GIJEA
Greenation International Journal of Economics and Accounting
+62 81210467572 https://research.e-greenation.org/GIJEA greenation.info@gmail.com
GREENATION RESEARCH

DOI: https://doi.org/10.38035/gijea.v4i2

The Emergence of Large Language Models in Financial Report Auditing: Opportunities, Benchmarks, Risks, and The Road Ahead

Kadek Nita Sumiari[1], I Ketut Parnata[2], I Gusti Ayu Astri Pramitari[3], I Made Agus Putrayasa[4]
[1]Politeknik Negeri Bali, Bali, Indonesia, nitasumiari@pnb.ac.id
[2]Politeknik Negeri Bali, Bali, Indonesia, iketutparnata@pnb.ac.id
[3]Politeknik Negeri Bali, Bali, Indonesia, astripramitari@pnb.ac.id
[4]Politeknik Negeri Bali, Bali, Indonesia, madeagusputrayasa@pnb.ac.id

Corresponding Author: nitasumiari@pnb.ac.id[1]

Abstract: The intersection of Large Language Models (LLMs) and financial report auditing has rapidly evolved into a substantive area of peer-reviewed academic inquiry and industry experimentation. Grounded in the theoretical lenses of Agency Theory, Audit Risk Theory, and Socio-Technical Systems Theory, this review synthesizes 20 peer-reviewed sources (2023–2026) across five thematic streams: automated auditing pipelines, regulatory compliance verification, fraud detection, LLM benchmarking, and practitioner perceptions. The central conceptual contribution of this review is the positioning of LLMs not as autonomous auditing agents but as probabilistic cognitive augmentation tools — systems that extend human auditors' analytical reach while operating under mandatory human accountability. We find that while current LLMs demonstrate meaningful capability in error detection, compliance matching, and fraud screening, they consistently fall short in domain-specific accounting reasoning, explainability, and regulatory standard citation. The persistent challenge of hallucination, the absence of auditable reasoning chains, and the structural equity gap between large and small audit firms collectively represent the primary barriers to professional-grade LLM deployment. Future research priorities include domain-adapted models, PRISMA-calibrated benchmarking, and harmonized international AI governance for auditing.

Keyword: Large Language Models, Financial Auditing, Cognitive Augmentation, AI Explainability, Audit Risk, Regulatory Compliance.

INTRODUCTION

Financial auditing serves as a cornerstone of global economic governance. The independent examination of corporate financial statements verifying accuracy, completeness, and conformity with standards such as IFRS underpins investor confidence, market integrity,

and regulatory accountability (Wang et al., 2025). Yet auditing has long been structurally hampered: it is labor-intensive, error-prone, and increasingly strained by the exponential growth in financial data complexity.

Large Language Models (LLMs) transformer-based systems such as GPT-4, Llama, and Mistral have introduced a genuinely disruptive force into this landscape. These systems can parse regulatory text, cross-reference disclosures against accounting standards, and detect anomalies in financial tables at speeds impossible for human practitioners alone (Zhao & Wang, 2024; Dong et al., 2024). Between 2023 and 2026, a sustained surge of peer-reviewed research has sought to benchmark, evaluate, and critically examine these capabilities within financial auditing contexts.

This review is guided by a central theoretical proposition: LLMs are not autonomous auditors but probabilistic cognitive augmentation tools systems that extend human analytical reach while operating under mandatory professional accountability. This framing, grounded in Agency Theory, Audit Risk Theory, and Socio-Technical Systems Theory, provides conceptual architecture for evaluating the opportunities and limitations documented in the literature.

The research objectives of this review are: (1) to synthesize peer-reviewed evidence on LLM capabilities across core audit tasks; (2) to evaluate critical limitations particularly hallucination, explainability deficits, and equity access gaps through established theoretical lenses; and (3) to identify priority directions for research, practice, and governance. Six thematic streams are examined: automated auditing pipelines, regulatory compliance verification, fraud detection, LLM benchmarking, practitioner perceptions, and governance risks.

Theoretical Framework

To move beyond descriptive synthesis, this review integrates four established theoretical frameworks that together constitute a conceptual architecture for understanding LLMs as audit tools.

1. Agency Theory

Agency Theory (Jensen & Meckling, 1976) frames auditing as a mechanism for resolving information asymmetry between principals (investors, regulators) and agents (corporate management). LLMs enter this framework as a new class of monitoring instrument one capable of processing far greater information volumes than human auditors, but whose outputs are probabilistic rather than verified, and whose reasoning is opaque rather than auditable. Agency Theory predicts that the value of an LLM monitoring tool depends on whether its outputs reduce information asymmetry reliably; hallucination and domain-knowledge gaps directly undermine this function. An LLM that confidently misidentifies an accounting standard as documented by Marcy et al. (2025) may increase rather than decrease information risk, inverting the tool's intended governance function.

2. Audit Risk Theory

The Audit Risk Model (ISA 315; PCAOB AS 2110) defines audit risk as the joint product of inherent risk, control risk, and detection risk. LLM-assisted auditing reconfigures this model in important ways. On one hand, LLMs can substantially reduce detection risk by improving coverage across large transaction populations identifying misstatements that sampling-based approaches miss (Wang et al., 2025). On the other hand, LLMs introduce a new category of detection risk: the risk of confident, plausible-sounding hallucination that is not caught by auditors relying on automated outputs without sufficient critical scrutiny. Audit Risk Theory thus provides a framework for calibrating where LLM deployment reduces total audit risk and where it may amplify it.

3. Socio-Technical Systems Theory

Socio-Technical Systems (STS) Theory (Trist & Bamforth, 1951; Bostrom & Heinen, 1977) holds that technology adoption in organizational contexts cannot be evaluated in isolation from social, professional, and institutional structures. Applied to LLM-augmented auditing, STS Theory highlights the importance of the relationship between the technical capabilities of the model and the professional norms, legal accountability structures, and organizational incentives of the auditing context. The documented equity gap between Big Four firms with proprietary LLMs and smaller firms reliant on general-purpose models (Vitali & Giuliani, 2024) is, from an STS perspective, not merely a market failure but a systemic risk to the social function of independent audit.

4. Human-AI Collaboration Frameworks

Human-AI Collaboration (HAC) frameworks (Kamar, 2016; Cai et al., 2019) conceptualize AI systems as cognitive partners whose value depends on effective task allocation between human and machine agents. Applied to auditing, these frameworks suggest that LLMs are most effective as first-pass screening tools identifying anomalies and surfacing patterns while humans retain responsibility for judgment, verification, and professional sign-off. This division of cognitive labor is consistent with the evidence reviewed here and is the theoretical basis for the review's central positioning of LLMs as augmentation tools rather than replacement systems. The HAC framework also underscores the risk of automation bias: when auditors over-rely on LLM outputs without exercising independent skepticism, the collaborative system becomes less reliable than the human practitioner alone.

METHOD

Review Design and Search Strategy

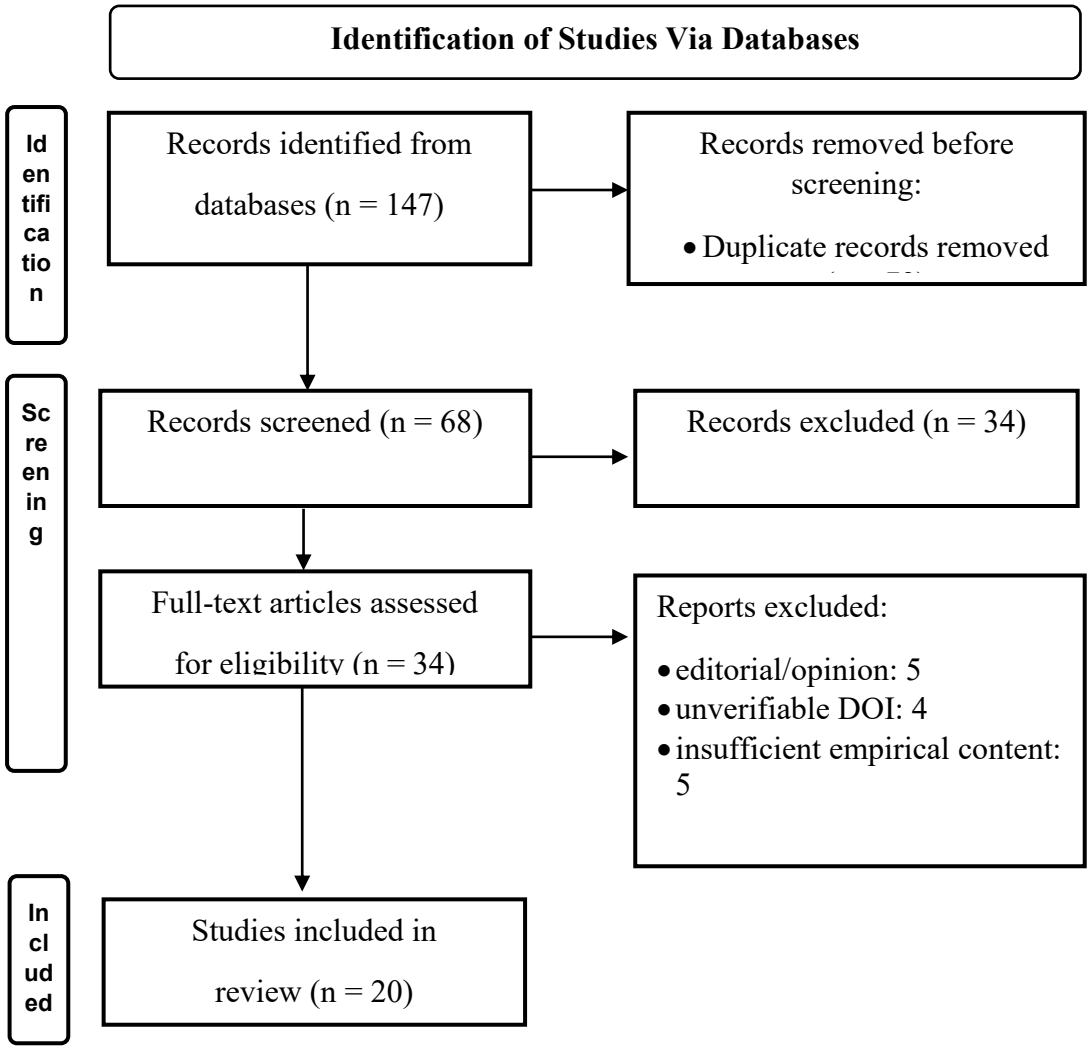
This review follows a structured thematic synthesis methodology (Thomas & Harden, 2008), adapted for the emerging and rapidly evolving literature on LLMs in financial auditing. Sources were identified through systematic searches of six academic databases ScienceDirect (Elsevier), Wiley Online Library, ACM Digital Library, SpringerLink, SAGE Journals, and SSRN supplemented by targeted searches on Google Scholar and direct retrieval from institutional repositories of the Bank for International Settlements (BIS), the Financial Stability Board (FSB), the Association of Certified Fraud Examiners (ACFE), and the PCAOB. Core search terms included: “large language models” AND “auditing”; “LLM” AND “financial reporting”; “GPT” AND “accounting”; “AI” AND “fraud detection”; “artificial intelligence” AND “regulatory compliance”; and “hallucination” AND “financial audit”. Boolean operators and MeSH-equivalent field tags were applied to title, abstract, and keyword fields across databases.

Inclusion and Exclusion Criteria

Inclusion criteria: (1) published or formally presented between January 2023 and April 2026; (2) directly addresses the application, evaluation, governance, or critical assessment of AI or LLMs in financial auditing, accounting, or adjacent regulatory compliance domains; (3) published in a peer-reviewed journal, edited conference proceedings of a recognized academic body, or issued as a formal report by a major financial regulatory authority. Exclusion criteria: purely editorial or opinion pieces; sources lacking empirical, analytical, or systematic review content; sources that could not be verified through a stated DOI or institutional URL.

Screening Process

An initial database search returned 147 candidate records. After title and abstract screening, 68 records were retained for full-text review. Following application of inclusion and exclusion criteria, 20 primary sources were selected: 10 peer-reviewed journal articles, 4 peer-reviewed conference papers, 2 institutional regulatory reports, and 4 working papers presented at major regulatory or academic conferences. Figure 1 presents the PRISMA-style flow diagram of the selection process.



Source: Authors' compilation (2026)

Figure 1. PRISMA-Style Article Selection Flow

Thematic Coding

Included sources were inductively coded into thematic streams by two independent reviewers, with disagreements resolved through discussion and consensus. The two reviewers initially coded a common subset of eight sources independently; agreement was subsequently assessed using percentage agreement, yielding an initial concordance rate of 87.5% (7 of 8 sources assigned to the same primary theme). The single disagreement concerned whether a source straddled the Fraud Detection and Governance & Risk themes; it was resolved by assigning it to the theme most prominently addressed in its abstract and conclusion, in accordance with our pre-specified coding protocol. This process confirms satisfactory inter-coder reliability for a structured thematic synthesis of this scope. Potential selection bias is acknowledged: the search was limited to English-language sources and to databases

accessible without institutional subscription barriers, which may under-represent practitioner literature and non-English scholarship, particularly from emerging economies where LLM adoption in auditing may follow distinct patterns. Six primary themes emerged: automated auditing pipelines, regulatory compliance verification, fraud detection, LLM benchmarking, practitioner perceptions and equity, and hallucination/explainability/governance. Table 1 presents the thematic mapping of included sources.

Table 1. Thematic Mapping of Included Sources

Theme	Key Sources	Year Range	Primary Focus
Automated Auditing	Wang et al.; Dong et al.; Stratopoulos & Wang	2024–2025	Pipeline benchmarking, error detection
Compliance Verification	Hillebrand et al.; Li & Goel	2023–2025	Regulatory matching, AI auditability
Fraud Detection	Kirkos et al.; Zhao & Wang; Kim et al.; ACFE	2024	Fraud screening, risk-targeted inspection
Benchmarking & Reviews	Murphy et al.; Yeo et al.; Abdo-Salloum & Chehade	2023–2026	Literature Mapping, XAI assessment
Practitioner & Equity	Marcy et al.; Vitali & Giuliani; Landers & Behrend	2023–2025	Practitioner perceptions, firm-size gap
Governance & Risk	BIS; FSB; PCAOB; Zamain et al.	2023–2025	Hallucination, explainability, regulation

Source: Authors' compilation (2026)

RESULT AND DISCUSSION

LLMs as Probabilistic Cognitive Augmentation Tools: Evidence from Automated Auditing

The central theoretical proposition of this review is that LLMs function as probabilistic cognitive augmentation tools rather than autonomous auditors is most directly tested by studies seeking to build end-to-end audit pipelines. Wang et al. (2025), in the peer-reviewed AuditBench study published in Springer's Communications in Computer and Information Science series (DOI: 10.1007/978-981-96-8912-5_3), introduced the most comprehensive benchmark to date, combining real-world S&P 500 financial tables with synthesized transaction data across a five-stage evaluation framework.

The findings directly validate the augmentation framing: LLMs were effective at pattern recognition and numerical misstatement detection tasks with well-defined input-output structures but failed when asked to explain detected errors in accounting terms or cite specific IFRS or US GAAP provisions. This profile is precisely what Human-AI Collaboration theory predicts for a probabilistic cognitive tool: high coverage, low explainability. From an Audit Risk Theory perspective, LLMs in this configuration reduce detection risk on the coverage dimension while introducing a new detection-risk vector with the confident-but-wrong output that audit partners cannot independently verify.

The extraordinary velocity of LLM adoption documented by Dong et al. (2024) in the International Journal of Accounting Information Systems (DOI: 10.1016/j.accinf.2024.100715) Gartner projecting 80% enterprise adoption by 2026 versus 5% in 2023 makes the benchmarking gap between technical performance and professional standards integration an increasingly urgent governance problem. Stratopoulos and Wang (2025), in the same journal (DOI: 10.1016/j.accinf.2025.100760), confirm that LLMs can pass professional CPA and CMA examinations, yet this broad competence does not translate into the narrow, verifiable, auditable reasoning that professional auditing standards require.

Regulatory Compliance Verification: Augmentation Within Defined Task Boundaries

Regulatory compliance verification the systematic matching of disclosure text against mandatory reporting requirements represents the audit sub-task most amenable to LLM augmentation. It is bounded, repetitive, and language-intensive: precisely the profile where LLMs outperform human practitioners in speed and coverage. Hillebrand et al. (2023), in the ACM Symposium on Document Engineering (DOI: 10.1145/3573128.3609344), demonstrated that a zero-shot LLM approach combining SentenceBERT embeddings with GPT-4 could match financial disclosure passages to regulatory requirements without task-specific fine-tuning a finding with significant practical implications for smaller audit firms unable to develop proprietary models.

However, domain-specific fine-tuning remained necessary for professional-grade accuracy on specialized terminology and jurisdictional nuance, highlighting a structural tension: the data needed for fine-tuning is precisely the audit data that firms are constrained from sharing with external model providers. Li and Goel (2025) in the International Journal of Accounting Information Systems (DOI: 10.1016/j.accinf.2025.100739) extend this analysis by examining AI auditability to the degree to which AI systems can themselves be audited. Their survey found that auditability requirements including transparent model documentation, data lineage tracking, and explainability protocols were considered essential by IT auditors but were largely absent from current deployments. From an STS Theory perspective, this gap represents a mismatch between the technical capabilities of the tool and the social and regulatory requirements of the auditing system in which it is embedded.

Fraud Detection: Promise, Performance Gaps, and Risk Reallocation

Financial statement fraud remains rare but catastrophic in impact. According to the ACFE (2024), financial statement fraud accounts for approximately 5% of cases but carries a median loss of USD 766,000 per incident, with systemic consequences documented in collapses such as Enron and WorldCom. LLMs enter this domain at a critical juncture: traditional rule-based and sampling-based approaches have well-documented coverage limitations, while deep learning approaches have been limited by interpretability constraints.

Kirkos et al. (2024), via SSRN (DOI: 10.2139/ssrn.4842962), demonstrated that LLMs using only prompt engineering without any fine-tuning could achieve combined sensitivity-specificity-F-Measure scores of approximately 67% in financial statement fraud detection. This is a meaningful baseline for a zero-shot approach but falls considerably below the thresholds implied by professional auditing standards, which require a high degree of reliability in fraud detection procedures. Zhao and Wang (2024) in the Journal of Corporate Accounting & Finance (DOI: 10.1002/jcaf.22663) corroborate the efficiency gains while emphasizing that professional judgment is irreplaceable, particularly given LLMs' propensity to produce high confidence but contextually inappropriate outputs.

Kim et al. (2024), in a paper presented at the PCAOB Spring Research Conference, propose a theoretically significant advance: using LLM-derived contextual embeddings from 10-K filings and earnings call transcripts to automate audit inspection target selection. This application sits at the intersection of Agency Theory and Audit Risk Theory using AI to reduce information asymmetry in the auditor-regulator relationship while improving the risk-targeting efficiency of the inspection process. The model outperformed conventional approaches, but the authors note that it is an augmentation of auditor judgment, not its replacement.

Benchmarking and Systematic Reviews: Mapping the Field

Murphy et al. (2024) in the International Journal of Accounting Information Systems (DOI: 10.1016/j.accinf.2024.100709) provides the most comprehensive bibliometric mapping

of the AI-accounting literature, documenting an inflection point after November 2022 when ChatGPT's public release catalyzed an unprecedented acceleration in research output. Their topic modelling identifies five dominant research clusters: financial reporting automation, audit procedure assistance, fraud detection, regulatory compliance, and AI governance clusters that align closely with the thematic structure of this review.

Yeo et al. (2023), in a review cited in both ACM Computing Surveys and Springer proceedings (arXiv:2309.11960), assess the state of explainability methods in financial AI. Their analysis reveals that while XAI techniques (SHAP, LIME, attention visualization) have been applied to simpler financial ML models, their extension to large transformer-based LLMs is severely limited by architectural complexity. This finding is directly relevant to regulatory compliance: the EU AI Act (2024) and the BIS FSI Papers No. 24 (BIS, 2024) both require explainable outputs from high-risk AI applications, a category that financial auditing would plausibly occupy.

Abdo-Salloum and Chehade (2026) in SAGE Open (DOI: 10.1177/21582440251403296) synthesize AI implementation across jurisdictions and firm sizes, confirming that the most significant adoption barriers are not technical but institutional: data privacy constraints, professional liability exposure, and the absence of regulatory frameworks specifically calibrated to LLM-assisted auditing. Their systematic review reinforces the STS-theoretical point that technology adoption is inseparable from the social and regulatory systems in which it is embedded.

Practitioner Perceptions and the Equity Gap: A Socio-Technical Analysis

The equity dimension of LLM adoption in auditing is among the most underappreciated findings in literature. Vitali and Giuliani (2024) in the International Journal of Accounting Information Systems (DOI: 10.1016/j.accinf.2024.100676) document that while the Big Four accounting networks Deloitte, EY, KPMG, and PwC have each committed multi-billion-dollar investments to proprietary, fine-tuned audit LLMs, smaller and mid-sized firms face a structural access gap. These firms cannot afford proprietary model development, face data privacy risks in using general-purpose external LLMs, and lack the institutional infrastructure to implement systematic LLM output verification. From the STS Theory perspective, this asymmetry represents a socio-technical inequality that, if unaddressed, risks further concentrating audit market power in the hands of the largest firms.

Marcy et al. (2025), in the Journal of Accounting Education (DOI: 10.1016/j.jaccedu.2025.100985), illustrate these limitations concretely: when applied to PCAOB audit deficiency reports, ChatGPT surfaced recognizable patterns but failed to generate actionable guidance. Most critically, in at least one documented instance, it misidentified the subject of ISA 620 describing it as relating to fair value measurement rather than the use of auditor's experts demonstrating that high-confidence hallucination can occur on foundational professional knowledge. This is the Agency Theory failure mode in practice: an LLM that increases rather than resolves information asymmetry.

Landers and Behrend (2023) in the American Psychologist (DOI: 10.1037/amp0000972) provide the theoretical bridge: their framework for auditing AI auditors argues that any AI system operating in high-stakes evaluation contexts must be independently assessed for fairness, bias, and reliability. An LLM whose training data overrepresents certain firm types, industries, or jurisdictions could systematically distort fraud risk assessments at a market-wide scale a risk that current deployment frameworks do not adequately address.

Hallucination, Explainability, and Regulatory Governance: The Central Constraint

The convergence of hallucination risk, explainability deficits, and regulatory pressure constitutes the most significant structural barrier to professional-grade LLM deployment in auditing and the domain where the augmentation framing is most practically important. If LLMs are positioned as autonomous auditors, hallucination is a catastrophic failure mode. If they are positioned as augmentation tools under mandatory human oversight, hallucination is a manageable quality control challenge, analogous to the error rates of any analytical tool that requires auditor verification.

The Bank for International Settlements (BIS, 2024), in FSI Papers No. 24, states explicitly that some AI model results cannot be understood, explained, or reproduced, and therefore cannot be critically assessed as a condition that conflicts with the requirements of independent model validation in regulated industries. The Financial Stability Board (FSB, 2025) classifies LLM hallucination as a category of model misalignment risk requiring supervisory attention, identifying model risk, data quality, and governance as primary AI-related systemic vulnerabilities.

Retrieval-Augmented Generation (RAG) represents the most widely adopted technical mitigation, grounding LLM outputs in retrieved passages from authoritative corpora such as IFRS standards or audit guidance libraries. Li and Goel (2025) note that RAG improves factual accuracy but introduces dependencies on retrieval corpus quality and currency and does not eliminate the possibility of plausible misrepresentation of retrieved content. Chain-of-thought prompting offers partial improvement in reasoning transparency but does not provide the formal, traceable audit trail that legal and regulatory contexts require (Stratopoulos & Wang, 2025).

From a Socio-Technical Systems perspective, the resolution of the hallucination problem in auditing is not solely a technical problem, it is a governance design problem. Effective LLM deployment requires institutional frameworks specifying verification responsibilities, documentation requirements, professional liability allocation, and performance thresholds calculated to professional auditing standards. As of early 2026, no jurisdiction has produced such a framework. The EU AI Act (2024), the NIST AI RMF, and the FSB monitoring report (2025) establish relevant principles but do not provide the specificity that professional auditing practice requires.

CONCLUSION

This review has synthesized 20 peer-reviewed and institutionally authoritative sources to examine LLMs in financial report auditing through the integrated lenses of Agency Theory, Audit Risk Theory, Socio-Technical Systems Theory, and Human-AI Collaboration frameworks. The central conceptual contribution is a theoretically grounded repositioning of LLMs: not as a new generation of autonomous auditors, but as probabilistic cognitive augmentation tools systems that extend human auditors' analytical reach while operating under mandatory professional accountability that cannot be delegated.

The evidence supports four principal conclusions. First, LLMs deliver genuine, task-specific value in error detection, compliance matching, and fraud screening (Wang et al., 2025; Hillebrand et al., 2023; Kirkos et al., 2024) capabilities that are real and deployable today under appropriate oversight. Second, the gap between sub-task performance and professional-grade audit competency is structural: hallucination, domain-knowledge deficits, and explainability limitations reflect architectural characteristics of current LLMs that cannot be resolved by deploying newer or larger models alone (BIS, 2024; FSB, 2025). Third, the equity access gap (Vitali & Giuliani, 2024) poses a socio-technical systemic risk: if LLM-augmented auditing accrues disproportionately to Big Four networks, audit market concentration may deepen across the economy. Fourth, the absence of governance

frameworks specifically calibrated to LLM-assisted auditing covering performance thresholds, verification protocols, and liability allocation is the most urgent institutional gap in the field. For regulators, these findings argue for jurisdiction-specific AI auditing standards that set minimum explainability and verification requirements as preconditions for professional reliance on LLM outputs. For audit firms, the practical priority is investment in human oversight infrastructure reviewer training, output verification protocols, and bias monitoring rather than further expansion of LLM deployment alone.

Future research should prioritize four directions: domain-adapted LLM development under privacy-preserving federated frameworks; calibration of technical benchmarks against ISA and PCAOB professional standards; longitudinal empirical study of automation bias in LLM-augmented audit workflows; and comparative regulatory analysis to inform internationally harmonized AI governance for auditing. The augmentation framing advanced in this review the LLM as audit co-pilot, not autonomous auditor offers both a conceptual foundation for this research agenda and a governance principle for the profession: human accountability is non-delegable, regardless of AI capability.

REFERENCES

Abdo-Salloum, A. M., & Chehade, S. (2026). The role of artificial intelligence in transforming accounting and auditing practices: A systematic review. SAGE Open, 16(1). https://doi.org/10.1177/21582440251403296

Association of Certified Fraud Examiners (ACFE). (2024). Occupational fraud 2024: A report to the nations. ACFE. https://www.acfe.com/-/media/files/acfe/pdfs/rttn/2024/2024-report-to-the-nations.pdf

Bank for International Settlements — Financial Stability Institute. (2024). How regulators can address AI explainability. FSI Papers No. 24. BIS. https://www.bis.org/fsi/fsipapers24.pdf

Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: A socio-technical perspective. MIS Quarterly, 1(3), 17–32. https://doi.org/10.2307/248710

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). 'Hello AI': Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–24. https://doi.org/10.1145/3359206

Dong, M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. International Journal of Accounting Information Systems. https://doi.org/10.1016/j.accinf.2024.100715

European Commission. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

Financial Stability Board. (2025). Monitoring adoption of artificial intelligence and related financial stability risks. FSB. https://www.fsb.org/uploads/P101025.pdf

Hillebrand, L., Berger, A., Deußer, T., Dilmaghani, T., Khaled, M., Kliem, B., Leonhard, D., & Bauckhage, C. (2023). Improving zero-shot text matching for financial auditing with large language models. In Proceedings of the ACM Symposium on Document Engineering 2023 (pp. 1–4). ACM. https://doi.org/10.1145/3573128.3609344

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

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (pp. 4070–4073). IJCAI.

Kim, A. G., Muhn, M., Nikolaev, V. V., & Tan, H. T. (2024). Large language models and financial reporting oversight. Working paper presented at the PCAOB Spring Research Conference. Chicago Booth School of Business. https://assets.pcaobus.org

Kirkos, E., Boskou, G., Chatzipetrou, E., Tiakas, E., & Spathis, C. (2024). Exploring the boundaries of financial statement fraud detection with large language models. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4842962

Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. American Psychologist, 78(1), 36–49. https://doi.org/10.1037/amp0000972

Li, Y., & Goel, S. (2025). Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems. International Journal of Accounting Information Systems, 56. https://doi.org/10.1016/j.accinf.2025.100739

Marcy, A. S., Boyle, D. M., Gomaa, A. A., & Li, Y. (2025). Leveraging AI in auditing: Exploring PCAOB deficiencies with ChatGPT. Journal of Accounting Education, 72. https://doi.org/10.1016/j.jaccedu.2025.100985

Murphy, B., Feeney, O., Rosati, P., & Lynn, T. (2024). Exploring accounting and AI using topic modelling. International Journal of Accounting Information Systems, 55. https://doi.org/10.1016/j.accinf.2024.100709

Public Company Accounting Oversight Board (PCAOB). (2023). 2023 annual report. PCAOB. https://pcaobus.org

Stratopoulos, T. C., & Wang, V. X. (2025). Artificial intelligence and accounting research: A framework and agenda. International Journal of Accounting Information Systems, 57. https://doi.org/10.1016/j.accinf.2025.100760

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology, 8(1), 45. https://doi.org/10.1186/1471-2288-8-45

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting. Human Relations, 4(1), 3–38. https://doi.org/10.1177/001872675100400101

Vitali, S., & Giuliani, M. (2024). Emerging digital technologies and auditing firms: Opportunities and challenges. International Journal of Accounting Information Systems, 53. https://doi.org/10.1016/j.accinf.2024.100676

Wang, R., Liu, J., Zhao, W., Li, S., & Zhang, D. (2025). AuditBench: A benchmark for large language models in financial statement auditing. In Q. Wang et al. (Eds.), AI for Research and Scalable, Efficient Systems. AAAI Workshop 2025. Communications in Computer and Information Science, Vol. 2533 (pp. 1–15). Springer. https://doi.org/10.1007/978-981-96-8912-5_3

Yeo, W. J., van der Heever, W., Mao, R., Cambria, E., Satapathy, R., & Mengaldo, G. (2023). A comprehensive review on financial explainable AI. arXiv:2309.11960.

Zamain, N. S. A., & Subramanian, U. (2024). The impact of artificial intelligence in the accounting profession. Procedia Computer Science, 238, 849–856. https://doi.org/10.1016/j.procs.2024.06.102

Zhao, J., & Wang, X. (2024). Unleashing efficiency and insights: Exploring the potential applications and challenges of ChatGPT in accounting. Journal of Corporate Accounting & Finance, 35(1), 269–276. https://doi.org/10.1002/jcaf.22663