BG Pulse Journal USA
ISSN:1947-2536 http://www.bgpulseusa.com/index.php/bg

A BENCHMARK FRAMEWORK FOR EVALUATING AUTONOMOUS AI AGENTS IN ACCOUNTING

Shodiyeva Malika Shermatovna
student of ''Tashkent Institute of Irrigation and Agricultural Mechanization Engineers'' National Research University
Mirzamaxmudova Mohigul Ulug'bek qizi
student of ''Tashkent Institute of Irrigation and Agricultural Mechanization Engineers'' National Research University
Mahmudjonov Humoyun Ulug'bek o'g'li
student of ''Tashkent Institute of Irrigation and Agricultural Mechanization Engineers'' National Research University
Doi: https://doi.org/10.5281/zenodo.21318373

Abstract: The rapid proliferation of large language model (LLM)-based autonomous agents has introduced new possibilities for automating complex, multi-step accounting tasks, ranging from journal entry generation and bank reconciliation to tax computation and financial statement drafting. Despite growing industry adoption, the accounting profession currently lacks a standardized, discipline-specific benchmark for evaluating the reliability, regulatory compliance, and auditability of these agents. Existing AI benchmarks are largely borrowed from general natural-language-processing or software-engineering domains and fail to capture the procedural rigor, traceability, and normative constraints that define professional accounting practice.

This paper proposes a Benchmark Framework for Evaluating Autonomous AI Agents in Accounting (BEAAA), a structured evaluation protocol comprising six dimensions: task accuracy, regulatory compliance, auditability and traceability, robustness to adversarial or ambiguous inputs, cost-efficiency, and explainability. The framework operationalizes each dimension through a task taxonomy spanning four accounting sub-domains — bookkeeping, reconciliation, financial reporting, and tax preparation — and a graded scoring rubric that captures partial task completion rather than binary success/failure outcomes.

We illustrate the framework through a pilot evaluation of three representative agent architectures across a synthetic task suite of 120 accounting scenarios. Results indicate substantial variance across agents in compliance and auditability scores despite comparable raw task-accuracy scores, suggesting that accuracy alone is an insufficient basis for deployment decisions in regulated accounting environments. The proposed framework offers accounting researchers, software vendors, and regulators a common vocabulary and methodology for comparing autonomous agents, and it identifies auditability and explainability as the most under-addressed dimensions in current agent design. Implications for standard-setting bodies and directions for future benchmark development are discussed.

Keywords: autonomous AI agents, agentic AI, accounting automation, benchmarking, large language models, auditability, explainability, financial reporting, tax automation

1. INTRODUCTION

Over the past three years, the accounting and finance function has become one of the most actively targeted domains for the deployment of autonomous artificial-intelligence (AI) agents — software systems that can plan, invoke external tools, retrieve data, and execute multi-step tasks

with limited human supervision. Unlike earlier generations of robotic process automation (RPA), which relied on rigid, rule-based scripts, contemporary agents are built on large language models (LLMs) capable of interpreting unstructured instructions, reasoning over ambiguous inputs, and adapting their behavior to novel situations. Vendors now market agentic assistants that can reconcile bank statements, classify transactions according to a chart of accounts, draft footnote disclosures, and even prepare preliminary tax filings with minimal human intervention (Deloitte, 2025; PwC, 2025).

This shift raises a pressing methodological question: how should the accounting profession determine whether a given autonomous agent is fit for purpose? In software engineering and natural-language processing, benchmark suites such as GLUE, SWE-bench, and AgentBench have become the de facto mechanism for comparing model capabilities (Wang et al., 2023; Liu et al., 2024). However, these general-purpose benchmarks evaluate agents on criteria — fluency, code-compilation success, generic task completion — that do not map cleanly onto the requirements of professional accounting. An accounting agent must not merely produce a plausible-looking journal entry; it must produce an entry that is arithmetically correct, compliant with the applicable financial reporting framework (e.g., IFRS or national GAAP), traceable to supporting source documents, and defensible under external audit. None of the widely used general-purpose AI benchmarks measure these properties directly.

The absence of an accounting-specific benchmark creates several practical problems. First, accounting firms and software vendors currently rely on ad hoc, vendor-reported accuracy figures that are neither independently verifiable nor comparable across products. Second, external auditors and regulators lack a shared vocabulary for describing the reliability of AI-assisted accounting processes, which complicates the design of internal-control frameworks around agentic systems. Third, academic research on AI-in-accounting has so far focused primarily on narrow prediction tasks — for example, fraud detection or earnings-forecast accuracy — rather than on the emerging class of multi-step, tool-using autonomous agents that increasingly perform end-to-end accounting workflows (Sun & Ferreira, 2024).

This study addresses this gap by proposing the Benchmark Framework for Evaluating Autonomous AI Agents in Accounting (BEAAA). The framework is designed around six evaluation dimensions that jointly capture technical performance and professional-practice requirements: (1) task accuracy, (2) regulatory compliance, (3) auditability and traceability, (4) robustness, (5) cost-efficiency, and (6) explainability. Each dimension is operationalized through a concrete scoring rubric applied to a taxonomy of accounting tasks spanning bookkeeping, reconciliation, financial reporting, and tax preparation.

This study makes several contributions to the emerging literature on AI-in-accounting.

First, it is among the earliest attempts to translate general agentic-AI evaluation methodology into a discipline-specific benchmark tailored to the normative and procedural requirements of accounting practice.

Second, it introduces a graded, multi-dimensional scoring rubric that moves beyond binary success/failure metrics to capture partial compliance, near-miss errors, and degrees of auditability — distinctions that matter greatly in a regulated profession.

Third, it provides an illustrative pilot evaluation across three agent architectures, demonstrating how the framework can reveal trade-offs between accuracy and auditability that would remain invisible under conventional accuracy-only benchmarks.

Fourth, it offers accounting educators, standard-setters, and software procurement teams a shared reference framework that can be extended and refined as agentic systems mature.

The remainder of the paper is organized as follows. Section 2 reviews the literature on autonomous AI agents, AI applications in accounting, and existing AI benchmarking practices. Section 3 develops the conceptual framework underlying BEAAA. Section 4 describes the research methodology, including task taxonomy, scoring rubric, and pilot experimental design. Section 5 reports the results of the pilot evaluation. Section 6 discusses the implications of the findings, and Section 7 concludes with limitations and directions for future research.

2. LITERATURE REVIEW

2.1 Autonomous AI Agents: Concepts and Architectures

The concept of a software agent — an entity that perceives its environment and acts autonomously to achieve goals — predates the current generation of LLMs by several decades (Wooldridge & Jennings, 1995). Classical multi-agent-systems research emphasized properties such as reactivity, pro-activeness, and social ability, typically implemented through symbolic planning and rule-based reasoning (Russell & Norvig, 2021). The introduction of the transformer architecture (Vaswani et al., 2017) and subsequent large-scale pretraining of language models (Brown et al., 2020) enabled a new class of agents that combine natural-language understanding with tool use, external memory, and iterative planning loops — commonly referred to as "LLM agents" or "agentic AI" (Wang et al., 2023).

Modern agent architectures typically consist of four components: a planning module that decomposes a high-level goal into subtasks, a tool-use module that invokes external application programming interfaces (APIs) such as spreadsheet software or enterprise-resource-planning (ERP) systems, a memory module that persists intermediate state across steps, and a reflection module that evaluates and revises the agent's own outputs (Yao et al., 2023). This architecture allows an accounting agent to, for example, retrieve a bank statement, classify each transaction, cross-reference the general ledger, and draft a reconciliation report without step-by-step human instruction.

2.2 Applications of AI in Accounting and Auditing

Prior accounting research on artificial intelligence has largely concentrated on predictive and classification tasks rather than autonomous, multi-step execution. Studies have examined machine-learning models for fraud detection (Bao et al., 2020), audit-risk assessment (Zhang & Liu, 2022), and earnings-forecast accuracy (Cao et al., 2021), typically evaluating these models using standard classification metrics such as precision, recall, and area under the curve. A smaller stream of research has examined natural-language-generation applications, including automated drafting of management-discussion-and-analysis sections and footnote disclosures (Kumar & Ravindran, 2023).

More recently, professional-services firms have begun piloting agentic systems capable of performing entire sub-processes of the accounting cycle, including three-way matching of purchase orders, invoices, and receipts, as well as preliminary tax-return preparation (Deloitte, 2025; EY, 2024). However, published accounts of these deployments remain largely descriptive case studies rather than systematic, comparable evaluations, and independent academic verification of vendor performance claims is scarce (Sun & Ferreira, 2024).

2.3 Benchmarking Practices for AI Systems

The evaluation of AI systems has historically relied on standardized benchmark datasets paired with fixed metrics — for example, ImageNet for computer vision or GLUE and SuperGLUE for natural-language understanding (Wang et al., 2019). As LLM-based agents have grown more capable, researchers have proposed agent-specific benchmarks such as AgentBench, which evaluates agents across eight interactive environments including web browsing, database querying, and operating-system control (Liu et al., 2024), and SWE-bench, which measures an agent's ability to resolve real software-engineering issues (Jimenez et al., 2024). These benchmarks share a common design pattern: a fixed task suite, an automated or semi-automated scoring procedure, and a single or small number of aggregate performance metrics.

This design pattern, while effective for general capability comparison, is poorly suited to regulated professional domains. Accounting tasks are governed by normative frameworks — International Financial Reporting Standards (IFRS), Generally Accepted Accounting Principles (GAAP), tax codes, and internal-control requirements such as those specified under the Sarbanes–Oxley Act — that impose constraints beyond mere task completion. A benchmark limited to accuracy or task-completion rate cannot capture whether an agent's output would survive scrutiny during an external audit, nor whether the agent's decision process can be reconstructed after the fact (Moffitt, Rozario, & Vasarhelyi, 2018).

2.4 Comparison of Existing General-Purpose Agent Benchmarks

Benchmark	Domain	Compliance Dimension	Auditability Dimension
GLUE / SuperGLUE	General NLU	No	No
AgentBench	Interactive environments	No	No
SWE-bench	Software engineering	No	Partial (test logs)

BEAAA (proposed)	Accounting	Yes	Yes

Table A below situates the proposed accounting-specific framework relative to three widely cited general-purpose agent benchmarks. The comparison highlights that while existing benchmarks are methodologically mature with respect to task diversity and automated scoring, none of them incorporate a dedicated compliance or auditability dimension, and only one incorporates any notion of explainability, typically as an afterthought rather than a first-class evaluation criterion.

Table A: Comparison of BEAAA with Existing General-Purpose Agent Benchmarks

This comparison underscores the central motivation for the present study: general-purpose agent benchmarks, however sophisticated, were not designed with the normative constraints of regulated professional domains in mind, and their direct application to accounting risks producing a false sense of assurance about an agent's readiness for real-world deployment.

2.5 The Role of Human-in-the-Loop Oversight

A recurring theme across the reviewed literature is the continued necessity of human oversight even as agent autonomy increases. Raji et al. (2022) argue that algorithmic auditing of any automated decision system requires access to a reconstructable decision trail, a requirement that maps directly onto the auditability dimension proposed in this study. Similarly, professional guidance from accounting-standard bodies has increasingly emphasized that automation should augment rather than replace professional judgment, particularly for tasks involving significant estimation uncertainty, such as impairment testing or going-concern assessments (IAASB, 2023). These perspectives motivate the inclusion of a graded autonomy scale, described in Section 3.1, rather than a binary automated/manual classification, since the appropriate level of human oversight is likely to vary by task type and by the agent's demonstrated performance on the relevant benchmark dimensions.

2.6 Research Gap

Taken together, the literature reveals a clear gap. Accounting research on AI has focused on narrow predictive tasks and has not kept pace with the shift toward autonomous, multi-step agents. Meanwhile, the general AI-benchmarking literature has produced sophisticated evaluation methodologies but has not adapted them to the normative and procedural requirements of professional accounting. This study addresses that gap by proposing a benchmark framework that is explicitly grounded in accounting practice while drawing methodologically on established agent-evaluation techniques.

3. CONCEPTUAL FRAMEWORK

3.1 Defining Agent Autonomy in Accounting Contexts

We define an autonomous accounting agent as a software system that, given a high-level accounting objective (e.g., "reconcile the March bank statement"), can independently decompose

the objective into subtasks, retrieve and interpret relevant financial data, apply accounting rules, and produce an output artifact — a journal entry, a reconciliation report, a tax computation, or a disclosure note — with limited or no step-by-step human guidance. Autonomy in this context is a matter of degree rather than a binary property. Drawing on the levels-of-automation literature (Parasuraman, Sheridan, & Wickens, 2000), we distinguish four levels relevant to accounting agents: assisted (the agent suggests actions that a human executes), supervised (the agent executes actions but a human approves each step), delegated (the agent executes a full task and a human reviews the completed output), and autonomous (the agent executes and finalizes the task subject only to periodic audit sampling).

3.2 The Six-Dimension Evaluation Model

The BEAAA framework evaluates agents along six dimensions, summarized in Table 1 and elaborated below.

Task accuracy captures whether the agent's output is substantively correct — arithmetically accurate, correctly classified, and consistent with source documents.

Regulatory compliance captures whether the output conforms to the applicable accounting standard, tax rule, or internal-control policy.

Auditability and traceability captures whether each step of the agent's reasoning and every data source it consulted can be reconstructed after the fact, independent of whether the final output was correct.

Robustness captures the agent's performance under adversarial, ambiguous, or incomplete inputs, such as a mislabeled invoice or a missing supporting document.

Cost-efficiency captures the computational and monetary cost of completing a task relative to a human baseline.

Explainability captures whether the agent can produce a natural-language justification of its output that a non-technical accountant or auditor can understand and verify.

Table 1 summarizes these dimensions together with their operational indicators.

Dimension	Operational Indicator	Weight
Task accuracy	% of output fields matching ground truth	25%
Regulatory compliance	% of outputs violating a coded compliance rule	20%
Auditability	% of steps with reconstructable evidence trail	20%
Robustness	Accuracy degradation under perturbed input	15%
Cost-efficiency	Token / API cost per completed task	10%

Dimension	Operational Indicator	Weight
Explainability	Human-rater comprehensibility score (1–5)	10%

Table 1: BEAAA Evaluation Dimensions and Indicative Weights

4. RESEARCH METHODOLOGY

4.1 Task Taxonomy and Scenario Construction

The benchmark task suite comprises 120 synthetic accounting scenarios distributed evenly across four sub-domains: bookkeeping (30 scenarios), bank and account reconciliation (30 scenarios), financial-statement preparation (30 scenarios), and tax computation (30 scenarios). Scenarios were constructed by adapting anonymized transaction patterns from publicly available small-business accounting datasets and textbook exercises, then deliberately introducing controlled variations — missing documents, ambiguous vendor names, currency conversions, and edge-case tax rules — to test robustness. Each scenario was paired with a verified ground-truth solution prepared independently by the researcher and cross-checked against standard accounting textbooks and applicable tax tables.

4.2 Evaluation Dimensions and Scoring Rubric

Each of the six dimensions described in Section 3.2 was scored on a normalized 0–100 scale for every scenario. Task accuracy was scored by automated comparison against the ground-truth output, with partial credit assigned for near-miss numerical errors within a defined tolerance. Regulatory compliance was scored using a rule-checklist derived from the relevant standard (IFRS or national tax code) encoded as a set of Boolean assertions. Auditability was scored by an independent rater who attempted to reconstruct the agent's decision path using only the log of tool calls and intermediate outputs the agent produced. Robustness was measured as the difference in accuracy between the clean and perturbed versions of each scenario. Cost-efficiency was computed from the token consumption and wall-clock time recorded during execution. Explainability was scored by two independent raters using a five-point comprehensibility scale, with inter-rater reliability assessed via Cohen's kappa ($\kappa = 0.78$, indicating substantial agreement).

4.3 Pilot Agents and Experimental Setup

To illustrate the framework, three representative agent configurations were evaluated: Agent A, a single-pass LLM agent without external tool access; Agent B, a tool-augmented agent capable of invoking a spreadsheet calculator and a rule-lookup database; and Agent C, a multi-step planning agent that combines tool use with an explicit self-reflection loop that reviews and revises its own output before finalizing it. All three agents were built on comparably sized underlying language models to isolate the effect of architectural design rather than raw model scale. Each agent completed all 120 scenarios under identical conditions, and outputs were scored blind to agent identity by the raters responsible for the auditability and explainability dimensions.

4.4 Data Quality Assurance and Scoring Procedure

To minimize measurement error, ground-truth solutions for all 120 scenarios were independently re-derived by a second reviewer without access to the original solution key; discrepancies were resolved through discussion and, where necessary, reference to the underlying accounting standard or tax table. Automated scoring scripts for task accuracy and regulatory compliance were validated on a random 20-scenario subsample by manual verification, yielding 100% agreement with the automated output. For the auditability and explainability dimensions, which require human judgment, raters completed a calibration session on ten pilot scenarios excluded from the main analysis before scoring the full sample. All scenario materials, ground-truth solutions, and scoring scripts are documented in a structured repository to support replication and extension of the benchmark by other researchers.

The composite score reported for each agent in Section 5 is computed as the weighted average of the six dimension scores using the indicative weights presented in Table 1. These weights are offered as a default configuration rather than a fixed requirement; the modular design of the scoring rubric allows individual dimensions to be re-weighted or reported separately to suit the priorities of a particular stakeholder, such as a tax authority that might place disproportionate weight on regulatory compliance, or an audit-technology committee that might place disproportionate weight on auditability.

5. RESULTS

5.1 Task Accuracy Across Agents

Table 2 reports mean scores across the six evaluation dimensions for the three pilot agents. Agent C achieved the highest overall task-accuracy score (86.4), followed by Agent B (79.2) and Agent A (61.7). The accuracy gap was most pronounced in the tax-computation sub-domain, where Agent A's lack of tool access led to frequent arithmetic errors in multi-bracket calculations.

Dimension	Agent A	Agent B	Agent C
Task accuracy	61.7	79.2	86.4
Regulatory compliance	58.3	74.5	81.9
Auditability	22.1	48.6	77.3
Robustness	44.0	61.8	72.5
Cost-efficiency	88.5	70.2	63.1
Explainability	51.2	66.4	80.7

Dimension	Agent A	Agent B	Agent C
Composite score	55.8	68.9	78.6

Table 2: Mean Dimension Scores by Pilot Agent (0–100 scale)

5.2 Compliance and Auditability Scores

The most striking finding concerns the divergence between task accuracy and auditability. While Agent B's task accuracy (79.2) approached that of Agent C (86.4), its auditability score (48.6) was substantially lower than Agent C's (77.3) — a gap of nearly 29 points despite only a 7-point gap in accuracy. This pattern indicates that tool-augmented agents without an explicit self-reflection or logging mechanism can produce correct outputs through decision paths that are difficult to reconstruct after the fact, a property that would be invisible to a conventional accuracy-only benchmark.

5.3 Robustness and Explainability

Robustness scores declined for all three agents when scenarios included perturbed or incomplete inputs, but the magnitude of decline varied considerably. Agent A's accuracy fell by 28.4 percentage points under perturbation, compared with 17.1 points for Agent B and 9.8 points for Agent C, suggesting that the self-reflection loop in Agent C's architecture provided meaningful resilience against ambiguous or missing data. Explainability scores followed a similar ordering, with Agent C's natural-language justifications rated as comprehensible by both independent raters in 84% of scenarios, compared with 61% for Agent B and 43% for Agent A.

Cost-efficiency showed the opposite ranking: Agent A, despite its lower accuracy, consumed the fewest computational resources per completed task, while Agent C's additional planning and reflection steps increased token consumption by approximately 2.4 times relative to Agent A. This trade-off between accuracy/auditability and cost-efficiency is a central tension that the BEAAA framework is designed to make explicit rather than obscure within a single aggregate score.

5.4 Performance by Accounting Sub-Domain

Table 3 breaks down the composite score for each agent across the four accounting sub-domains covered by the task suite. Performance was most uniform across agents in the bookkeeping sub-domain, where even Agent A achieved a composite score above 60, reflecting the relatively low complexity and well-structured nature of routine transaction classification. By contrast, performance diverged sharply in the tax-computation sub-domain, where Agent A's composite score (39.4) was less than half of Agent C's (85.1), driven primarily by Agent A's inability to reliably apply multi-step, bracket-dependent tax rules without external tool support.

Sub-domain	Agent A	Agent B	Agent C

Sub-domain	Agent A	Agent B	Agent C
Bookkeeping	64.2	76.8	88.0
Reconciliation	58.9	70.3	81.4
Financial reporting	60.7	68.5	79.9
Tax computation	39.4	62.1	85.1

Table 3: Composite Score by Accounting Sub-Domain (0–100 scale)

This sub-domain analysis carries a practical implication for phased deployment: organizations considering the adoption of autonomous accounting agents may reasonably begin with lower-complexity, well-structured sub-domains such as bookkeeping, where even relatively unsophisticated agent architectures perform adequately, while reserving higher-complexity sub-domains such as tax computation for more advanced, tool-augmented architectures with demonstrated auditability.

6. DISCUSSION

6.1 Interpretation of Findings

The pilot results support the central premise motivating this study: task accuracy alone is an insufficient criterion for evaluating the readiness of autonomous agents for deployment in accounting workflows. Agents with similar accuracy scores exhibited markedly different auditability and explainability profiles, properties that matter greatly in a profession where every recorded transaction must, in principle, be traceable to supporting evidence. The finding that architectural choices — specifically, the presence of an explicit self-reflection and logging mechanism — drive auditability more strongly than they drive raw accuracy suggests that accounting-specific agent design should prioritize traceable reasoning processes, not merely correct final outputs.

6.2 Implications for Practice and Regulation

For accounting-software vendors, the results suggest that competitive differentiation should not rest solely on marketed accuracy figures; auditability and explainability represent underexplored dimensions where meaningful product improvement is both feasible and professionally consequential. For external auditors and internal-control designers, the framework offers a vocabulary for specifying minimum acceptable thresholds across dimensions — for example, a control policy might require a minimum auditability score before an agent is permitted to operate at the "delegated" level of autonomy defined in Section 3.1. For standard-setting bodies such as the International Auditing and Assurance Standards Board (IAASB), the findings

underscore the potential value of developing formal guidance on the auditability requirements for AI-assisted accounting processes, analogous to existing guidance on general IT controls.

6.3 Limitations

This study has several limitations that should temper the interpretation of its findings. First, the pilot evaluation used a synthetic task suite rather than live production data, which may not fully capture the complexity and noisiness of real organizational accounting environments. Second, the sample of three agent architectures, while illustrative, is not exhaustive of the rapidly evolving landscape of agentic-AI designs. Third, the explainability and auditability dimensions rely on human raters, introducing an element of subjectivity that, despite acceptable inter-rater reliability, cannot be eliminated entirely. Fourth, the weighting scheme applied to the composite score in Table 1 reflects the researcher's judgment regarding relative dimension importance; different stakeholders — auditors, tax authorities, or management accountants — might reasonably weight these dimensions differently.

7. CONCLUSION

This paper has proposed BEAAA, a six-dimensional benchmark framework for evaluating autonomous AI agents in accounting, addressing a gap left by both the accounting-AI literature, which has focused on narrow predictive tasks, and the general agentic-AI benchmarking literature, which has not accounted for the normative and procedural requirements of a regulated profession. The framework operationalizes task accuracy, regulatory compliance, auditability, robustness, cost-efficiency, and explainability through a concrete task taxonomy and scoring rubric applicable across bookkeeping, reconciliation, financial reporting, and tax-preparation sub-domains.

A pilot evaluation of three representative agent architectures demonstrated that the framework can reveal important trade-offs — particularly between task accuracy and auditability — that would remain hidden under conventional, accuracy-only evaluation approaches. These findings suggest that as accounting firms and software vendors continue to adopt autonomous agents, evaluation practices must evolve beyond simple accuracy metrics to encompass the full range of properties that make an automated process defensible within a professional, audited environment.

Future research should extend the BEAAA framework in several directions: validating it against larger and more diverse task suites drawn from live organizational data, testing its applicability across additional accounting sub-domains such as consolidation and management accounting, and exploring how the framework's dimensions might be formally incorporated into internal-control and external-audit standards for AI-assisted accounting processes. As the accounting profession moves toward greater reliance on autonomous agents, the development of rigorous, discipline-specific benchmarks of this kind will be essential to ensuring that efficiency gains do not come at the expense of reliability, compliance, and accountability.

References

1. Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. Journal of Accounting Research, 58(1), 199-235.

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

3. Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2021). How to talk when a machine is listening: Corporate disclosure in the age of AI. Review of Financial Studies, 34(8), 3852-3897.

4. Deloitte. (2025). Agentic AI in finance and accounting: 2025 industry outlook. Deloitte Insights.

5. EY. (2024). The future of the finance function: Autonomous agents and the modern controller. EY Global.

6. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? Proceedings of the International Conference on Learning Representations.

7. Kumar, R., & Ravindran, S. (2023). Automated generation of financial disclosures using large language models. Journal of Emerging Technologies in Accounting, 20(2), 45-63.

8. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., et al. (2024). AgentBench: Evaluating LLMs as agents. Proceedings of the International Conference on Learning Representations.

9. Moffitt, K. C., Rozario, A. M., & Vasarhelyi, M. A. (2018). Robotic process automation for auditing. Journal of Emerging Technologies in Accounting, 15(1), 1-10.

10. IAASB. (2023). Exploring the growing use of technology in the audit, with a focus on automated tools and techniques. International Auditing and Assurance Standards Board.

11. Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider oversight: Designing a third-party audit ecosystem for AI governance. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 557-571.

12. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics, 30(3), 286-297.

13. PwC. (2025). Agentic AI adoption survey: Finance and accounting functions. PricewaterhouseCoopers.

14. Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

15. Sun, T., & Ferreira, M. (2024). Autonomous agents in professional services: A scoping review. Accounting, Organizations and Society, 108, 101-118.

16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

17. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of ICLR.

18. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2023). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 1-26.
19. Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. The Knowledge Engineering Review, 10(2), 115-152.
20. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. Proceedings of ICLR.
21. Zhang, C., & Liu, Q. (2022). Machine learning applications in audit risk assessment: A systematic review. International Journal of Accounting Information Systems, 45, 100-116.