



INDEPENDENT EVIDENCE ASSURANCE | WORKED FICTIONAL EXAMPLE

AI Vendor Performance & Claims Audit

Illustrative matter: Northbridge Customer Operations LLC

Review subject: Veridian Support AI

THE BUYER DECISION

Do the evaluation design, denominators, errors and observed workflow support the vendor's commercial performance claims?

ILLUSTRATIVE EVIDENCE POSITION

Reject reliance on the unqualified 98% statement; commission a blinded production-like pilot with jointly agreed metrics.

DEMO REFERENCE

DEMO-05-2026-001

EVIDENCE DATE

11 OCT 2026

STATUS

ILLUSTRATIVE ONLY

Prepared for prospective US clients. No real client, counterparty or evidence packet.



Document control and limits of reliance

Aperture Research Works | AI Vendor Performance & Claims Audit | Demonstration Edition, 11 October 2026.

Control	Record
Matter reference	DEMO-05-2026-001
Scenario	Wholly fictional; invented organisations, documents, persons, dates and figures
Prepared for	Technology purchasers, independent investors and owner-led enterprises
Evidence cut-off	11 October 2026 for public methodological references only; no target-specific sources retrieved
Approval status	Demonstration document. Human approval required before public release.
Professional boundary	No certification of AI safety, model-weights access, penetration testing or guarantee of future model behaviour.

READ THIS FIRST: All MOCK exhibit labels and findings describe invented records created to show the investigation method. They are not allegations about, or findings concerning, any real person or entity.

The decision this work supports

A fictitious software vendor claims 98% accurate ticket resolution, 60% cost savings and fully autonomous handling. Its disclosed evaluation is limited to curated, single-turn data and omits human escalation from the denominator.

Central evidence question: Do the evaluation design, denominators, errors and observed workflow support the vendor’s commercial performance claims?

Commercial moment: Before relying on a vendor’s claimed accuracy, savings or automation rate

Economic context in the fictional case: \$145,000 proposed annual AI subscription



Executive position: what should change now?

Reject reliance on the unqualified 98% statement; commission a blinded production-like pilot with jointly agreed metrics.

The scenario illustrates why a clean-looking presentation is not a complete evidence basis. The purpose is to identify source-attributed decision issues early, distinguish conflicting explanations and direct additional work only where it changes the decision.

Three practical buyer benefits

- Turn vague accuracy claims into a repeatable test protocol.
- Identify who counted exceptions, interventions and hidden operating costs.
- Give procurement measurable acceptance and pilot-termination criteria.

From a claim to decision-ready evidence



The evidence trail is part of the deliverable, not an invisible step.

Each claim is matched to an invented source object and tested for entity attribution, date, origin, contrary evidence, plausible alternatives and the specific document that would close the gap.

Finding classification used in this example

SUPPORTED means adequate evidence for the narrow stated point. QUALIFIED means partial support with a material limitation. MATERIAL CONFLICT or CONTRADICTED means the invented record contains an explicit inconsistency. UNRESOLVED and NOT ESTABLISHED are evidence insufficiency classifications, not findings of misconduct.



Claim-to-evidence matrix

All rows below are constructed data, not actual research results. Exhibit references identify simulated document objects, not attachments or verified public records.

# / claim	Illustrative record outcome	Evidence state
1. The system resolves 98% accurately	The 98% rate is on filtered FAQ examples, not full production cases.	QUALIFIED
2. It operates without humans	25% of the sample was excluded after human escalation.	MATERIAL CONFLICT
3. Average cost falls by 60%	Software cost, QA, migration and exception handling were omitted.	CONTRADICTED
4. Results generalise to our industry	Only non-regulated FAQ tasks are represented.	NOT ESTABLISHED
5. No confidential text enters third-party models	Subprocessor appendix has an unspecified LLM endpoint.	UNRESOLVED
6. Training data is representative	No sampling basis or minority-case analysis is documented.	NOT ESTABLISHED
7. False completions are rare	An error category is merged into successful resolutions.	MATERIAL CONFLICT



# / claim	Illustrative record outcome	Evidence state
8. Response time satisfies service commitment	P95 excludes retrieval failures and timeout retries.	QUALIFIED
9. Security testing is complete	Reports describe planned rather than completed review.	NOT ESTABLISHED
10. Vendor pilot is directly comparable	Pilot includes vendor-curated examples and assisted mode.	UNRESOLVED

Material findings with original-document tests

01. The system resolves 98% accurately

Constructed record: MOCK-01 pitch / MOCK-02 benchmark. State: QUALIFIED.

Worked observation: The 98% rate is on filtered FAQ examples, not full production cases.

Counter-evidence test: look for independently attributable, contemporaneous records, confirm the relevant entity and effective dates, and actively search for a reasonable alternative before describing the discrepancy as adverse.

Release gate for the buyer: Specify denominator, ground truth and failure policy.

02. It operates without humans

Constructed record: MOCK-03 workflow map. State: MATERIAL CONFLICT.

Worked observation: 25% of the sample was excluded after human escalation.

Counter-evidence test: look for independently attributable, contemporaneous records, confirm the relevant entity and effective dates, and actively search for a reasonable alternative before describing the discrepancy as adverse.

Release gate for the buyer: Count interventions in end-to-end resolution rate.

03. Average cost falls by 60%

Constructed record: MOCK-04 ROI worksheet. State: CONTRADICTED.

Worked observation: Software cost, QA, migration and exception handling were omitted.

Alternative explanations to test: the stated discrepancy could reflect a recording delay, a superseded document, an affiliated entity or a legitimately scoped exception. These are candidate explanations only; none is proven by



this illustration.

Release gate for the buyer: Compute full annual operating cost under realistic case mix.

04. Results generalise to our industry

Constructed record: MOCK-05 customer case studies. State: NOT ESTABLISHED.

Worked observation: Only non-regulated FAQ tasks are represented.

Counter-evidence test: look for independently attributable, contemporaneous records, confirm the relevant entity and effective dates, and actively search for a reasonable alternative before describing the discrepancy as adverse.

Release gate for the buyer: Run sector-specific blinded test set.

05. No confidential text enters third-party models

Constructed record: MOCK-06 dataflow diagram. State: UNRESOLVED.

Worked observation: Subprocessor appendix has an unspecified LLM endpoint.

Counter-evidence test: look for independently attributable, contemporaneous records, confirm the relevant entity and effective dates, and actively search for a reasonable alternative before describing the discrepancy as adverse.

Release gate for the buyer: Obtain subprocessor register, retention and routing terms.

06. Training data is representative

Constructed record: MOCK-07 dataset card. State: NOT ESTABLISHED.

Worked observation: No sampling basis or minority-case analysis is documented.

Alternative explanations to test: the stated discrepancy could reflect a recording delay, a superseded document, an affiliated entity or a legitimately scoped exception. These are candidate explanations only; none is proven by this illustration.

Release gate for the buyer: Get evaluation protocol, slice metrics and provenance.

07. False completions are rare

Constructed record: MOCK-08 outcome matrix. State: MATERIAL CONFLICT.

Worked observation: An error category is merged into successful resolutions.

Counter-evidence test: look for independently attributable, contemporaneous records, confirm the relevant entity and effective dates, and actively search for a reasonable alternative before describing the discrepancy as



adverse.

Release gate for the buyer: Separate factual error, incorrect close and escalation.

08. Response time satisfies service commitment

Constructed record: MOCK-09 latency chart. State: QUALIFIED.

Worked observation: P95 excludes retrieval failures and timeout retries.

Counter-evidence test: look for independently attributable, contemporaneous records, confirm the relevant entity and effective dates, and actively search for a reasonable alternative before describing the discrepancy as adverse.

Release gate for the buyer: Recompute tail latency including failures.

09. Security testing is complete

Constructed record: MOCK-10 assurance summary. State: NOT ESTABLISHED.

Worked observation: Reports describe planned rather than completed review.

Alternative explanations to test: the stated discrepancy could reflect a recording delay, a superseded document, an affiliated entity or a legitimately scoped exception. These are candidate explanations only; none is proven by this illustration.

Release gate for the buyer: Obtain dated test evidence and independent scope.

10. Vendor pilot is directly comparable

Constructed record: MOCK-11 pilot brief. State: UNRESOLVED.

Worked observation: Pilot includes vendor-curated examples and assisted mode.

Counter-evidence test: look for independently attributable, contemporaneous records, confirm the relevant entity and effective dates, and actively search for a reasonable alternative before describing the discrepancy as adverse.

Release gate for the buyer: Use buyer-controlled sample and acceptance criteria.

Reproducible quantitative check

The following values are fictional and their arithmetic is shown so a client can understand the checking method. No valuation, legal loss, statutory liability or security rating is inferred.



Resolution-rate denominator test

Illustrative test tickets	400
Escalated / removed from vendor accuracy denominator	100
Correct autonomously closed tickets	219
Autonomous correct-close rate (%)	54.75

$219 / 400 = 54.75\%$ correct autonomous closure on the invented full sample. This does not estimate any actual vendor.

Re-performance control: the calculations above are to be recomputed from the underlying constructed inputs. Where the third or fourth number is a derived value, test its inclusion and whether any value is already a subset of another amount. A buyer must never add an item twice merely because it appears under two risk headings.

Alternative hypotheses and challenge log

Hypothesis	What would support it	What would undermine it
Benign documentation gap	Time-stamped correction, matching independent counterparty confirmation	Conflicting identifiers or unexplained account changes
Scope or definition mismatch	Contemporaneous definitions and consistent signed annexes	Retroactive redefinition after the discrepancy was identified
Material misstatement / control weakness	Repeated exceptions, inconsistent original records, ineffective approvals	Authenticated correction, complete reconciled trail and credible independent corroboration

No adverse conclusion is reached in this hypothetical example. A real matter requires applicable fairness, right-of-reply considerations, lawful collection methods and jurisdiction-specific professional review.

Evidence provenance and chain of reliance

For every material source record, the production team should preserve: native filename and format; collector and lawful access method; capture timestamp and time zone; document hash where possible; issuing organisation; entity and period covered; reliability grading; amendments or superseded versions; contradiction trail; and reviewer approval.



Evidence object	Why it matters	Reliance boundary
MOCK-01 pitch	The system resolves 98% accurately	Specify denominator, ground truth and failure policy.
MOCK-03 workflow map	It operates without humans	Count interventions in end-to-end resolution rate.
MOCK-04 ROI worksheet	Average cost falls by 60%	Compute full annual operating cost under realistic case mix.
MOCK-05 customer case studies	Results generalise to our industry	Run sector-specific blinded test set.
MOCK-06 dataflow diagram	No confidential text enters third-party models	Obtain subprocessor register, retention and routing terms.
MOCK-07 dataset card	Training data is representative	Get evaluation protocol, slice metrics and provenance.

What the buyer can do with the result

Stage	Decision use	Appropriate next owner
Initial screening	Do not elevate unverified representations to confirmed facts.	Buyer / internal project lead
Before irreversible payment or contract	Resolve documentary release gates and obtain written conditions.	Buyer, finance leader and counsel
Negotiation or remediation	Send a focused discrepancy schedule rather than an unstructured accusation.	Authorised counterparty contact
Specialist escalation	Separate accounting, legal, technical or regulated determinations.	Qualified professional in relevant jurisdiction
Final record	Preserve traceable limitations and a source-backed decision note.	Buyer and nominated reviewer



Priority release gates for the illustrative engagement

PRIORITY / EVIDENCE RELEASE PATH

- Specify denominator, ground truth and failure policy.
- Count interventions in end-to-end resolution rate.
- Compute full annual operating cost under realistic case mix.
- Run sector-specific blinded test set.
- Obtain subprocessor register, retention and routing terms.
- Get evaluation protocol, slice metrics and provenance.

Illustrative evidence-object index

X-01 | MOCK-01 pitch / MOCK-02 benchmark | Purpose: The system resolves 98% accurately. Result: The 98% rate is on filtered FAQ examples, not full production cases. | Constructed illustration only.

X-02 | MOCK-03 workflow map | Purpose: It operates without humans. Result: 25% of the sample was excluded after human escalation. | Constructed illustration only.

X-03 | MOCK-04 ROI worksheet | Purpose: Average cost falls by 60%. Result: Software cost, QA, migration and exception handling were omitted. | Constructed illustration only.

X-04 | MOCK-05 customer case studies | Purpose: Results generalise to our industry. Result: Only non-regulated FAQ tasks are represented. | Constructed illustration only.

X-05 | MOCK-06 dataflow diagram | Purpose: No confidential text enters third-party models. Result: Subprocessor appendix has an unspecified LLM endpoint. | Constructed illustration only.

X-06 | MOCK-07 dataset card | Purpose: Training data is representative. Result: No sampling basis or minority-case analysis is documented. | Constructed illustration only.

X-07 | MOCK-08 outcome matrix | Purpose: False completions are rare. Result: An error category is merged into successful resolutions. | Constructed illustration only.

X-08 | MOCK-09 latency chart | Purpose: Response time satisfies service commitment. Result: P95 excludes retrieval failures and timeout retries. | Constructed illustration only.

X-09 | MOCK-10 assurance summary | Purpose: Security testing is complete. Result: Reports describe planned rather than completed review. | Constructed illustration only.

X-10 | MOCK-11 pilot brief | Purpose: Vendor pilot is directly comparable. Result: Pilot includes vendor-curated examples and assisted mode. | Constructed illustration only.



Methodological source register

The links below identify real public reference material used to shape the methodology. They are NOT documents obtained about the fictional organisations or verification of any mock finding.

1. NIST AI Risk Management Framework, AI RMF 1.0
<https://www.nist.gov/itl/ai-risk-management-framework>
2. NIST AI RMF Generative AI Profile
<https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence>
3. FTC final action regarding Workado AI detection claims
<https://www.ftc.gov/legal-library/browse/cases-proceedings/2323092-content-scale-ai>

Scope, professional boundaries and limitations

No certification of AI safety, model-weights access, penetration testing or guarantee of future model behaviour.

The proposed fee is an indicative quotation hypothesis for a bounded service: \$3,750. It is not evidence of prior completed sales or a guaranteed price. Actual scope, geographic coverage, legal authority, source costs, timing, conflict checks and responsible-review resources must be approved in writing before work begins.

This report is intentionally a constructed demonstration, not a certified assessment. No claims of actual source retrieval, customer success, audit assurance, regulated investigative licensing or market-leading accuracy are made.

What to send for a scope review

Start with a short written account of the decision to be made, the legal entities involved, the date by which the decision is needed, the approximate geography and the single most material unresolved claim. Do not send confidential or personal documents in an initial public enquiry.

Aperture Research Works can assess whether a narrowly defined evidence review would be useful and lawful. Every engagement is separately accepted and scoped. The buyer retains decision responsibility.

Visit: apertureforensic.com | [Discuss a Matter](#) | info@apertureforensic.com