GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

Agents for Financial Accounting and Reporting

Application tested

Tester

earlybird × novel Keep Novelty Going

Earlybird's AI-native accounting infrastructure continuously reconciles and maintains accurate, audit-ready books in real time. The system combines autonomous AI agents with expert accountants to continuously process, classify, reconcile, and maintain financial records across invoices, receipts, bank feeds, and payment systems

Knovel Engineering is a Singapore-based AI consultancy and solution provider that helps enterprises and government adopt and operationalise AI in real-world environments. We deliver applied AI, cloud, and data-driven systems; with a strong focus on AI assurance and trust as a key differentiator

How were LLMs used in application?

Agentic | Classification | Detection | Recommendation

What risks were considered relevant and tested?

- Hallucination and Inaccuracy
- Robustness Under Real-World Conditions
- Data Leakage
- Vulnerability to Adversarial Prompts

How were the risks tested?

- **Accuracy:** Real and synthetic financial documents were used to evaluate extraction, classification, and reconciliation accuracy against human-verified data
- **Robustness:** Documents were degraded with realistic blur, skew, rotation, and compression to assess reliability under real-world conditions
- **Data leakage and Adversarial prompts:** Manipulated documents and embedded malicious instructions were used to test resilience against prompt injection, and persistence-related workflow attacks
- **Testing both independent and memory-influenced behaviour:** The system was evaluated both on standalone accounting tasks where each output should not depend on past interactions, and on workflows where past actions and transaction history could influence future decisions through the memory layer

How were test design and evidence evaluated?

- Automated comparison with ground truth and manual testing for stateful tests

Challenges

01 **Testing persistence-enabled workflows:** The system's memory layer was comparatively difficult to manipulate because financial inputs are highly structured and constrained, unlike open-ended chatbot environments with unrestricted free-form text inputs This reduced the attack surface for direct prompt injection and required more sophisticated adversarial techniques targeting extracted financial metadata and document-based workflows

Insights

01 **AI-native accounting needs broader assurance:** Testing must go beyond model accuracy to evaluate workflow reliability, safeguards, persistence behaviour, and human oversight across continuous financial operations

02 **Realistic operational testing is critical:** Testing under real-world document conditions, reconciliation edge cases, and adversarial scenarios provided a more representative assessment of operational reliability

03 **Human oversight and safeguards remain important:** Governance controls and expert review are still critical for persistence-enabled workflows, exception handling, and complex financial decisions

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

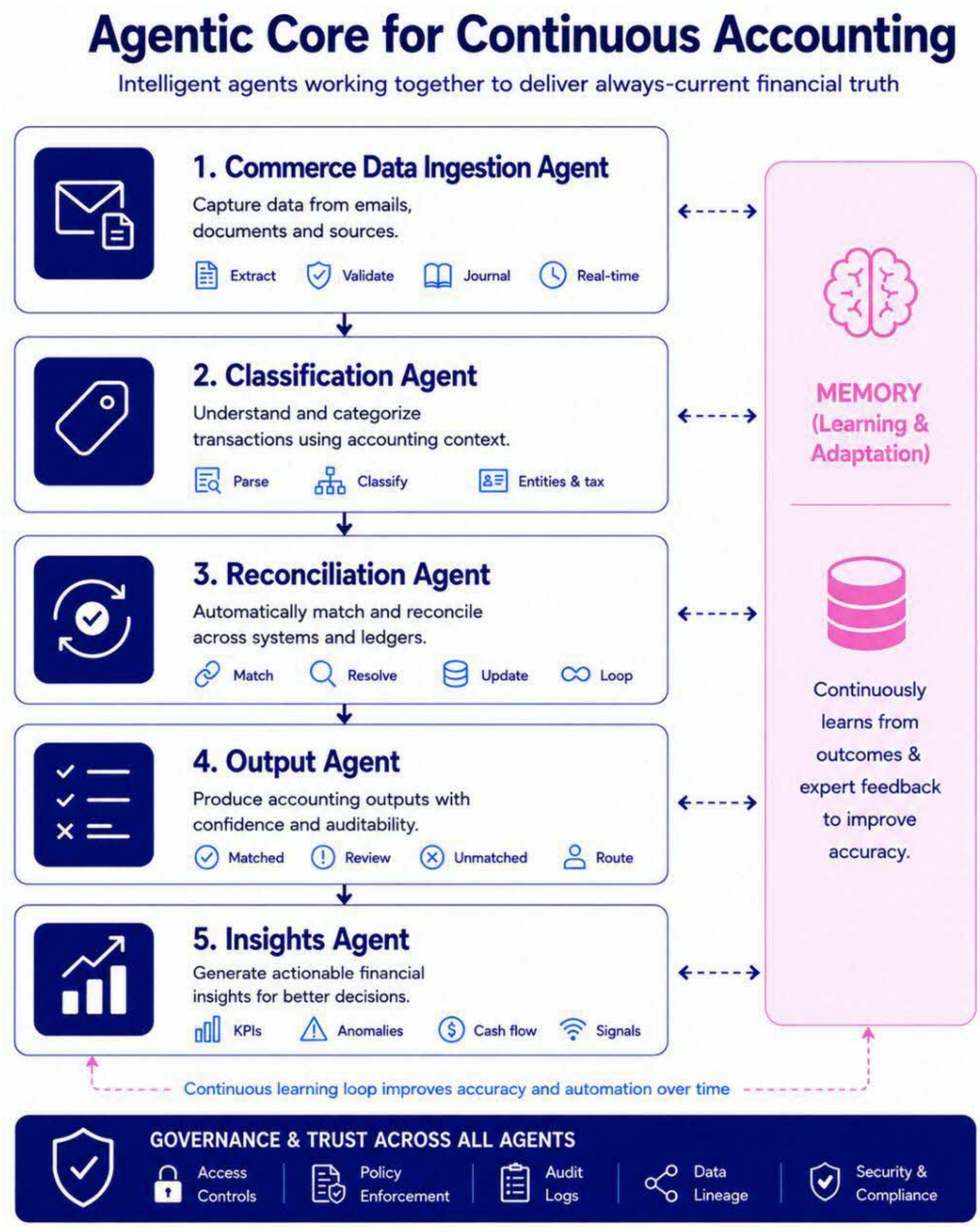
Agents for Financial Accounting and Reporting

Earlybird is an AI-native accounting infrastructure for modern commerce businesses. The platform combines autonomous AI agents with expert accountant oversight to continuously extract, classify, reconcile, and maintain financial records across invoices, receipts, emails, bank feeds, and payment systems — delivering continuously updated, audit-ready books in real time.

Use Case

Earlybird runs continuous accounting workflows including bookkeeping, reconciliation, and financial close processes. The system transforms raw financial inputs into compliant, filing-ready financial records and live operational visibility for businesses. Human oversight is maintained for reconciliation exceptions, compliance-sensitive actions, and financial approvals to ensure accuracy and regulatory integrity.

High-Level Architecture



Deployer and Application

Step 1 **Email-to-ledger agent (automated data-entry)**

An email arrives with a receipt, invoice, or billing info in the body or as an attachment. The agent reads it, extracts the transaction details, classifies the flow of money, and writes a journal entry directly into the ledger. No manual data entry needed.

Step 2 **Classification agent**

Once data is in the system (whether from email or a direct bank/ledger upload), the classification agent parses every line. It matches each bank transaction and each ledger entry against the Chart of Accounts to label what kind of expense, revenue, or liability it represents. A memory layer sits on top, if a transaction pattern has been seen before, it's recognised instantly and skipped for re-analysis (i.e., routing to the reconciliation agent), saving time on recurring items like payroll or SaaS subscriptions.

Step 3 **Reconciliation agent** (4 sub-steps). This is the core of the engine. It runs in four passes:

- **Step 3a (Memory check)**
  Before doing any matching score analysis, it checks if the bank-to-ledger pairing has been seen before. Known patterns are matched instantly without invoking the AI.
- **Step 3b (Pattern matching)**
  For new pairs, it identifies real-world matching patterns, a single bank payment might map to multiple ledger lines (1-to-many), or several payments might combine into one ledger entry (many-to-1).
- **Step 3c (Best match selection)**
  All candidate pairings are evaluated. Duplicates are eliminated and the highest-confidence match is kept for each line.
- **Step 3d (Final sweep)**
  Any items still unmatched get one more pass using the memory store. Confirmed pairings are then written back into memory for next time.

Step 4 **Output buckets.** Every bank transaction ends up in one of four states

- Matched (clean, done)
- Needs review (low confidence, flagged for a human)
- Unmatched (no pairing found, flagged for a human)
- Not required (intentionally excluded, e.g., internal transfers)

Step 5 **AI Insights**

The reconciled data is fed to an LLM that surfaces actionable signals, cash flow anomalies, expense trends, risks, so business owners and finance teams can understand performance without manually compiling reports.

02 Testing Partner and Testing Approach

Knovel Engineering is a Singapore-based AI consultancy that helps enterprises and government agencies operationalise trusted AI. Knovel Engineering specialise in assessing agentic AI systems; designing bespoke, risk-based tests and delivering actionable recommendations to ensure real-world readiness and reliability.

Testing Approach

Knovel Engineering utilised a structured testing framework combining expert crafted adversarial test cases with systematic human evaluation, focusing on metadata extraction from unstructured invoice/bank statements, reconciliation of invoice/bank statement pairs, poisoning of memory layer for persisted cross-document attacks.

03 Risk Assessment and Testing Scope

Unlike traditional accounting systems that relies heavily on manual processing and periodic updates, Earlybird continuously processes live financial information from invoices, receipts, and bank statements to maintain near real-time financial records and operational visibility. As a result, the assessment prioritised assurance areas critical to maintaining accuracy, resilience, governance, and operational trust within continuously operating accounting systems.

Importantly, the objective of the testing was to rigorously assess and validate the platform's reliability, safeguards, and operational resilience under realistic and adversarial conditions — helping strengthen trust in AI-assisted accounting operations.

Earlybird identified the following core assurance domains (with risk terminology aligned with IMDA's Starter Kit for Testing LLM-Based Applications for Safety and Reliability):

- **Hallucination and Inaccuracy**
  (financial extraction & reconciliation integrity)

The autonomous system may incorrectly extract, classify, or reconcile financial information from invoices, receipts, bank feeds, or payment systems — particularly where records are incomplete, ambiguous, or inconsistently formatted.

Why this matters: Accurate extraction and reconciliation integrity are foundational to maintaining reliable financial records, compliant reporting, and trustworthy accounting operations. Assurance testing therefore focused not only on point-in-time extraction accuracy, but also on downstream reconciliation reliability across continuously operating workflows.

- **Robustness Under Real-World Document Conditions**

The system may experience reduced extraction reliability when processing low-quality or degraded financial documents, including blurred receipts, skewed scans, rotated images, compressed files, or partially obscured records.

Why this matters: In real-world business environments, financial documents are often captured through mobile devices, scans, screenshots, or compressed uploads. Systems operating at scale must therefore remain reliable and minimise operational friction for users even when document quality is imperfect.

Risk Assessment and Testing Scope

Data Leakage

The system may unintentionally expose sensitive financial information through prompts, outputs, integrations, logs, or downstream operational workflows.

Why this matters: AI-native accounting systems process highly sensitive financial and operational records. Maintaining confidentiality, access controls, and responsible handling of financial data is essential for regulatory compliance, customer trust, and operational governance.

Vulnerability to Adversarial Prompts
(persistence & adversarial workflow resilience)

Adversarial document inputs or manipulated financial records may influence extracted metadata, transaction classifications, or memory-assisted accounting workflows over time.

Why this matters: Persistence-enabled accounting systems introduce distinct assurance considerations compared to isolated document-processing systems. Assurance testing therefore evaluated not only immediate extraction behaviour, but also the resilience, safeguards, and governance controls surrounding memory-assisted workflows and continuous accounting operations.

Scope of Testing

The testing focused on the following assurance domains across the accounting workflow:

- Financial extraction and reconciliation integrity
- Robustness under realistic operational conditions
- Adversarial document resilience
- Memory-assisted workflow behaviour
- Sensitive financial data governance considerations

The scope was intentionally designed to reflect realistic accounting operations and operational edge cases encountered in production environments.

Test Design

Technical tests were designed to evaluate the identified assurance risks using a combination of automated and manual testing approaches.

Hallucination and Inaccuracy (financial extraction & reconciliation accuracy)

- Knovel Engineering designed tests to evaluate the system's ability to accurately extract and reconcile financial information from invoices and bank statements under realistic accounting conditions. This included:
  - Extract financial metadata accurately
  - Classify financial information correctly
  - Reconcile invoices and bank statement records consistently
  - Maintain accounting accuracy across varying document structures and transaction patterns
- Human-reviewed reference datasets were used to validate extraction outputs and reconciliation behaviour. Metrics focused on:
  - Metadata extraction correctness
  - Classification accuracy
  - Reconciliation consistency
  - Workflow-level reliability

Robustness Under Real-World Conditions

To simulate realistic operational conditions, Knovel Engineering introduced controlled document degradations designed to mirror common real-world capture issues. This included:

- motion blur
- rotation and skew
- compression artefacts
- partial document degradation

The testing focused on visually plausible real-world conditions rather than unrealistic stress-testing scenarios. Metrics focused on extraction reliability and accuracy under degraded document conditions.

Data Leakage and Vulnerability to Adversarial Prompts

To evaluate system's resilience, Knovel Engineering designed scenarios that attempted to influence extraction behaviour or workflow outcomes such as

- adversarial document manipulation scenarios
- persistence-related workflow testing
- validation of safeguards around extracted financial metadata and memory-assisted classification behaviour

Metrics focused on behavioural consistency, validation effectiveness, and resilience against adversarial influence.

Test Implementation

Execution of Tests

The tests were conducted on Earlybird's secure staging environment under controlled access conditions. Testing combined:

- Automated evaluation workflows
- Manual behavioural assessments
- Human-reviewed validation datasets
- Adversarial scenario testing

Key Findings

The testing exercise demonstrated operational reliability across core accounting workflows under realistic operating conditions. Key observations included:

- Extraction and reconciliation workflows performed reliably under standard operating conditions and realistic document-quality variations.
- Isolated adversarial document scenarios were successfully contained at the document level without persistence into subsequent workflows.
- Additional testing of memory-assisted workflows highlighted the importance of layered validation controls, human oversight, and safeguards around persisted classification behaviours within continuous accounting systems.
- The testing exercise reinforced the importance of evaluating both immediate extraction accuracy and longer-term workflow resilience in AI-assisted accounting infrastructure.

Data Used in Testing

Approximately 100 anonymised financial documents, receipts, invoices, and bank statements were used throughout the assessment. The dataset included:

- Representative financial documents
- Synthetic accounting records
- Varied document structures
- Realistic operational edge cases

Cost of Testing

The testing exercise involved collaborative participation from Earlybird and Knovel Engineering, including:

- Environment preparation and technical integration
- Human-reviewed validation dataset preparation
- Execution of automated and manual assurance tests
- Resilience and workflow behaviour analysis

From a resourcing perspective, Earlybird spent 3 days and Knovel Engineering contributed 25 days.

05

Test Implementation

Challenges in Implementation

Testing persistence-enabled workflows

Evaluating how adversarial or manipulated documents could influence memory-assisted accounting workflows over time required system-level testing beyond isolated document checks. The system's memory layer was comparatively difficult to manipulate because financial inputs were highly structured and constrained, reducing the attack surface for direct prompt injection and requiring more sophisticated adversarial techniques targeting extracted metadata and document-based workflows. The exercise highlighted the importance of:

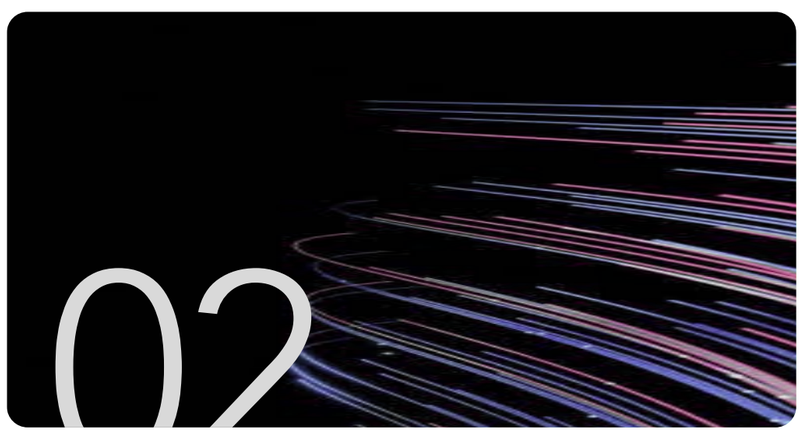
- Layered validation controls
- Human oversight for exception handling
- Safeguards around persisted classification behaviour
- Governance mechanisms within continuous accounting systems

06

Insights/Lessons Learned

Assurance for AI-native accounting requires a broader lens

AI-native accounting systems require a fundamentally different assurance approach from traditional software systems or general-purpose AI applications. Unlike conventional automation workflows, AI-native accounting platforms operate continuously across live financial workflows — processing invoices, receipts, bank feeds, reconciliations, classifications, and compliance-sensitive accounting operations in real time. This introduces unique assurance requirements around accuracy, persistence, operational resilience, governance, and human judgement within continuously running financial systems.



Realistic workflow testing matters

The testing exercise highlighted the importance of evaluating systems under realistic conditions, including:

- Real-world document-quality variation
- Financial reconciliation edge cases
- Adversarial document scenarios
- Persistence-related behaviours

This provided a more representative understanding of how the platform performs in day-to-day accounting operations.

AI-assisted accounting workflows have reached strong operational maturity

The testing demonstrated that a substantial portion of day-to-day accounting workflows can be automated reliably through AI-assisted extraction, classification, reconciliation, and workflow orchestration under realistic operating conditions. The assessment thus highlighted the growing role of persistence-enabled and memory-assisted systems in reducing repetitive manual effort over time.

Accounting remains highly context-sensitive and judgement-driven, particularly across complex multi-channel business environments. Financial classification, reconciliation, and exception handling often depend on operational context, historical transaction behaviour, accounting standards, and evolving business workflows. This reinforced the importance of combining:

- AI-native workflow automation
- memory-assisted operational systems
- structured accounting controls
- and expert financial oversight

within continuously operating accounting infrastructure.

Moving toward system-level AI assurance

Importantly, the exercise reinforced that assurance for AI-native accounting systems must evolve beyond isolated model evaluation toward broader system-level governance, workflow resilience, and operational assurance testing. More broadly, the collaboration represents an important step toward establishing stronger assurance methodologies, governance standards, and resilience frameworks for the next generation of AI-native accounting services and continuously operating financial infrastructure.

AI VERIFY FOUNDATION