From Sampling to Continuous Assurance: A Practical Framework for Using AI in Audit Transaction Testing

Santosh Vasudevan, Lead Data Scientist, Caterpillar, Chicago, IL, USA
David Wang, DePaul University, Chicago, IL, USA

Executive Summary

Audit transaction testing has traditionally relied on sampling. Auditors review a subset of customer statements and compare key fields against the system of record. While this approach is well established, it leaves most transactions untested and becomes increasingly difficult to scale as organizations generate millions of customer documents each reporting cycle. Advances in AI-driven document intelligence now make full-population testing practical. Customer-facing documents can be converted into structured data and reconciled against authoritative records at scale, allowing auditors to focus on exceptions rather than routine verification. This article introduces the Risk-Based Population-Level Assurance Framework (RPLAF), a seven-stage, technology-agnostic approach for integrating AI-assisted evidence collection into audit planning, testing, exception management and continuous assurance under a confidence-aware, human-in-the-loop governance model. In a credit-card statement testing implementation, the framework expanded coverage from approximately 0.5% of the population to full-population testing, identified six confirmed discrepancies and reduced recurring testing costs by more than 94%. The framework is guided by a simple principle: AI gathers the evidence; auditors make the decisions.

Why Sampling Is Becoming a Limitation

Audit transaction testing answers a fundamental question: Do customer-facing documents accurately reflect what is recorded in the system of record? In industries such as banking, insurance and utilities, that question extends beyond operational accuracy to regulatory compliance. Account statements, invoices, loan disclosures and confirmations are issued in enormous volumes, and organizations are expected to ensure that these communications are accurate, complete and consistent with authoritative records. The Sarbanes-Oxley Act requires effective internal control over financial reporting, while consumer-protection regulations require customer disclosures to be accurate and not misleading.[19] When these controls fail, organizations may face regulatory findings, penalties and costly remediation efforts.[20]

For decades, transaction-testing controls have relied on sampling. Auditors select a subset of customer documents, verify key fields and reconcile them to the underlying system of record. At smaller scales, this approach can provide reasonable assurance. At enterprise scale, however, it faces two fundamental limitations.

Limited coverage. Sampling examines only a small fraction of the population. A control that reviews 500 statements from a population of approximately 100,000 tests only about 0.5% of the documents issued during the period. Errors that exist in the remaining 99.5% of the population may never be examined. This residual sampling risk is inherent in the methodology rather than the result of insufficient effort.

Limited scalability. Manual review requires auditors to read and validate each document individually. As document volumes increase, audit effort grows almost proportionally.[6, 7] Because full manual review is often impractical, organizations accept limited coverage and the risks associated with sampling.

The result is a control that is simultaneously incomplete and resource intensive. As document volumes continue to grow and regulatory expectations increase, the limitations of sample-based testing become more pronounced.

What AI Changes

The limitation is not the auditor's expertise; it is the nature of the evidence being tested. The source transactions reside in structured systems, while the documents customers receive are typically unstructured PDFs. Traditional audit analytics can validate data within structured systems, but they often cannot verify whether the information presented to customers matches what was intended. Yet that verification is the very purpose of the control.[6, 12]

The traditional process consists of three steps: selecting a sample, manually reviewing documents and reconciling the results against the system of record. Each step is constrained by the time and effort required for human review.

AI-enabled document intelligence changes the economics of document testing. Rather than requiring auditors to read each statement individually, AI can process an entire population of documents, extract key fields such as minimum payment amount, due date and statement balance, and convert them into structured data. [13, 14, 15, 16] That information can then be reconciled against authoritative records across the full population rather than a limited sample. The auditor receives a prioritized list of exceptions for review instead of hundreds or thousands of documents requiring manual inspection. Commercial document-intelligence platforms increasingly provide this capability at enterprise scale.[18]

From an audit perspective, three important changes occur:

Coverage. Testing can expand from a small sample to the full population, significantly reducing sampling risk for the fields examined.

Effort. Auditor time shifts from routine document review to evaluating exceptions, investigating discrepancies and exercising professional judgment.

Timing. Testing can move from periodic, retrospective reviews toward near-real-time monitoring as documents are generated.

Importantly, the technology does not function as a black box. Extracted values can be linked back to the originating document and accompanied by confidence scores indicating the model's level of certainty. Auditors therefore review evidence rather than simply accept automated conclusions.[9, 17] This transparency is essential in assurance environments, where conclusions must be supported, explainable and auditable.

The challenge, therefore, is not whether organizations can apply AI to transaction testing, but how to do so in a way that remains auditable, explainable and governed.

The Risk-Based Population-Level Assurance Framework (RPLAF)

The RPLAF converts AI-enabled document intelligence into a repeatable audit process.[5] The framework is technology-agnostic and expressed in audit rather than vendor-specific terms. It positions AI as an evidence-collection and prioritization tool while preserving human responsibility for investigation,

escalation and audit conclusions. Seven stages guide the process from risk assessment through continuous assurance. The framework consists of seven stages, each designed to strengthen assurance quality while maintaining human accountability.

Stage 1: Risk Identification

Purpose: Decide what to assure and why.

Key activities: Identify the processes whose customer communications carry the greatest regulatory, financial or reputational exposure; define the control objectives (typically that each field on the statement matches the authoritative record); and set coverage, frequency and tolerance.

Outputs: a risk inventory and a control mapping that links each assertion to the field-level test that will evidence it.

Governance: Traditional audit planning remains the foundation of the framework. Risk identification ensures that all downstream testing remains aligned with business and regulatory priorities.[21, 22]

Stage 2: Data Readiness Assessment

Purpose: Confirm that required evidence sources are complete, accessible and reliable before testing begins.

Key activities: Inventory the evidence sources, separating structured systems of record from the unstructured customer documents; assess quality and accessibility; and identify what extraction is needed to make the documents analyzable.

Outputs: a data inventory and an evidence catalog naming, for each assertion, the authoritative source and the extraction task.

Governance: a candid readiness assessment prevents confident conclusions being drawn over incomplete or unreliable evidence.

Stage 3: AI-Enabled Evidence Collection

Purpose: Convert unstructured documents into structured, testable data.

Key activities: Train an extraction model using a small representative set of annotated documents. Subject-matter experts define target fields, extraction logic and expected values. The model is then applied across the population, with extracted values, confidence scores and source-document references stored for audit review.

Outputs: a structured evidence repository in which every value is traceable to its origin and carries an explicit measure of model certainty.

Governance: two properties matter: a small training corpus lowers the barrier to adoption, and persisting confidence with every value is what makes risk-based triage and confidence-aware routing possible later.

Stage 4: Population-Level Testing

Purpose: Test the entire population instead of a sample.

Key activities: Reconcile every extracted value against the system of record under a consistent normalization policy; confirm matches; and flag mismatches, missing values and unreadable extractions as candidate exceptions.

Outputs: a complete ledger of confirmations and candidate exceptions, each tied to its source document and confidence score.

Governance: The shift is more than a matter of scale. Full-population testing removes the blind spots inherent in sampling and replaces them with measurable detection risk that can be evaluated and improved over time. Under sampling, an error in an untested item is a permanent blind spot. Under full-population testing that blind spot disappears for the fields tested; what remains is detection risk: the chance the extraction or reconciliation logic misses a true discrepancy, which, unlike sampling risk, can be measured against ground truth and reduced over time.

Stage 5: Risk-Based Exception Prioritization

Purpose: Make total coverage manageable.

Key activities: Rank candidate exceptions using three signals: materiality (financial or regulatory significance), likelihood (how probable a true discrepancy is) and model confidence (how certain the extraction was), and surface the ranked queue through a dashboard that lets reviewers filter, sort and drill into individual statements.

Outputs: a prioritized exception list that turns full coverage into a bounded, triage-ordered workload.

Governance: Low confidence is treated as a risk signal, directing reviewer attention to items most likely to require investigation.

Stage 6: Human Investigation and Escalation

Purpose: Preserve auditor judgment and accountability.

Key activities: The AI gathers evidence, reconciles it, scores confidence and ranks exceptions; the auditor interprets, investigates root cause, escalates material findings and signs off. A confidence-aware routing gate sends low-confidence or high-materiality items to mandatory human review, while high-confidence routine matches may be processed automatically and spot-checked.

Outputs: validated findings, documented dispositions and sign-off.

Governance: This stage defines the framework's central governance principle: AI supports the audit process, but auditors remain responsible for interpretation, escalation and conclusions.[9, 10, 11]

Stage 7: Continuous Assurance Feedback Loop

Purpose: Turn a one-time test into a standing control.

Key activities: Feed confirmed exceptions back as control-improvement signals; adjust thresholds and rules based on validated exceptions and reviewer feedback; and use confidence trends to decide when retraining or prompt refinement is warranted.

Outputs: refined thresholds, improved controls and updated risk models for the next cycle.

Governance: because new documents can be tested as they are generated, assurance moves from a quarterly event to near-real-time monitoring, and the same loop that improves controls also maintains the model.[1, 2, 3, 4, 8]

Case Study Results

The framework was applied to a customer-statement accuracy control that verified whether minimum payment amounts, payment due dates and statement balances matched the authoritative system of record. The implementation analyzed 500 statements representing 500 unique customers from a single issuer. Although conducted in a controlled environment with a largely standardized statement format, the testing scenario is representative of issuer-level control testing commonly performed within financial-services audit functions.[23]

Coverage: The implementation tested 100% of the population. The control it replaced had been run on a quarterly sample of 500 statements drawn from a population of roughly 100,000, about 0.5%, leaving 99.5% of the population untested. Expanding coverage from approximately 0.5% of the population to full-population testing substantially reduces the limitations inherent in sample-based testing and provides assurance over every tested item.

Exception detection: Reconciliation surfaced six candidate exceptions: two in minimum payment, two in payment due date and two in statement balance. Auditor review confirmed all six as genuine discrepancies between the customer-facing statement and the authoritative record. No false positives were observed during validation. These are exactly the kind of divergence a 0.5% sample would almost certainly have missed.

Confidence behavior: The model reported an overall average confidence of 0.781, varying by field: roughly 0.89 for minimum payment, 0.78 for statement balance and 0.68 for payment due date. The lower figure on dates reflects their greater layout variability. Manual validation against ground truth showed near-perfect agreement. Interestingly, observed accuracy exceeded the model's reported confidence levels. This conservative behavior can be beneficial in audit settings because borderline items are more likely to be routed for human review rather than accepted automatically.

Economics: The recurring cost reduction exceeds 94% versus manual testing, with payback inside the first month and cumulative three-year savings above US$350,000, while coverage rises from 0.5% to 100%. The implementation improved both assurance coverage and operating efficiency simultaneously. Table 1 contrasts the two approaches across the audit dimensions that matter, and Table 2 summarizes the three-year economics.

Audit Dimension	Manual Approach	AI-enabled Approach
Documents Reviewed	Sample-based (~0.5%)	Full population (100%)
Confidence Reporting	Not available	Model-generated (avg. 0.781)
Scalability	Limited by human effort	Scales with compute
Review Latency	Periodic (batch audits)	Near real time
Sampling Risk	Present	Eliminated
Audit Effort	High manual effort	Focused exception review
Explainability	Human judgment	Field-level values + confidence

Table 1. Manual transaction testing versus AI-enabled population-level testing across key audit dimensions.

Year	Manual cost	Automated cost	Cumulative savings
Year 0	$0	$9,000	($9,000)
Year 1	$127,500	$5,000	$113,500

Year 2	$127,500	$5,000	$236,000
Year 3	$127,500	$5,000	$358,500

Table 2. Three-year cost comparison and cumulative savings (assumptions per the enterprise implementation).

For audit leaders, the key takeaway is that broader coverage, faster detection and lower cost can be achieved simultaneously while auditor time shifts from routine verification to higher-value investigative work.

Governance Lessons Learned

Population-level automation does not eliminate the auditor's role; it changes it. That transition succeeds only when governance is deliberate, documented and consistently applied. One principle underpins the framework: **AI gathers the evidence; auditors make the decisions**.

Human-in-the-loop oversight. Accountability for the audit conclusion stays human. The machine gathers, reconciles, scores and ranks; the auditor interprets, investigates, escalates and signs off. None of those last four can be delegated to a model.

Confidence-based routing. Automated disposition is conditional, never blanket. Low-confidence or high-materiality items always reach a person; only high-confidence routine confirmations are processed automatically, and even those are spot-checked for quality control.

Accountability and ownership. The division of labor can be expressed as a formal accountability matrix across four roles: the AI system, the auditor, the audit manager and the model-risk or governance function (Table 3). The defining governance principle is that AI systems may perform tasks, but they are never accountable for outcomes. Accountability always rests with a designated human owner. The RACI matrix shown in Table 3 formalizes this division of responsibility across the AI system, auditor, audit manager and governance function.

Activity	AI system	Auditor	Audit manager	Model risk / governance
Model training	R	C	I	A
Evidence extraction	R	I	I	C
Reconciliation to source of truth	R	C	I	I
Confidence scoring	R	I	I	C
Exception prioritization	R	C	I	I
Exception investigation	–	R/A	C	I
Escalation of material findings	–	R	A	I
Sign-off on conclusions	–	R	A	I
Threshold management	C	C	A	R

Model monitoring	–	I	C	R/A

Table 3. RPLAF accountability matrix. R = Responsible, A = Accountable, C = Consulted, I = Informed, – = not involved.

Model governance. AI introduces model risk, which must be managed, monitored and documented rather than assumed away. Organizations control training data, model versions and access; monitor for drift as templates evolve; and set thresholds that trigger revalidation.

Prompt governance. Where extraction is driven by natural-language prompts, an uncontrolled prompt edit can change behavior with no code change. Prompts are therefore treated as controlled audit artifacts, versioned and change-approved on the same footing as the model.

Hallucination control. Generative components can produce plausible but unsupported values. The framework addresses this risk structurally. Every extracted value is reconciled against an authoritative source, causing unsupported outputs to surface as exceptions rather than being accepted as evidence. Low-confidence extractions are routed for human review.

Documentation, evidence retention and auditability. Training data, prompts, model versions, thresholds and evaluation results must be documented, version-controlled and reproducible. Reproducibility, the ability to re-derive a past conclusion from preserved inputs, is what converts an AI output into audit evidence on which reliance can be placed.

Change management. Moving from sampling to population-level assurance changes roles, skills and workflows. Auditor effort shifts from document review to investigation and analysis, requiring deliberate change management and training.

Regulatory expectations. As regulators increase scrutiny of AI used in regulated processes, organizations must demonstrate that AI-assisted controls are governed, explainable and auditable. Evidence records, confidence scores, dispositions and accountability assignments help establish that assurance.

Across every one of these controls, the same refrain holds: the AI gathers the evidence, and the auditor makes the decision.

Implementation Checklist for Audit Leaders

Before relying on AI-generated audit evidence, audit leaders should confirm that the following are in place:

- ☐ A defined authoritative source of truth for every field to be tested
- ☐ Documented control objectives mapped to field-level tests (Stage 1)
- ☐ A completed data readiness assessment covering quality, completeness and accessibility (Stage 2)
- ☐ Established governance ownership, including a named model-risk or governance function
- ☐ A documented RACI accountability model in which no AI activity is “Accountable”
- ☐ Defined confidence thresholds for routing items to mandatory human review
- ☐ Defined auditor sign-off requirements for conclusions and material findings
- ☐ Established escalation procedures for material exceptions
- ☐ A model validation process measured against auditor-confirmed ground truth

☐ Implemented model and confidence monitoring, including drift detection

☐ Established retraining and prompt-refinement triggers

☐ Version control and change approval for prompts and model versions

☐ Defined evidence-retention standards (extracted values, confidence scores, source links and dispositions)

☐ Reproducibility controls enabling any past conclusion to be re-derived from preserved inputs

☐ A change-management plan for redirecting auditor effort from review to investigation

Key Takeaways

Governance is the enabler, not the afterthought. Full-population testing is only as trustworthy as the controls around the model. Build the accountability, routing and documentation before relying on the output.

Assurance quality improves structurally. Replacing sample-based testing with full-population testing reduces blind spots and shifts the remaining uncertainty toward measurable and continually improvable detection risk.

Scale and cost move together. The case study delivered 100% coverage and a recurring cost reduction exceeding 94% at the same time: coverage and economics are not a trade-off.

Auditor judgment is preserved by design. Interpretation, escalation and sign-off remain human. The AI gathers the evidence; the auditor makes the decision.

Confidence is a triage tool, not a verdict. Use confidence scores to direct attention, and treat conservative scores as a feature that protects the conclusion.

Start with the source of truth and the accountability model. These two foundations determine whether everything downstream is defensible.

Treat it as a standing control. The greatest value comes from continuous monitoring and a feedback loop that improves both the controls and the model over time.

Endnotes

1. M. A. Vasarhelyi and F. B. Halper, "The Continuous Audit of Online Systems," Auditing: A Journal of Practice & Theory, vol. 10, no. 1, 1991, pp. 110–125.

2. M. G. Alles, A. Kogan and M. A. Vasarhelyi, "Putting Continuous Auditing Theory into Practice: Lessons from Two Pilot Implementations," Journal of Information Systems, vol. 22, no. 2, 2008, pp. 195–214.

3. M. G. Alles, A. Kogan and M. A. Vasarhelyi, "Continuous Monitoring of Business Process Controls: A Pilot Implementation of a Continuous Auditing System at Siemens," International Journal of Accounting Information Systems, vol. 7, no. 2, 2006, pp. 137–161.

4. M. A. Vasarhelyi, M. G. Alles and A. Kogan, "Principles of Analytic Monitoring for Continuous Assurance," Journal of Emerging Technologies in Accounting, vol. 1, 2004, pp. 1–21.

5. A. R. Hevner, S. T. March, J. Park and S. Ram, "Design Science in Information Systems Research," MIS Quarterly, vol. 28, no. 1, 2004, pp. 75–105.

6. F. Huang, W. G. No, M. A. Vasarhelyi and Z. Yan, "Audit Data Analytics, Machine Learning, and Full Population Testing," Journal of Finance and Data Science, vol. 8, 2022, pp. 138–144.

7. Y. Chen, Y. Liu and J. Zhang, "A Full Population Auditing Method Based on Machine Learning," Sustainability, vol. 14, no. 24, art. 17008, 2022.

8. P. L. Joshi and G. Marthandan, "Continuous Internal Auditing: Can Big Data Analytics Help?" International Journal of Accounting, Auditing and Performance Evaluation, vol. 16, no. 1, 2020, pp. 25–42.

9. J. Kokina and T. H. Davenport, "The Emergence of Artificial Intelligence: How Automation Is Changing Auditing," Journal of Emerging Technologies in Accounting, vol. 14, no. 1, 2017, pp. 115–122.

10. H. Issa, T. Sun and M. A. Vasarhelyi, "Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation," Journal of Emerging Technologies in Accounting, vol. 13, no. 2, 2016, pp. 1–20.

11. H. Brown-Liburd, H. Issa and D. Lombardi, "Behavioral Implications of Big Data's Impact on Audit Judgment and Decision Making," Accounting Horizons, vol. 29, no. 2, 2015, pp. 451–468.

12. G. Salijeni, A. Samsonova-Taddei and S. Turley, "Understanding How Big Data Technologies Reconfigure Audit Practices," Accounting, Auditing & Accountability Journal, vol. 34, no. 8, 2021, pp. 1751–1778.

13. S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong and T. Choudhury, "Exploring AI-Driven Approaches for Unstructured Document Analysis and Future Horizons," Journal of Big Data, vol. 11, no. 92, 2024.

14. Y. Huang, Y. Lv, Y. Cao, L. Li and F. Wei, "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," Proc. ACM Int. Conf. Multimedia (ACM MM), 2022.

15. S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie and R. Manmatha, "DocFormer: End-to-End Transformer for Document Understanding," Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021.

16. G. Kim et al., "OCR-Free Document Understanding Transformer," Proc. European Conf. Computer Vision (ECCV), 2022, pp. 498–517.

17. A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, vol. 58, 2020, pp. 82–115.

18. Snowflake Inc., "Snowflake Document AI," https://docs.snowflake.com/en/user-guide/snowflake-cortex/document-ai/overview (accessed 6 December 2025).

19. Sarbanes-Oxley Act of 2002, Pub. L. No. 107–204, 116 Stat. 745, 2002.

20. J. S. Hammersley, L. A. Myers and C. Shakespeare, "Market Reactions to the Disclosure of Internal Control Weaknesses and the Costs of Remediation," Journal of Accounting and Economics, vol. 46, no. 2–3, 2008, pp. 291–321.

21. Committee of Sponsoring Organizations of the Treadway Commission (COSO), Internal Control: Integrated Framework, 2013.

22. Institute of Internal Auditors, International Professional Practices Framework (IPPF): Global Internal Audit Standards, 2024.

23. S. Vasudevan and V. Natarajan, "Automated Population-Level Audit Assurance via AI-Based Document Intelligence," Proc. IEEE SoutheastCon 2026, Huntsville, AL, USA, 2026, pp. 1–8.