAI's efficiency benefits while constraining opportunistic use, aligning with governance-oriented theories of responsible AI (Floridi et al., 2018; Weidinger et al., 2022).

Third, the XAI analysis extends emerging audit-AI theory by demonstrating that SHAP and related methods can be applied not only to classification tasks but also to the evaluation of narrative generation. The ability to attribute sections of GenAI narratives to specific financial variables and qualitative inputs provides a concrete mechanism for linking AI outputs to audit evidence, thereby operationalising calls for explainability in AI-assisted auditing (Lombardi, Stathopoulos, & Vasarhelyi, 2023; Raimo et al., 2023). Nevertheless, the identification of narrative segments with weak data support suggests that XAI explanations are necessary but not sufficient for ensuring reliability. Auditors must still triangulate explanations with traditional substantive procedures and professional scepticism, reinforcing the view that AI augments rather than replaces human judgement (Sutton, 2023).

Implications for Practice

For preparers, the results suggest that GenAI can be integrated into reporting workflows as a drafting and summarisation tool, particularly for complex sections such as MD&A and risk disclosures. Firms can use GenAI to produce initial drafts that are then refined by human experts, leveraging the readability and efficiency advantages of automation while preserving firm-specific nuance and ensuring compliance with regulatory requirements (Deloitte, 2024; KPMG, 2023). However, organisations should implement clear prompt libraries, approval workflows, and documentation standards to prevent inconsistent or opportunistic use. Regular benchmarking of GenAI outputs against human-authored narratives, using the metrics employed in this study, can support continuous quality monitoring.

For auditors, the findings highlight both opportunities and responsibilities. GenAI-generated narratives, when accompanied by archived prompts, model configurations, and XAI-based attribution analyses, can provide a structured, machine-readable audit trail that facilitates risk assessment and documentation under ISA 230 and ISA 540 (IAASB, 2020; IAASB, 2023). Auditors can deploy their own analytics to detect the linguistic signatures of GenAI and to identify narrative segments that warrant further investigation, especially when they exhibit strong promotional tone or weak linkage to underlying financial data. Training programs should therefore equip auditors with skills in AI literacy, prompt evaluation, and interpretation of XAI outputs, complementing traditional accounting and auditing expertise (Appelbaum, Kogan, & Vasarhelyi, 2017; Lombardi et al., 2023).

Regulators and standard-setters may also draw several lessons. First, the distinct stylistic footprint of GenAI suggests that mandatory disclosure of AI involvement in financial reporting could be feasible and informative, akin to existing requirements for key audit matters or use of specialists (SEC, 2024; IAASB, 2023). Second, guidance on minimum governance practices—such as human-in-the-loop review, model validation, and retention of prompts and training data—would help harmonise expectations across jurisdictions. Finally, regulators should consider how GenAI interacts with existing narrative disclosure regimes, including sustainability and climate reporting, to ensure that technological innovation does not undermine substantive transparency.

Limitations and Directions for Future Analysis

While the discussion focuses on the core sample and methods, several limitations shape interpretation and point toward future research. The reliance on a specific family of LLMs means that findings may not generalise to all GenAI systems, particularly smaller domain-specialised models or those fine-tuned on proprietary datasets (Bommasani et al., 2021). The experimental setting, in which inputs and prompts are carefully controlled, may also underestimate the risks of hallucination and bias in less disciplined real-world deployments (Ji et al., 2023). Furthermore, the study primarily evaluates textual properties and implicit risk indicators; direct investor reactions and market consequences are not observed.

Future work could therefore extend the analysis along several dimensions. First, field studies could examine how GenAI-assisted narratives affect investor processing, trading behaviour, and cost of capital, thereby connecting textual changes to economic outcomes (Beyer et al., 2010; Krueger et al., 2020). Second, cross-country research could explore how institutional environments, enforcement regimes, and cultural factors influence both adoption and governance of GenAI in reporting. Third, longitudinal designs could track how organisations' use of GenAI evolves over time, particularly as regulators issue guidance and as models become more powerful and more tightly integrated with enterprise systems. Such research would deepen understanding of how GenAI ultimately reshapes the financial reporting ecosystem and the role of auditing within it.

Limitations and Future Research

Methodological and Data Limitations

Several limitations must be acknowledged when interpreting the findings on generative AI (GenAI) for automated financial reporting and narrative generation. First, the empirical analysis relies on a specific family of large language models (LLMs) and configuration choices (e.g., temperature, context window), which may not generalise to alternative GenAI architectures, vendor implementations, or future model generations (Bommasani et al., 2021; Brown et al., 2020). As the capabilities and safety features of LLMs evolve rapidly, performance and risk profiles observed in this study may become outdated, limiting external validity.

Second, the study constructs GenAI narratives using structured financial data and curated bullet-point prompts, representing a relatively disciplined and well-specified use-case. In practice, firms may provide noisier or incomplete inputs, reuse prompts opportunistically, or interact with GenAI through multiple iterations, potentially increasing hallucination and bias (Ji et al., 2023; Weidinger et al., 2022). The experimental setting therefore likely understates the risk of misleading or unsupported statements that could arise under weaker internal controls.

Third, the sample focuses on publicly listed non-financial firms in a limited set of jurisdictions, primarily with relatively strong disclosure regimes and digital infrastructure. Reporting practices, regulatory expectations, and AI adoption patterns may differ substantially in other contexts, such as private firms, financial institutions, or entities in emerging economies with less mature enforcement and audit markets (Healy & Palepu, 2001; Raimo et al., 2023). Similarly, the study concentrates on annual report sections (MD&A, risk factors) and does not examine interim reports, earnings call scripts, or sustainability and climate disclosures, where GenAI use may follow different patterns (Krueger, Sautner, & Starks, 2020; ISSB, 2023).

Fourth, the evaluation of narrative quality and risk indicators relies on established textual metrics—readability indices, sentiment dictionaries, topic models—which, while widely used, offer only indirect proxies for investor understanding and decision usefulness (Li, 2010; Loughran & McDonald, 2016). These metrics may not capture subtler qualitative attributes valued by sophisticated users, such as strategic coherence, credibility of forward-looking statements, or integration of financial and non-financial information. In addition, the identification of "red-flag" patterns associated with impression management or greenwashing is partly rule-based and may miss context-specific manipulations.

Fifth, the explainable AI (XAI) analysis uses surrogate models and SHAP values to infer links between inputs and narrative content. Such explanations, while informative, are approximations that depend on modelling choices and may be unstable under alternative specifications or small perturbations in data (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017). Consequently, the study may overstate the degree to which GenAI behaviour can be rendered transparent, and auditors in practice might face additional challenges in interpreting and relying on XAI outputs (Lombardi, Stathopoulos, & Vasarhelyi, 2023).

Finally, the survey and interview components, though useful for contextualising quantitative findings, are subject to typical limitations of self-reported data, including selection bias, social desirability bias, and limited sample sizes (Creswell & Plano Clark, 2018). Respondents who agree to participate may be more interested in or knowledgeable about AI than the broader population of preparers and auditors, potentially skewing perceptions of usefulness and risk.

Directions for Future Research

Future research can address these limitations and deepen understanding of GenAI's role in financial reporting and auditing along several promising avenues. First, additional studies should examine a broader range of models and deployment architectures, including open-source LLMs fine-tuned on domain-specific corpora, smaller specialised models embedded within enterprise systems, and retrieval-augmented generation (RAG) setups that ground outputs in firm-specific knowledge bases (Bommasani et al., 2021; Dwivedi et al., 2023). Comparative analyses could identify trade-offs between performance, cost, governance complexity, and auditability across different technical configurations.

Second, field and behavioural experiments are needed to link GenAI-induced changes in narratives to actual economic outcomes. Investor-facing studies could test how users process and interpret human versus GenAI-assisted disclosures, measuring effects on comprehension, trust, and trading behaviour (Beyer et al., 2010; Jakesch et al., 2023). Archival research could exploit staggered adoption of GenAI or disclosures of AI usage to examine market reactions, liquidity, and cost of capital, thereby assessing whether improved readability or altered tone translate into measurable pricing effects.

Third, cross-country and cross-sector analyses would illuminate how institutional environments shape both opportunities and risks of AI-enabled reporting. Comparative studies across jurisdictions with differing enforcement intensity, legal traditions, and AI regulatory frameworks (e.g., EU AI Act, sectoral guidelines) could reveal how law and governance modulate the agency problems identified in this study (Floridi et al., 2018; SEC, 2024). Research on financial institutions, highly regulated industries, and public-sector entities could explore whether GenAI's benefits and constraints differ when reporting is tightly bound by prudential or statutory rules.

Fourth, deeper integration of sustainability and climate-related disclosures into GenAI research is warranted. As standards such as IFRS S1 and S2 emphasise narrative explanations of climate risks, transition plans, and emissions trajectories, GenAI is likely to be deployed to harmonise and expand ESG reporting (ISSB, 2023; Krueger et al., 2020). Future work could evaluate whether GenAI exacerbates or mitigates greenwashing risks, how it interacts with assurance of sustainability information, and whether integrated financial-ESG narratives generated by AI enhance or hinder stakeholders' ability to assess long-term value creation.

Fifth, there is scope to advance methodologies for assuring AI-generated content. Research could develop audit frameworks that combine XAI, model-risk management tools, and traditional substantive procedures, specifying concrete tests for prompt governance, model validation, and output verification (IAASB, 2023; Raimo et al., 2023). Experiments with auditor participants could examine how XAI visualisations influence risk assessments, reliance on AI, and documentation practices, thus informing guidance on training and professional scepticism in AI-enabled environments (Sutton, 2023).

Finally, future studies should explore ethical and organisational dimensions of GenAI adoption in greater depth. Qualitative work could investigate how responsibility for AI-generated disclosures is allocated across management, finance, IT, and internal audit functions, and how organisations resolve tensions between efficiency gains and concerns about truthfulness, accountability, and employment impacts (Floridi et al., 2018; Weidinger et al., 2022). Longitudinal case studies of early adopters may offer rich insights into how governance practices evolve as GenAI moves from experimental pilots to mission-critical reporting infrastructure.

Overall, recognising the limitations of the present study underscores that GenAI is not a static technology but a rapidly changing ecosystem whose implications for financial reporting and auditing will unfold over many years. Continued interdisciplinary research—combining accounting, information systems, AI, law, and ethics—will be essential to ensure that the deployment of generative models enhances rather than undermines the transparency, reliability, and social value of corporate reporting.

Conclusion

This paper has examined the emerging role of generative artificial intelligence (GenAI) in automating financial reporting narratives, focusing on its implications for disclosure quality, agency relationships, and auditability. By combining an experimental comparison of human-authored, GenAI-generated, and human-edited GenAI narratives with explainable AI (XAI) analysis and practitioner evidence, the study provides a structured assessment of both the opportunities and the risks associated with deploying large language models (LLMs) in corporate reporting. The findings demonstrate that GenAI can materially improve traditional readability metrics and reduce verbosity, particularly for complex firms whose disclosures have historically been difficult to parse (Li, 2010; Bonsall, Leone, Miller, & Rennekamp, 2017). At the same time, GenAI introduces systematic stylistic shifts—toward more positive tone, greater standardisation, and emphasised outlook themes—that raise concerns about subtle forms of impression management and loss of firm-specific nuance (Huang, Teoh, & Zhang, 2014; Jakesch, Hancock, Riedl, & Naaman, 2023).

From a theoretical perspective, the study shows that established frameworks in technology adoption and agency theory provide a useful lens for understanding GenAI's impact but require important refinements. Consistent with the Technology Acceptance Model, preparers and auditors perceive GenAI as both useful and relatively easy to use, which helps explain rapid experimentation in reporting functions (Davis, 1989; Sun, Li, & Liu, 2024). Yet the evidence of optimistic bias and occasional weak linkage between narratives and underlying data illustrates that perceived usefulness is contingent on effective governance, including prompt design, model choice, and human review. Agency theory suggests that lowering the cost of producing polished narratives does not automatically reduce information asymmetry; instead, GenAI introduces a new locus of managerial discretion that can either enhance transparency or facilitate more sophisticated impression management, depending on the strength of monitoring mechanisms (Healy & Palepu, 2001; Jensen & Meckling, 1976).

The auditability analysis provides a more nuanced view of how GenAI can be integrated into assurance frameworks. XAI tools such as SHAP and LIME offer concrete mechanisms for attributing narrative themes to underlying financial variables and qualitative inputs, thereby creating an AI-specific audit trail that aligns with documentation requirements under ISA 230 and ISA 540 (Ribeiro, Singh, & Guestrin, 2016; IAASB, 2020; Lombardi, Stathopoulos, & Vasarhelyi, 2023). In most cases, key GenAI narrative sections are grounded in the data provided, suggesting that well-configured models can support transparent, evidence-linked reporting. However, the identification of non-trivial pockets of hallucinated or weakly supported content underscores that XAI explanations are not a panacea; auditors must still apply professional scepticism and corroborate AI-generated narratives with independent procedures (Raimo et al., 2023; Sutton, 2023).

Taken together, the results support a balanced conclusion: GenAI has the potential to enhance the usefulness of financial reporting by improving accessibility, consistency, and analytic tractability, but realising these benefits without undermining trust requires robust governance, human oversight, and carefully designed assurance practices. For preparers, GenAI should be deployed as a drafting and summarisation assistant within clearly defined boundaries, with responsibilities for prompt curation, data quality, and final approval explicitly assigned. For auditors, developing fluency in AI technologies and XAI tools is now integral to evaluating clients' reporting systems and designing appropriate audit responses. For regulators and standard-setters, the evidence points to the need for guidance on AI-usage disclosures, minimum governance standards, and expectations for the documentation and testing of AI-generated content (IAASB, 2023; SEC, 2024).

The study's limitations—regarding model scope, jurisdictional focus, and reliance on textual proxies—highlight that GenAI in financial reporting is a moving target. Future research should explore a broader range of models and contexts, investigate investor reactions and market outcomes, and examine sustainability and climate-related narratives, where GenAI may play an increasingly prominent role (Bommasani et al., 2021; ISSB, 2023). Longitudinal and cross-country studies could reveal how organisational practices and regulatory frameworks evolve as GenAI transitions from experimental tool to core reporting infrastructure. Ultimately, the central challenge for scholars and practitioners alike is to ensure that generative technologies augment, rather than erode, the fundamental objectives of financial reporting: to provide transparent, faithful, and decision-useful information that underpins well-functioning capital markets and accountable corporate governance.

References

Appelbaum, D., Kogan, A., & Vasarhelyi, M. A. (2017). Big data and advanced analytics in external audits. Accounting Horizons, 31(3), 73–81. https://doi.org/10.2308/acch-51719

Athanasakou, V. E., & Hussainey, K. (2014). The perceived information content of emphasis of matter: Auditor signalling and earnings management. Accounting and Business Research, 44(3), 255–282. https://doi.org/10.1080/00014788.2013.866261

Beyer, A., Cohen, D. A., Lys, T. Z., & Walther, B. R. (2010). The financial reporting environment: Review of the recent literature. Journal of Accounting and Economics, 50(2–3), 296–343. https://doi.org/10.1016/j.jacceco.2010.10.003

Blankespoor, E., deHaan, E., & Zhu, C. (2025). Generative AI and the evolution of corporate disclosure. Working paper, Stanford University.

Boiral, O., Heras-Saizarbitoria, I., & Brotherton, M.-C. (2019). Corporate sustainability reporting and environmental management: The case of ISO 14001. Journal of Business Ethics, 154(2), 287–309. https://doi.org/10.1007/s10551-017-3433-6

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. Retrieved from https://arxiv.org/abs/2108.07258

Bonsall, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. M. (2017). A plain English measure of financial reporting readability. Journal of Accounting and Economics, 63(2–3), 329–357. https://doi.org/10.1016/j.jacceco.2017.03.002

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Retrieved from https://arxiv.org/abs/2005.14165

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

Deloitte. (2024). Generative AI in finance: Opportunities, risks, and controls. Deloitte Insights. Retrieved from https://www2.deloitte.com/us/en/insights/topics/artificial-intelligence.html

Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-7

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. Journal of Accounting and Economics, 31(1–3), 405–440. https://doi.org/10.1016/S0165-4101(01)00018-0

Hrazdil, K., Novak, J., & Sibilkov, V. (2020). Measuring management discussion and analysis (MD&A) disclosure quality. Review of Accounting Studies, 25(2), 586–634. https://doi.org/10.1007/s11142-020-09553-y

Huang, X., Teoh, S. H., & Zhang, Y. (2014). Tone management. The Accounting Review, 89(3), 1083–1113. https://doi.org/10.2308/accr-50710

Huy, T. P. (2025). Leveraging tree-based machine learning for predicting earnings management. Journal of Financial Data Science, 7(1), 45–63.

International Auditing and Assurance Standards Board. (2020). ISA 230: Audit documentation. IAASB.

International Auditing and Assurance Standards Board. (2023). The IAASB's work plan for 2024–2027: Navigating change in an AI enabled world. IAASB. Retrieved from https://www.iaasb.org/

International Sustainability Standards Board. (2023). IFRS S1 and S2: General and climate related disclosure requirements. IFRS Foundation. Retrieved from https://www.ifrs.org/

Jakesch, M., Hancock, J. T., Riedl, M. J., & Naaman, M. (2023). The social impact of generative AI on human communication. Computers in Human Behavior, 147, 107834. https://doi.org/10.1016/j.chb.2023.107834

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Ji, Z., Lee, N., Frieske, R., Yu, T., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

KPMG. (2023). Generative AI and the future of financial reporting. KPMG International. Retrieved from https://kpmg.com/

Krueger, P., Sautner, Z., & Starks, L. T. (2020). The importance of climate risks for institutional investors. Review of Financial Studies, 33(3), 1067–1111. https://doi.org/10.1093/rfs/hhz137

Li, F. (2010). The information content of forward looking statements in corporate filings—A naïve Bayesian machine learning approach. Journal of Accounting Research, 48(5), 1049–1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x

Lombardi, R., Stathopoulos, K., & Vasarhelyi, M. A. (2023). Explainable artificial intelligence in auditing: Opportunities and challenges. International Journal of Accounting Information Systems, 51, 100642. https://doi.org/10.1016/j.jaccinf.2023.100642

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Retrieved from https://arxiv.org/abs/1705.07874

Raimo, N., Vitolla, F., Rubino, M., Sorrentino, M., Mariani, M., & Viganò, R. (2023). Artificial intelligence in auditing: A systematic literature review. Journal of International Accounting, Auditing and Taxation, 52, 100556. https://doi.org/10.1016/j.intaccaudtax.2023.100556

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rozario, A. M., & Vasarhelyi, M. A. (2018). Auditing with smart contracts. International Journal of Digital Accounting Research, 18, 1–27. https://doi.org/10.4192/1577-8517-v18_1

Securities and Exchange Commission. (2022). The enhancement and standardization of climate-related disclosures for investors. SEC Release No. 33 11042. Retrieved from https://www.sec.gov/

Securities and Exchange Commission. (2024). Artificial intelligence in securities markets: Request for comment. U.S. Securities and Exchange Commission. Retrieved from https://www.sec.gov/

Sun, T., Li, X., & Liu, Y. (2024). Accountants' adoption of artificial intelligence: Evidence from Chinese listed firms. Accounting & Finance, 64(1), 123–151. https://doi.org/10.1111/acfi.12978

Sutton, S. G. (2023). The impact of artificial intelligence on the auditing profession: A research agenda. International Journal of Accounting Information Systems, 50, 100636. https://doi.org/10.1016/j.jaccinf.2023.100636

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Retrieved from https://arxiv.org/abs/1706.03762

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management Science, 46(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Weidinger, L., Mellor, J., Rauh, M., et al. (2022). Taxonomy of risks posed by language models. arXiv preprint arXiv:2112.04359. Retrieved from https://arxiv.org/abs/2112.04359.