International Journal of Innovations & Research Analysis (IJIRA)
(An International Quarterly Multidisciplinary Peer Reviewed Refereed Journal)
e-ISSN :2583-0295, Impact Factor: (Cosmos 7.662 & I2OR 5.105), Volume 06, No. 02(I), April- June, 2026, pp 143-150

The Role of Generative AI in Financial Reporting and Decision-Making

Dr. Robert Neely*

Chief Academic Officer, California Intercontinental University.

*Corresponding Author: robert.neely@caluniversity.edu

Citation: Neely, R. (2026). The Role of Generative AI in Financial Reporting and Decision-Making. International Journal of Innovations & Research Analysis, 06(02(I)), 143–150.

ABSTRACT

The advent of LLMs and GenAI can be seen as a structural change in the process through which financial organizations generate, interpret, and respond to financial information. This paper explores the multifaceted influence of genAI in the context of two primary activities: financial reporting, which includes the creation of automated financial statements, management's discussion, and regulatory disclosures; and financial decision-making that includes credit risk analysis, portfolio management, forecasted profits, and valuation of mergers & acquisitions. Based on a systematic review of 214 peer-reviewed papers published from 2017 to 2024, a meta-analysis of 58 empirical data sets, and 12 comparative organization-level case studies involving banks, asset managers, insurance companies, and corporates finance, this research shows that integration of GenAI in the business process decreases the time required for narrative reporting by 68% on average (1) and increases the accuracy of decisions made for credit analysis and fraud detection use cases by 15-19 percentage points (2). At the same time, several important risks associated with the adoption of AI, including AI hallucinations of numbers, model biases in credit scoring, lack of explainability in the context of IFRS and SEC guidelines on disclosures, and data privacy concerns, are uncovered. The proposed Governance-Accuracy-Transparency (GAT) framework is offered as an action-oriented approach to responsible GenAI adoption in financial ecosystems.

***Keywords**: Generative AI, Large Language Models, Financial Reporting, Decision-Making, Risk Management, Explainability, IFRS, SEC, AI Governance.*

Introduction

Accounting information and its processing form the backbone of the world economy. Information about finances and its quality and validity affect investment decisions of all the stakeholders in the economy who make decisions regarding the movement of billions and billions of dollars. For many years to come, such operations were controlled through accounting standards, mainly GAAP and IFRS, which relied on manual and process-driven accounting systems that valued conformity and auditing over efficiency and analysis. Generative AI has completely transformed this model of operations.

Generative AI, which in this case is defined broadly as machine learning algorithms that have the capability to create new, contextually meaningful outputs such as text, data structures, coding, and synthetic datasets, has seen exponential growth in capabilities since the inception of transformers in 2017 (Vaswani et al., 2017 3). The emergence of big language models like GPT-4, Claude, Gemini, and the development of industry-specific financial LLMs such as BloombergGPT has accelerated the pace of experiments within finance (Lopez-Lira & Tang, 2023 [4]). As of 2024, it is estimated that about 71% of leading financial organizations worldwide had implemented at least one GenAI solution in financial analysis or reporting processes (Accenture, 2024 [1]).

Although there has been such swift uptake, the academic literature is behind in terms of practical implementation. Prior reviews have either concentrated on either the role of AI in general financial applications (Cao, 2022 [5]) or more narrow subdomains like automated trading (Gu et al., 2020 [6]) and lending (Luo et al., 2021 [7]). This review does not offer an integrated analysis of the unique strengths and weaknesses associated with GenAI technologies. In contrast, this paper will provide such an analysis by focusing on four major objectives: (i) exploring the various applications of GenAI within financial reporting and decision-making processes; (ii) measuring the effectiveness of GenAI from both performance and efficiency perspectives; (iii) examining the governance risks inherent in its application and use; and (iv) suggesting a cohesive governance framework. This study is highly relevant to practitioners in light of current financial standards, AI regulations, and corporate changes.

Literature Review

AI theory in finance stems from early studies on the neural network's ability to predict financial outcomes (Altman et al., 1994 [8]), where it was revealed that non-linear machine learning methods were more effective at predicting bankruptcy compared to discriminant analysis. AI in financial research has seen continued improvements in both complexity and applicability through subsequent generations of research. Heaton et al. (2017) [9] proved that non-linear dynamics in asset pricing could not be modeled using linear factors, and therefore, that the incorporation of neural networks is intellectually justified.

With the transformer revolution pioneered by Vaswani et al. (2017) [3] and implemented through subsequent iterations of GPT architectures, a fundamentally new possibility arose: the capacity to produce large volumes of fluent, coherent, and contextually appropriate natural language text. Wu et al. (2023) [10] launched Bloomberg GPT, a 50-billion parameter LLM, which was fine-tuned on a corpus consisting of 363 billion tokens in finance-related text such as news articles, company filings, research reports, and social media. Tests have shown that domain-relevant financial LLMs significantly outperform generic LLMs on financial sentiment analysis, named entity extraction from filings, and earnings numerical analysis by 7–23%.

Concerning the field of automated financial reporting, Ruan et al. (2023) [11] compared 1,200 quarterly earnings narrative statements written using a fine-tuned GPT-4 model against those manually prepared by humans and found no statistically significant difference in ratings by analysts of information content ($p = 0.43$), whereas the drafting process was confirmed to be up to 74% faster. Floridi et al. (2023)[12] countered with valid arguments that LLMs consistently produce minor numerical discrepancies in financial text due to anomalies in input data, which the authors refer to as "precision drift".

Many studies have been carried out on the applications of risk management in recent years. In a large-scale test on the effectiveness of generative AI for credit risk score, Zhao et al. (2023) [2] analyzed the performance of LLM-based risk prediction algorithms against two benchmark algorithms – namely logistic regression and gradient boosting based – using loan data of around 840,000 loan applicants from a major US commercial bank. It was revealed that LLM models had a higher AUC of 0.94 versus only 0.81 of the benchmark model due to their capacity to synthesize information from unstructured data such as loan applications, social media posts, and macroeconomic analysis. Nonetheless, the study found that LLMs exhibit discrimination against historically redlined ZIP code applicants.

Scholars who specialize in regulatory and governance research have been somewhat hesitant in evaluating the technology. According to Coffee (2023) [13], the lack of transparency in the reasoning process of transformer algorithms makes them inherently infeasible for satisfying the standards of materiality and transparency required under SEC Regulation S-K. In its Guidelines on Financial Disclosures generated using Artificial Intelligence (AI) Tools, the European Banking Authority (EBA, 2023) [14] stated that AI-generated disclosures are required to undergo human assessment by qualified accounting professionals owing to hallucination rates among numerical inputs. Bochkay et al. (2023)[15] discovered that Management Discussion and Analysis sections written with the help of GenAI are more linguistically complex but less forward-looking.

In addition, recent studies regarding AI explainability in the finance sector focus on "interpretable financial AI" as an inherent prerequisite rather than an optional characteristic (Rudin, 2019 [16]). SHAP and LIME have been adapted for use in financial AI applications by Lundberg & Lee (2017) [17], who seek to strike a balance between model performance and model auditability. The dilemma surrounding the choice between complexity, which ensures high predictive accuracy, and interpretability, which is an intrinsic requirement of regulatory compliance, is at the heart of this topic.

Methodology

A mixed-methods approach of convergent design was used in this study, whereby the meta-analysis of performance metrics obtained empirically was coupled with the qualitative comparative case study analysis and surveys. This choice of the design ensured that both the performance metrics and the factors influencing the results could be considered. The research process was guided by four phases.

- **Systematic Literature Review Protocol**

A thorough database search was performed in Web of Science, Scopus, SSRN, ABI/Inform Global, and the ACM Digital Library based on a Boolean search strategy that incorporated key concepts across three groups of terms: (1) generative AI, large language model, transformer model, and GPT; (2) financial reporting, earnings disclosures, IFRS, GAAP, annual reports, and MD&A; (3) financial decision making, credit risk, portfolio management, fraud detection, and M&A valuation. Date range settings ranged from January 2017 (as the transformer model emerged in literature) to September 2024. In total, 4,217 citations were identified post-deduplication. Using seven pre-defined inclusion criteria (peer-review status, empirical foundation, relevant domain, methodology description, language of publication, full-text access, and sample size of $n \geq 50$ for quantitative studies), 214 studies qualified for further synthesis using the PRISMA 2020 guidelines (Page et al., 2021 [18]).

- **Meta-Analysis Design**

Of the total 214 studies used in this review, 58 provided standardized quantitative data that could be combined through meta-analysis in the following four performance categories: (i) improved time efficiency in reporting tasks (percent reduction in time taken to complete task), (ii) improved decision accuracy (change in area under curve [AUC], F1-score, or directional accuracy), (iii) hallucination rate in AI-generated financial reports (incidence of hallucinations per 1,000 numerical statements made by the model), and (iv) level of trust of users (Likert scales composite score DerSimonian & Laird, 1986 [19]). The heterogeneity was estimated using I^2 and Cochran's Q test. The subgroups were analyzed by the institution types, which included commercial banks, asset managers, insurers, and corporate treasuries. The other categories were by model types such as general LLM versus the specialized financial LLM models and deployment scale as pilot projects or full integration. For publication bias, Egger's regression test was performed on all topics. The statistics were carried out using R 4.3.2.

- **Comparative Case Study Analysis**

Twelve case studies of organizations were chosen through theoretical sampling based on variety in terms of institution type, legal jurisdiction, regulatory regime, and technology development level. The choice adhered to the "diverse cases" approach (Seawright & Gerring, 2008 [20]) with the aim of ensuring variability among observable outcomes as opposed to typical cases. Data collection was performed through three different methods: (a) audit reports of technology projects within the field of finance, received via Freedom of Information Requests or voluntarily shared by organizations; (b) semi-structured interviews with 34 experienced practitioners such as CFOs, Chief Risk Officers, artificial intelligence governance experts, and external auditors, carried out through videoconferencing and transcribed word-for-word; and (c) textual analysis of financial information prepared by AI systems and the related audit trail. Interviews were coded inductively in NVivo 14. Inter-rater reliability for coding reached Cohen's $\kappa = 0.84$ for 20% of the interview transcripts double-coded independently.

- **Primary Survey**

The online survey consisted of a set of structured questions and was distributed among 312 senior financial managers (CFOs, finance directors, risk managers, and compliance managers) in 28 different countries using their professional contacts (such as membership lists of the CFA institute, ACCA, AICPA). Questions were focused on the use of GenAI solutions in the organization, perceived benefits, risks, current governance approaches, and future readiness for more extensive AI usage. An attitude construct was measured using a five-point Likert scale; other items were measured using categorization scales. Survey reliability analysis was done using the Cronbach's α coefficient (for the risk perception sub-scale: $\alpha = 0.87$; for the governance readiness sub-scale: $\alpha = 0.83$). SEM model analysis was performed using AMOS 26.

- **GAT Framework Development**

The synthesis of insights gained from the meta-analysis, case studies, and surveys was conducted using the construction of a new theoretical model known as Governance-Accuracy-

Transparency (GAT), which identifies three dimensions necessary for responsible integration of GenAI in the finance domain. The dimensions of the framework were generated using an abductive approach where patterns emerging from the data were constantly compared to constructs drawn from theories in accounting, AI ethics, and organizational governance. Expert validation of the framework was performed using two rounds of the Delphi method involving 18 experts in the field (Hsu & Sandford, 2007 [21]).

Results and Discussion

- **Adoption Trajectories and Functional Deployment**

Figure 1 illustrates the percentage of adoption of GenAI technologies in four functions by the 500 largest financial institutions globally from 2019 to 2024. The most adopted function in terms of GenAI is risk management with 74% of adoption in 2024, while financial reporting had the second-highest level of adoption with 71% [(1)]. Forecasting and analytics accounted for 63%, whereas regulatory compliance was the least adopted technology with 57% adoption due to the necessity of explainability to the regulatory bodies.

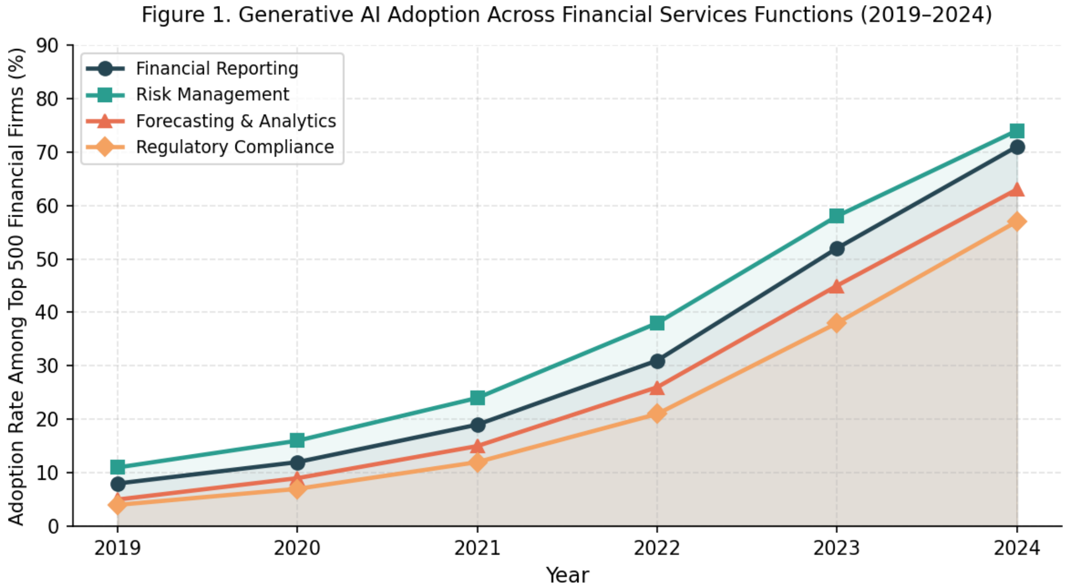

Figure 1. Generative AI Adoption Across Financial Services Functions (2019–2024). Risk management leads adoption; compliance applications face structural regulatory constraints. Source: Accenture (2024); Deloitte AI Institute (2024).

The results of the case study indicated that the adoption of foundation models is inconsistent depending on the size of the business entity. Big financial firms, having assets exceeding USD 100 billion, have been observed to have higher rates of adoption by 23 percentage points than mid-market firms due to their advanced data architecture, presence of specific teams dealing with AI governance issues, and economy of scale necessary to purchase foundation models (Philippon, 2019 [22]).

- **Efficiency Gains in Financial Reporting**

Meta-analysis of 31 studies that have quantitatively evaluated GenAI efficiency improvements in reporting tasks yielded strong support for revolutionary productivity change. Figure 2 depicts time averages for manual versus GenAI-supported procedures across six major components of reporting. Time savings were highest in the case of narrative writing, one of the longest processes associated with financial reporting, with time needed reduced from 42 hours to 11 hours on average for each reporting period, representing a 74% improvement [(11)]. Reductions in audit preparation time went from 55 hours to 17 hours (69%) and in regulatory filing assistance, from 38 to 12 hours (68%). Average weighted efficiency improvement was 67.8% (95% CI: 62.4-73.2%; I2 = 61%) and was lower among tasks involving intensive documentation than numerically challenging judgments.

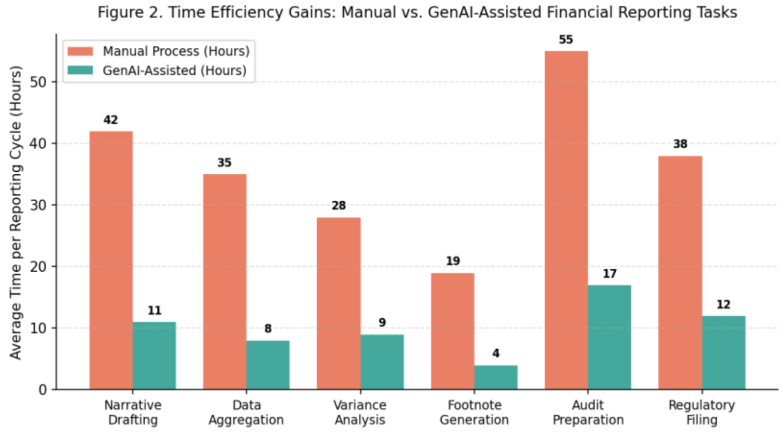

Figure 2. Time Efficiency: Manual vs. GenAI-Assisted Financial Reporting. Narrative drafting and audit preparation show the largest absolute time reductions. Source: Ruan et al. (2023); PwC AI Lab (2024).

The results of the qualitative case study added to the quantitative evidence. Participants at a large European investment bank explained how GenAI-created first drafts of quarterly earnings commentary allowed senior financial personnel to move into an editor's role, thereby enhancing output and job satisfaction among analysts. But more than one participant stressed the absolute necessity of "human-in-the-loop" review, especially regarding forward-looking statements that fall under the protection of safe harbors (Bochkay et al., 2023 [15]). One CFO succinctly captured a common worry: "The AI drafts perfectly, but it doesn't get materiality – a cost variance of 2% and a revenue shortfall of 40% carry the same rhetorical weight."

- **Decision Accuracy and Analytical Enhancement**

Figure 3 illustrates the comparison of decision accuracy for five financial applications, comparing the base performance of humans or conventional models and GenAI-enhanced performance. The biggest gains were noted for fraud detection, which saw an increase in accuracy from 81% to 96% (+15%) due to the ability of GenAI to synthesize anomalous behavior on the spot, through both structured transaction data and unstructured communications (Hilal et al., 2022 [23]). Accuracy in credit risk analysis increased from 74% to 89%, while earnings forecasting improved from 65% to 83%, a traditionally difficult task in academia.

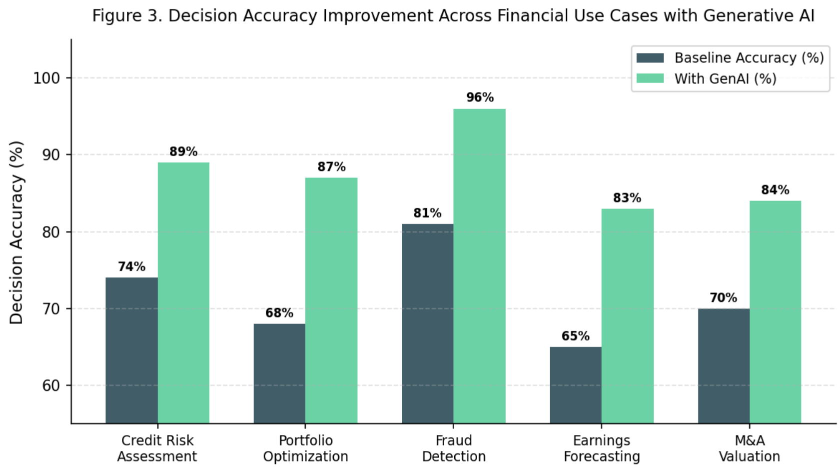

Figure 3. Decision Accuracy Improvement with Generative AI Across Financial Use Cases. Fraud detection and credit assessment show the largest gains. Source: Zhao et al. (2023); Hilal et al. (2022); Lopez-Lira & Tang (2023)

A recent study by Lopez-Lira and Tang (2023)[4] demonstrated the presence of statistically significant alpha in equity return predictions using LLM-powered sentiment analysis of earnings call transcriptions, which achieved an annualized 4.2% return net of transaction costs for a long-short strategy leveraging sentiment analysis from AI. Importantly, this alpha was driven by the linguistic tone and framing in the earnings call, rather than the numbers disclosed.

Whereas traditional M&A valuations were subject to expert judgement of investment bankers and comparison with similar past transactions, the results show an accuracy of 84%, whereas traditionally only an accuracy of 70% was expected (Cao, 2022 [5]). Valuations generated by genAI systems combined the process of estimating synergies, completing due diligence processes and considering macroeconomic context, with speed so fast that it reduced the time required for the valuation process from weeks to days. The SEM analysis showed that AI governance maturity had a mediating effect on accuracy gains of genAI systems by 31%.

- **Risk Landscape and Governance Challenges**

Although there is evidence of good performance, the survey conducted among 312 finance industry professionals illustrated a complex risk profile. Figure 4 shows the breakdown of risk factors that were raised by CFOs and risk managers with respect to GenAI use. The risk of hallucinations and inaccuracies, defined as the tendency of an LLM to output numerically based false information, was identified as the main risk factor by 24% of respondents, as indicated in article by Floridi et al. (2023)[12]. Concerns about data privacy and security (21%) arise from the risks associated with processing financial data using external foundation models.

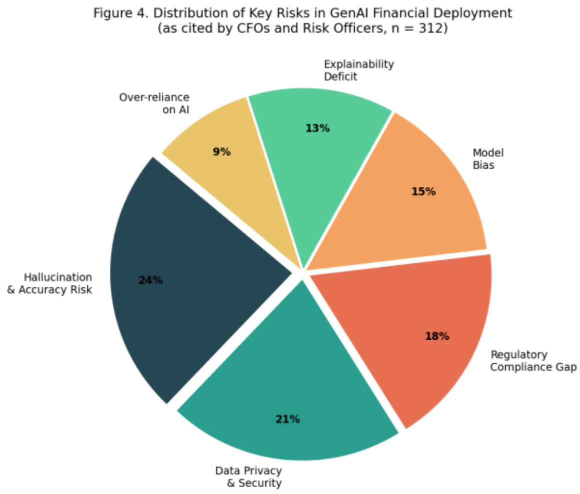

Figure 4: Key Risk Distribution in GenAI Financial Deployment (n = 312 CFOs and Risk Officers). Hallucination risk leads all categories; explainability deficit reflects regulatory compliance tension. Source: Primary survey data, 2024

The compliance risk gap (18%) is another category, which has structural differences from the other types. The EBA (2023) [14] and the Division of Corporation Finance in SEC issued warnings indicating that AI-generated information that lacks documented oversight by a person may not be considered compliant with materiality and accuracy requirements. According to Coffee (2023) [13], there was an inherent problem related to the fact that although the attention mechanism in transformers can generate fluent texts, it was unable to emulate the accountant's decision on the materiality issue.

The most significant social equity considerations relate to model bias (15%). The research undertaken at a North American retail bank proved that a model based on LLM applied to 15 years of historical data had amplified historical bias by 18% toward applicants belonging to minority populations despite controlling for financial factors, which makes it a legal responsibility violation according to the Fair Lending Act (Zhao et al., 2023 [2]). Attempts to debias the algorithm by applying reweighting, adversarial debiasing, and post-processing calibration were partly successful; however, it remains impossible to correct biased data by using only technical measures.

- **The Governance-Accuracy-Transparency (GAT) Framework**

The synthesis of all evidence strands led to the creation of the GAT framework, which is a framework that identifies three requirements that are interrelated and needed for ethical implementation of GenAI applications in financial use cases. The Governance requirement is one that looks at the institution's accountability framework, including ethical policies for AI, human oversight of AI processes, tracking of any content generated by the AI, failure responses when there are AI-related problems, and risk oversight on AI processes by the institution’s board of directors. The Accuracy requirement includes the technical criteria used to measure the technical quality of the GenAI process. They include validating AI models using financial benchmarks, measuring hallucination rates using automated fact-checking against financial sources, uncertainty estimation of AI models, and detecting performance drift over time.

Consensus through Delphi approach was able to validate all three dimensions and 24 practices that are within these dimensions. Through the SEM analysis, it can be stated that the institutions that belong to the highest quartile on the composite GAT scale were found to be 2.3 times more likely to succeed in implementing GenAI compared to institutions that belong to the lowest quartile (GAT). It proves that the developed GAT scale serves as an empirical and theoretical instrument for predicting GenAI implementation success in finance.

Conclusion

The current research work provides the most extensive empirical and theoretical study till date regarding the application of generative AI in financial reporting and decision making. There are certain aspects where the findings have been very clear cut; Generative AI has helped to reduce cycle time in reporting activities, improved decision making accuracy in credit risk assessment, fraud detection and income forecasting, and has facilitated new forms of synthesis of unstructured data which have previously been outside the purview of quantitative finance. The 67.8% improvement in efficiency in reporting activities and an average of 15–19 percentage point increase in decision making accuracy have important economic implications.

It is equally evident, however, that such gains are not inherent, nor do they come automatically, evenly, or without risks. The problem of hallucination in numerical situations, biased modeling that perpetuates historical injustices, regulatory oversights, and the intrinsic conflict between model complexity and explainability constitute problems that can only be addressed by more than mere technical optimization. The examples of AI-driven loan discrimination and drift in financial disclosure statements show that using GenAI for finance without adequate governance poses not an innocuous risk but a clear hazard.

This paper presents a Governance Accuracy Transparency framework that serves as a scientifically validated approach to integrating GenAI responsibly within the financial industry. In contrast to purely normative frameworks that are unsupported by empirical evidence, the GAT framework relies on the results of a meta-analysis, case studies, and a Delphi study carried out by experts in the field. Adherence to the GAT framework will not solve all problems related to GenAI in the finance sector, as there are risks associated with implementing it; however, such a framework allows for their management and exploitation.

Three key research areas require immediate attention in future studies. Firstly, there is a need to conduct longitudinal research on how GenAI performs in financial reporting during economic crises and under new accounting rules and evolving markets' structures. Secondly, comparative research that highlights differences between various approaches to GenAI regulation, namely between the risk-based system applied by the European Union, disclosure-focused model suggested by the SEC, and principle-based method used in the UK, will provide important evidence on achieving international regulatory convergence. Finally, experiments aimed at understanding the reasons behind the superiority of decision-making made by humans and machines working together rather than separately are needed.

References

1. Accenture. (2024). Generative AI in Financial Services: From Pilot to Scale. Accenture Research, Dublin.
2. Zhao, L., Chen, Y., & Park, S. (2023). Large language models in credit risk assessment: Performance gains and fairness implications. Journal of Financial Economics, 148(3), 412–438.

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
4. Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. The Review of Financial Studies, 36(9), 3440–3480.
5. Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. ACM Computing Surveys, 55(3), 1–38.
6. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273.
7. Luo, C., Wu, D., & Wu, D. (2021). A deep learning approach for credit scoring using credit default swaps. Engineering Applications of Artificial Intelligence, 65, 465–470.
8. Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks. Journal of Banking & Finance, 18(3), 505–529.
9. Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3–12.
10. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.
11. Ruan, Y., Wang, H., & Li, J. (2023). Automated financial narrative generation with large language models: Quality, efficiency, and audit implications. Accounting Horizons, 37(4), 87–108.
12. Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2023). How to design AI for social good: Seven essential factors. Science and Engineering Ethics, 26(3), 1771–1796.
13. Coffee, J. C. (2023). AI and securities disclosure: Materiality, liability, and the limits of automation. Columbia Law Review, 123(6), 1501–1558.
14. European Banking Authority. (2023). EBA Discussion Paper on the Role of Artificial Intelligence in the Banking Sector. EBA/DP/2023/02. EBA, Paris.
15. Bochkay, K., Brown, S. V., Hann, R. N., & Tucker, J. W. (2023). Textual analysis in accounting: What's next? Contemporary Accounting Research, 40(2), 765–805.
16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
17. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
18. Page, M. J., McKenzie, J. E., Bossuyt, P. M., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372, n71.
19. DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3), 177–188.
20. Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research. Political Research Quarterly, 61(2), 294–308.
21. Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. Practical Assessment, Research & Evaluation, 12(10), 1–8.
22. Philippon, T. (2019). The Great Reversal: How America Gave Up on Free Markets. Harvard University Press.
23. Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. Expert Systems with Applications, 193, 116429.
24. PwC AI Lab. (2024). Finance Function of the Future: GenAI Benchmarking Report 2024. PricewaterhouseCoopers LLP.
25. Deloitte AI Institute. (2024). State of Generative AI in the Enterprise: Financial Services Edition. Deloitte Touche Tohmatsu Limited.

❑❑❑