The Reality Ledger

Evidence preservation and decision assurance for AI assisted workflows

Christos Birkitt | 3 November 2026 | Version 1.0

An AI-assisted decision can become unsound without an invented fact. A valid source can support a narrower claim than the final memo makes. A provisional approval can become an unconditional clearance. An unsuccessful search can become an assertion that no adverse record exists. The words remain plausible while the basis for acting on them changes.

The Reality Ledger is a proposed audit method for examining those changes across document intake, extraction, retrieval, synthesis, human review, and release. Its unit of analysis is a decision transaction: a particular output, the evidence available to produce it, the transformations it underwent, and the action it was intended to support.

The method asks whether the final decision preserves the meaning, limits, and authority of its evidence. It requires an explicit evidence boundary, links between claims and their support, a record of qualifications that must survive each handoff, and release rules that cannot be satisfied by averaging away a critical failure.

The central rule

> *A decision-relevant qualification remains an open obligation until it is preserved in the downstream decision, resolved by identified evidence, or assigned an explicit restriction on the permitted action. Silence does not resolve it.*

This rule turns caveats, contradictory records, missing evidence, and human judgments into inspectable audit objects. A reviewer can identify where a condition disappeared, which recommendation depended on it, who was authorized to resolve it, and what must change before the output may proceed.

What this specification provides

The following pages define the audit record, a five-vector rubric with every score from 0 to 4, separate diagnostic and release decisions, an illustrative acquisition case, implementation requirements, and a validation study. The accompanying reference scorer implements the routing rules; it does not inspect documents or determine whether evidence is true.

Status: proposed method, version 1.0. Weights and thresholds are initial policy parameters, not empirically validated estimates. The case is an instructional fixture, not a completed client engagement. Regulatory references are an illustrative crosswalk, not a claim of certification or conformity. A separate author note addresses development provenance and rights.

1 Purpose and contribution

A technical evaluation is useful only for the risks it measures. A system may perform well on a benchmark while a particular deployment omits a mandatory document, loses a time limit, or permits an unsupported approval. The resulting gap is a mismatch between evaluation coverage and the evidence required for the decision.

Existing research already addresses atomic factual support, retrieval quality, citation support, and diagnostic evaluation of RAG systems. FActScore decomposes text into atomic facts; RAGAs evaluates retrieval and generation; ALCE examines citation quality; and RAGVUE provides more detailed diagnostics. The Reality Ledger builds on those capabilities. It does not claim that existing evaluations are uniformly limited to formatting or isolated responses. [1–4]

The proposed contribution is an operational specification connecting evidence requirements to workflow transitions and release decisions. Its distinctive emphasis is the obligation to account for a material condition at each transformation, including the human decision. Whether this integration improves detection, review quality, or cost must be tested against existing methods.

Four questions an audit must keep separate

Question	What a satisfactory answer establishes
Was the source represented faithfully?	The output preserves the source's actual assertion, scope, and qualifications.
Is the source adequate for this question?	Its authority, date, coverage, and limitations support the intended use.
Does the evidence justify the inference?	The reasoning has identified premises and alternatives; inference is not presented as an observed fact.
Is the proposed action permitted?	Applicable evidence requirements, review duties, and release conditions have been met.

Faithful quotation does not establish that a source is correct. An accurate account of an allegation remains an allegation. A current document can repeat an outdated assessment. The ledger therefore preserves the status of evidence rather than promising custody of an unquestionable ground truth.

Where to apply the method

Use it when an AI-assisted output supports a consequential decision and the supporting record can be bounded: acquisition diligence, legal file synthesis, research assessment, policy analysis, or a regulated operational review. Start with one workflow and one defined action. A case-level finding does not establish the reliability of every use of the model.

Throughout this specification, must describes a requirement of the proposed Reality Ledger profile. It does not assert that a legislature, regulator, or standard requires that particular implementation. A system may satisfy this profile and still require separate testing for security, discrimination, privacy, robustness, or other domain risks.

2 Define the decision and evidence boundary

Before examining the model's answer, the workflow owner and a domain reviewer define what the answer is allowed to support. The boundary is approved independently of the output so that the model cannot determine the evidence requirements by selecting the easiest documents to summarize.

The required profile

Profile field	Required content
Decision and audience	The action, decision-maker, intended use, and prohibited uses.
Evidence universe	Repositories, document classes, entities, jurisdictions, time period, and known exclusions.
Required questions	What must be established, qualified, or escalated before the action can proceed.
Materiality	Which errors could change the action, conditions, timing, eligibility, escalation, or required disclosure.
Review authority	Who may verify evidence, accept residual risk, resolve disagreements, and authorize release.
Control version	Rubric, gate definitions, weights, thresholds, retention policy, and change owner.

A material claim is one whose correction could reasonably change the action or its conditions. Materiality also includes mandatory disclosures and controls. Assess defects individually and together: several minor omissions may collectively change a decision. Record each designation and its basis before scoring; do not downgrade a failed claim after seeing its score.

The required-evidence manifest lists expected records and records actually received. It distinguishes an absent document from a file that arrived but failed extraction, an extracted record excluded by filtering, and a relevant record omitted from the retrieval result. These are different failure locations with different owners.

Three boundaries must remain visible

- **Corpus boundary:** what records were available, expected, excluded, inaccessible, or outside the search.
- **Temporal boundary:** the observation cutoff and each record's effective date, expiry, amendment, or supersession status.
- **Decision boundary:** what the evidence permits now, what remains conditional, and what requires further work.

A fixed corpus is reproducible, but not necessarily complete. The owner must justify why the selected sources are adequate for the proposed decision. If that justification is unavailable, the audit reports the limitation and withholds a release recommendation. If the intended action changes, the profile must be reconsidered before the prior assessment is reused.

RL-BASE-1.0 fixes the scoring rules and requires human review, all five vectors, and all five gates. The local action and evidence profile has a separate versioned identity. A changed action requires reassessment. Different weights, gates, or review requirements need separately approved and validated rules.

3 Record claims and evidence obligations

Decompose the output into assertions small enough to evaluate without hiding a disputed premise inside an otherwise accurate sentence. Preserve the original text and location. A sentence may contain several claims; a claim may require several sources. Repeated claims receive linked identifiers so repetition cannot inflate the denominator.

Classify each item as a reported fact, source-attributed allegation, inference, prediction, recommendation, or negative assertion. An inference is permissible when its premises and limits are explicit. General model knowledge must be identified as such and checked to the standard required by the profile before it supports a material decision.

Minimum claim record

Record component	Example from a provisional easement
Identity and text	C-01; exact output span; stage and output version.
Type and dependency	Reported fact; supports the utility-access premise for an acquisition recommendation.
Scope and status	Parcel B; provisional; renewal outstanding; effective through the stated expiry date.
Evidence links	D-01 version 1, paragraph 2; supporting and contradictory records linked separately.
Support assessment	Supported, qualified, unsupported, contradicted, or not assessed; with a short reason.
Obligation and owner	O-01; retain the expiry and renewal condition; owner is the designated acquisition reviewer.

An evidence obligation specifies what must be established or carried forward for a material claim or required question. It contains an identifier, the relevant evidence, the condition that must survive, the downstream decision it affects, a responsible owner, and a closure criterion. Obligations may originate in source limitations, the approved task profile, or a discovered contradiction.

Permitted obligation states

An obligation is **open** until assessed. It becomes **carried** when its limitation remains explicit and the permitted action respects it; **resolved** when identified evidence justifies closure; **restricted** when the action is narrowed or deferred; or **violated** when a downstream claim or action exceeds the evidence. Record each transition, its reason, and its author.

For example, a renewal condition cannot disappear because a summary is shorter. It may be carried into the memo, resolved by an authenticated renewal within its scope, or reflected in a recommendation to defer closing. A promise to check later does not satisfy a requirement to verify before release.

A correction to one material premise triggers inspection of every linked recommendation, approval, and downstream reuse. This dependency record limits the chance that a repaired sentence leaves an unsupported conclusion elsewhere in the workflow.

4 Inspect meaning at each handoff

For each material claim and obligation, compare the state before a transformation with the state after it. Inspect source to extraction, extraction to retrieved context, context to synthesis, synthesis to reviewed output, and reviewed output to the action record. A final-output check can detect a defect; paired states help locate where it first became visible.

Change to inspect	Example of a material change
Scope	A finding about Parcel B becomes a statement about the whole site.
Certainty or status	An unconfirmed report becomes a verified fact.
Condition	Access subject to renewal becomes unrestricted access.
Time	A 2022 assessment becomes the current risk classification.
Disagreement	A disputed conclusion becomes the only conclusion shown.
Support	A directly supporting record is replaced by a citation on the same topic.
Decision permission	A recommendation to investigate becomes approval to proceed.

Record the exact before and after spans, the changed attribute, the evidence that permits or contradicts the change, and the affected decision. A paraphrase is acceptable when it preserves meaning. A change in wording is not itself a failure. Conversely, keeping the word may does not make a statement faithful if its subject, time range, or implication has broadened.

The burden for closing a condition

A condition may be retired only when the ledger identifies the resolving evidence, explains why it applies to the same entity and time period, and records the authorized resolver. If a newer report conflicts with an older report, evaluate applicability and authority before treating it as controlling. Date sorting alone cannot determine precedence.

A finding should identify the earliest observed defective transition and any later control that failed to catch it. This supports remediation without assuming a single cause. An OCR defect, a retrieval filter, an overbroad synthesis, and an inadequate review can contribute to the same transaction.

Evidence for the problem and limits of the claim

Research on scientific summarization has directly examined generalization beyond the scope of source claims. That supports investigating qualification loss; it does not establish its frequency in acquisition, legal, or financial workflows. Here, epistemic smoothing is an operational label for observed strengthening or broadening of evidence across a handoff. [5]

Source comparisons may be assisted by entailment models or language models. Their findings remain subject to evaluation and disagreement review. An automated judge must not certify its own adequacy merely because it returns a confident explanation.

5 Test completeness and negative assertions

An output cannot reveal every relevant fact that a system never received. Completeness must therefore be tested against an independently defined set of required questions, expected sources, known conflicts, and search limits. A fluent summary of retrieved passages cannot establish that the retrieval set was adequate.

Reconcile the evidence path

For each required document, reconcile receipt, page or record count, extraction status, indexing status, retrieval eligibility, retrieval result, and downstream use. Check for empty extraction, partially missing pages, damaged tables, lost footnotes, missing attachments, and permission-related exclusions. A nonempty text field is evidence of extraction, not proof that extraction was complete or correct.

Where document families contain amendments, addenda, or superseded versions, record the relationship. Test entity identifiers and effective dates. A later file may concern a different parcel; an earlier instrument may remain legally operative. The precedence rule must reflect the domain rather than automatically prefer the newest timestamp.

Requirements for a bounded negative finding

Required element	Question the auditor must answer
Universe	Which repositories, document types, and jurisdictions were searched?
Identity	Which entity identifiers, names, aliases, and relevant relationships were checked?
Time	What dates and observation cutoff does the finding cover?
Access and processing	Were relevant records inaccessible, missing, unreadable, or excluded?
Search adequacy	What queries or matching rules were used, and how were known failure modes tested?
Wording	Does the output limit its conclusion to the actual search and disclose material gaps?

No matching record was located in the defined search is a claim about a search result. No litigation exists is a claim about the world. The second requires a different and usually much stronger basis. Completing ingestion of the received files does not bridge that difference.

> *Example of a bounded statement: No matching civil action was identified in the specified county index under the listed parcel-owner identifiers as of the recorded cutoff. One supplied docket attachment remains unreadable. Litigation clearance is therefore unresolved.*

An unresolved required question may still permit an informational memo that explicitly recommends deferral. It does not permit the originally requested unconditional approval. Record the changed action and assess it under the applicable profile. This distinction prevents the audit from treating every acknowledged limitation as a reason to suppress useful information.

6 Evaluate the human decision

Human review is effective only when the reviewer has the evidence, competence, authority, and practical ability to change the outcome. A logged approval proves that an interaction occurred. It does not, by itself, establish that the person understood a condition, checked a contradiction, or could stop release.

Required evidence of review

- The identified reviewer and role, the output version reviewed, and the applicable authority.
- The material claims and obligations checked, with links to the relevant evidence.
- The disposition of each material conflict or gap and any restriction carried into the decision.
- A concise decision-specific justification, the accepted residual risk, and any required second approval.
- The final action, timestamp, escalation path, and ability to reject, defer, or reverse the recommendation.

The justification should explain the observable basis for the action. It need not expose private deliberations or a model's hidden reasoning. A reusable phrase such as reviewed and approved is inadequate when the profile requires resolution of identified exceptions.

How to use interaction telemetry

Elapsed time, source-link openings, scrolling, and override frequency can identify transactions for inspection. They must not serve as stand-alone proof of careful review or negligence. A reviewer may have examined the evidence elsewhere; a user can click every link without understanding it. Interpret telemetry in context and retain only what is necessary under the applicable privacy and access policy.

For selected high-consequence tasks, test whether requiring an initial assessment before showing the AI recommendation improves decisions. Research on cognitive forcing provides a basis for that experiment, including possible usability costs. It does not establish that mandatory clicks or delays improve this workflow. [6]

Failure and uncertainty are different findings

If an authenticated event log shows that a mandatory review was bypassed, the review gate fails. If the auditor cannot establish whether a review occurred because the record is inaccessible, the gate is unknown. If a complete review packet exists but gives an unsupported rationale, the audit evaluates that rationale. All three situations may prevent release, but they require different remediation.

Routine analyst approval cannot override a hard gate. A designated authority may authorize an alternative action through a documented process, but cannot change a failed evidence finding into a pass. The original finding, subsequent decision, conditions, and accountable owner remain visible.

7 Preserve the state needed for reconstruction

An audit record must preserve what the system and reviewer actually received and produced. Recording a model name and a prompt is insufficient when source versions, retrieved context, tool results, or the final edited output can change.

Record family	Minimum useful record
Source and ingestion	Source identifier, version, origin, effective date, content fingerprint where appropriate, page/record counts, extraction results, and exceptions.
Retrieval	Query, filters, index/retriever version, returned passages, ordering, document versions, and excluded or failed sources relevant to the decision.
Generation	Available provider/model identifiers, prompt/template version, actual input and output, relevant parameters, tool calls, and timestamps.
Review and action	Reviewed output version, reviewer role, evidence-linked findings, edits, approvals, restrictions, and final action.
Audit and change	Profile version, assessor, adjudication, scores, gate evidence, remediation links, and reassessment identifiers.

Store the actual artifacts or controlled references that permit authorized reconstruction. When a vendor does not expose an exact model build or nondeterministic behavior cannot be reproduced, record that limitation. Reconstruction of the decision state is different from regenerating identical text.

Content hashes can help detect changes only when compared with a trusted record. A hash alone does not establish source truth, authorship, or the time of creation. Appropriate authentication, access controls, retention, and change history remain necessary. The architecture may use an append-only event store or another controlled record system; blockchain is not a requirement.

Lineage and evidence are related but distinct

A complete history can faithfully document an unsupported conclusion. Conversely, a correct conclusion may be unauditable if its decision state was not retained. Keep these findings separate. Provenance models such as W3C PROV provide an established vocabulary for entities, activities, agents, and derivation; the ledger can use that vocabulary without treating lineage as proof of factual accuracy. [7]

Restrict access to confidential and personal information according to the workflow's legal and organizational requirements. Minimize unnecessary copies and record retention limits. A custodian may verify protected evidence for an authorized audit, but the report must describe the resulting scope and access limitation.

Never overwrite the original transaction with a corrected output. Link the original finding to the replacement, explain the change, and record a new assessment. A remediation history is part of the audit evidence, not an embarrassment to remove.

8 Score provenance and uncertainty preservation

The vectors are related dimensions. A documented defect limits the rating to its lower anchor; if several defects apply, use the lowest. Otherwise award the highest fully evidenced anchor. Score 4 requires score 3 plus the specified independent checks. Do not average correct claims to dilute a material defect. Insufficient assessment evidence remains unknown.

V1 Provenance and support fidelity Weight 30 percent

Score	Anchor
0	A material source or citation is fabricated or substituted, or affirmative evidence establishes that the supporting record cannot be recovered.
1	Sources are identifiable, but a material claim is unsupported, contradicted, misattributed, or presented as fact when it is only an inference.
2	Material claims are adequately supported; a localized nonmaterial attribution or inference-labeling defect remains.
3	Material claims link to adequate, versioned evidence; allegations, inferences, and recommendations are distinguished and limitations are recorded.
4	Score 3 is met and an independent check verifies the material support links, source adequacy, and plausible competing interpretations.

Assess the relationship between claim and evidence, not merely whether the citation resolves. Several sources may jointly support one inference. The record must explain their roles and must preserve any source disagreement.

V2 Uncertainty and scope preservation Weight 20 percent

Score	Anchor
0	The workflow's material qualification history is demonstrably destroyed or corrupted, so the change in evidentiary status cannot be established.
1	A material condition, uncertainty, scope limit, expiry, or dispute is removed or strengthened without a documented resolving basis.
2	Material qualifications survive; a nonmaterial qualification is inconsistently expressed or a local record is incomplete.
3	Required qualifiers are preserved or resolved with evidence at each relevant handoff, and the action respects unresolved limits.
4	Score 3 is met and an independent paired-state review verifies semantic preservation, including relevant scope, temporal, and contradiction challenges.

This vector evaluates preservation of evidentiary status. It does not measure probabilistic confidence calibration. Any claim about calibration requires separate predictions, observed outcomes, and a defined statistical assessment.

For score 4, a second qualified assessor who did not create or approve the transaction documents the checks and initial findings before reconciliation. The profile specifies any additional organizational independence. This is verification within scope, not accredited certification.

9 Score completeness and human review

V3 Required evidence and contextual completeness Weight 20 percent

Score	Anchor
0	The required evidence boundary is demonstrably absent or manipulated so the adequacy of the decision record cannot be assessed.
1	A required source, material conflict, applicable superseding record, or required search limitation is omitted from the decision basis.
2	Decision-critical requirements are met; a disclosed noncritical context or reconciliation gap remains.
3	The required manifest is reconciled, relevant conflicts and precedence are assessed, and negative claims remain within the verified search boundary.
4	Score 3 is met and an independent reviewer reconciles the relevant source inventory and tests plausible omission, identity, and supersession failures.

Completeness refers to the approved boundary and the adequacy of that boundary for the action. A percentage of indexed files is not a completeness score. One missing disqualifying record can matter more than thousands of successfully processed pages.

V4 Human review and override integrity Weight 15 percent

Score	Anchor
0	A required approval is forged, the entire mandatory review stage is demonstrably bypassed, or an unauthorized override determines release.
1	An approval or review event is recorded, but the required decision-specific verification or resolution of material exceptions is absent or unsupported.
2	Material checks and authority are documented, but a noncritical rationale, escalation, or review-record gap remains.
3	An authorized reviewer records the evidence checked, dispositions of material exceptions, decision rationale, and final action or restriction.
4	Score 3 is met and an independent review verifies the rationale, exception handling, and practical ability to reject or correct the recommendation.

A transaction may receive a low review score without a conclusion about the reviewer's mental state. The finding concerns the required control and its evidence. Allegations of automation bias require their own basis.

Adjudicate disagreements explicitly

Two assessors record initial scores independently for calibration samples and designated high-consequence cases. Preserve disagreements, the evidence reviewed, and the adjudicated result. Unresolved disagreement about a material support or gate finding remains a hold unless a confirmed failure already requires blocking. Do not silently choose the more favorable score.

These anchors distinguish identifiable but inadequate controls at score 1 from specified breakdowns at score 0. Both can prevent release. The severity of the business consequence is reported separately from the vector number.

10 Score lineage and calculate the diagnostic result

V5 Lineage and reconstruction Weight 15 percent

Score	Anchor
0	Affirmative evidence shows that essential records are falsified, destroyed, or unavailable so the material decision state cannot be reconstructed.
1	Some records exist, but known missing versions or artifacts prevent reliable reconstruction of a material step.
2	The material decision basis can be reconstructed, with a disclosed noncritical metadata or retention-control gap.
3	Versioned source, context, output, review, and action records reconstruct the material decision state; unavailable provider details are explicit.
4	Score 3 is met and an independent reconstruction check verifies the records and the relevant integrity, access, and change controls.

The Epistemic Integrity Score

When all five scores are known, calculate EIS on a 0–100 scale. Each score is an integer from 0 through 4. The baseline weights are 0.30, 0.20, 0.20, 0.15, and 0.15 and sum to one.

$$EIS = 25\left(0.30\,S_1 + 0.20\,S_2 + 0.20\,S_3 + 0.15\,S_4 + 0.15\,S_5\right)$$

The calculation is a transparent policy index over ordinal ratings. It is not a probability that the output is true, a percentage of legal compliance, or a validated interval scale. Report the five scores, evidence, and disposition alongside the composite. Decimal precision comes from arithmetic and does not imply measurement precision.

Unknown is not zero

Use unknown when required assessment evidence is insufficient or a material disagreement remains unresolved. A verified failure receives the applicable score or failed gate. An unassessed vector receives no numeric value. If any vector is unknown, suppress the composite; do not replace the missing value with zero, assume a pass, or redistribute the weight.

A confirmed gate failure can still block a transaction with other unknown findings. Retain both the failure and the information gaps. The absence of a numeric composite does not weaken a documented reason to stop release.

Why the gate is reported separately

Multiplying the score by zero discards information about the other controls. The ledger instead reports a result such as EIS 32.5, BLOCKED, with provenance and coverage failures identified. Remediation can then be evaluated against the original score and findings without suggesting that the original output was safe to use.

11 Apply gates before considering release

Each required gate is pass, fail, or unknown and must have a rationale. Pass and fail findings require evidence identifiers. A pass means the specified test was completed within scope; it is not a general assurance about the system. Unknown prevents a release recommendation.

Gate	Failure condition
PROVENANCE	A decision-critical claim is unsupported, contradicted, misattributed, or strengthened by material qualification loss.
COVERAGE	Required evidence or a material conflict is unresolved for the proposed action, or a negative finding exceeds the verified search.
REVIEW	A required authorized review or decision-specific verification is demonstrably absent, bypassed, or invalid.
LINEAGE	Essential evidence of the material decision state is demonstrably missing or untrustworthy.
PROFILE	A specific predeclared domain, safety, or policy prohibition is violated. The assessed requirements and their authority must be identified.

Evaluate the following rules in order. A known score of zero, or V1 below 2, is also a baseline blocking condition even if a gate was incorrectly recorded as passing.

Disposition	Rule and permitted next step
BLOCKED	Any gate fails, any known vector is 0, or known V1 is below 2. Stop the affected release and investigate; retain unknown findings.
HOLD	No blocking finding, but a gate or vector is unknown. Obtain evidence or adjudication before reassessment.
REMEDIATE	All gates pass and scores are known, but EIS is below 70 or any vector is below 2. Correct and reassess.
SUPERVISED CANDIDATE	All gates pass, EIS is at least 70, and every vector is at least 2; the stricter rule below is not met. Route for designated enhanced review.
STANDARD REVIEW CANDIDATE	All gates pass, EIS is at least 85, and every vector is at least 3. Route through the separately authorized review and release process.

A score of 90 with one vector at 2 remains a supervised candidate. If a rating establishes a gate failure, the assessor must record that gate as failed. The scorer cannot infer evidence content; its score floors withhold candidate routing even when entries are inconsistent. No score authorizes deployment or replaces the responsible release authority.

PROFILE is limited to its predeclared requirements. Correct inconsistent score and gate entries before relying on the result. The below-2 remediation branch is a defensive stop for supplied assessments; a confirmed material gate failure requires BLOCKED. Overlapping findings are not independent incidents.

12 Illustrative acquisition evidence packet

Instructional fixture RL-ILL-001. Every record, excerpt, event, and outcome in this case is constructed for demonstration. No model run, acquisition, or client audit is represented as having occurred. The assumed transaction concerns a $45 million facility acquisition and a proposed memo recommending unconditional approval.

The profile requires current evidence for utility access, flood exposure, and litigation review before unconditional approval. The corpus contains the five records below. For this exercise, D-03 explicitly supersedes the flood finding in D-02 for the same parcel. D-04 is a docket attachment, not proof of a particular legal outcome.

ID and state	Constructed source excerpt or record
D-01 Received and extracted	Parcel B utility access is provisional. The easement expires on 31 December 2026 unless renewed in writing. No renewal is recorded in this packet.
D-02 2022 baseline report	Parcel B is classified as Low-to-Moderate flood exposure for this assessment.
D-03 2025 addendum	This addendum replaces the 2022 flood classification for Parcel B. Updated mapping places Parcel B in the High Flood Risk category.
D-04 Scanned docket attachment	A civil proceeding affecting the recorded owner appears on the attached docket. The current status and effect on acquisition require legal review.
D-05 Workflow and review events	D-04 received: 8 pages; extraction result: 0 characters; no extraction alert raised. Retrieval returned D-01 and D-02. D-03 was excluded by a stale metadata filter. An approval event was recorded with no required exception-resolution fields completed.

Output presented for approval

> *Parcel B carries full commercial utility clearance. Flood exposure is Low-to-Moderate. No active civil litigation or encumbrances were found across county filings. The acquisition may proceed without further diligence conditions.*

The available lineage records preserve the selected passages, the generated text, the review event, and the action request. The fixture assumes that no additional review record exists. A fourteen-second approval interval may be recorded as context, but the review finding rests on the missing required verification, not on elapsed time.

No 94/100 judge result is reported. A format-focused score would not establish support for these claims, but this fixture does not simulate or measure a competing evaluator's performance.

13 Findings and disposition for the fixture

Finding	Observed defect and decision impact
F-01 Qualification loss	D-01's provisional status, expiry, and renewal condition are absent from full commercial utility clearance. The unconditional approval premise is unsupported. O-01 remains violated.
F-02 Supersession omission	D-03 is available in the declared corpus but excluded from retrieval. The memo presents a superseded classification as current. O-02 requires the applicable addendum to govern the flood finding.
F-03 Unsupported negative	The system did not process D-04, yet asserted a negative across county filings without an adequate search record. O-03 requires recovery and legal review or an explicit restriction on action.
F-04 Review deficiency	An approval event exists, but the required decision-specific verification and exception dispositions do not. The evidence supports a control failure; it does not establish the reviewer's psychological state.

Vector assessment

Vector	Score	Reason
V1	1	Identifiable sources do not support the material clearance and negative assertions.
V2	1	A material provisional condition is removed without resolving evidence.
V3	1	An applicable addendum and required docket review are omitted.
V4	1	Approval is recorded without the required verification basis.
V5	3	Records reconstruct the material decision state; independent reconstruction testing is not asserted.

$$EIS = 25(0.30 + 0.20 + 0.20 + 0.15 + 0.45) = 32.5$$

PROVENANCE, COVERAGE, and REVIEW fail. PROFILE also fails because the predeclared unconditional-approval rule requires resolution of these diligence conditions. LINEAGE passes. V1 below 2 independently requires BLOCKED. The raw EIS remains 32.5.

The affected memo and its release are blocked. The incident owner preserves the original record, checks whether any dependent action occurred, and identifies the affected scope. The finding does not automatically justify shutting down every unrelated workflow using the model.

The earlier draft's 36.25 was an arithmetic error. These same five scores and weights produce 32.5. The retained diagnostic score and separate disposition are the revised method's reporting convention.

14 Repair the evidence path and reassess the action

A repair must address both the initiating defect and the later controls that allowed it to reach the decision. Assign each action an owner, acceptance evidence, and reassessment identifier. Promising a future fix does not close an existing finding.

Owner	Required correction and acceptance evidence
Ingestion owner	Recover D-04, reconcile all 8 pages, verify the relevant extracted content against the image, and retain the exception and recovery record.
Retrieval owner	Repair the filter that excluded D-03; demonstrate that the applicable addendum reaches the evidence set for the same parcel.
Synthesis owner	Preserve provisional utility access and expiry; replace the stale flood classification; remove the unsupported universal negative.
Domain reviewer	Assess the recovered docket, record the significance and remaining limits, and identify the action the evidence permits.
Release owner	Require evidence-linked resolution or an explicit action restriction for each material obligation; preserve the original failure and new assessment.

A corrected decision statement

> *Parcel B's utility access remains provisional and expires on 31 December 2026 unless renewed in writing. The 2025 addendum supersedes the 2022 flood classification and identifies High Flood Risk. The recovered docket requires legal assessment of its current status and acquisition implications. This memo supports continued diligence and escalation; it does not support unconditional acquisition approval.*

The corrected memo is useful because it states what can be concluded and what action remains unavailable. The audit may pass the evidence gates for a restricted informational release if the profile permits that use and the required review is complete. It must not relabel the same unresolved evidence as sufficient for unconditional approval.

Illustrative reassessment

Assume a subsequent evidence packet establishes that all five vectors meet score 3, all gates pass for the revised informational action, and the owner has recorded the changed decision boundary. The resulting EIS is 75, with disposition SUPERVISED CANDIDATE. This is an assumed worked outcome, not a measured improvement or release authorization.

The included case_after.json encodes that assumption. Its evidence identifiers are example references. An actual reassessment must inspect the repaired records and review evidence before assigning those values. A stronger score would require the additional independent checks specified in the rubric.

Changing the permitted action is recorded explicitly. The original acquisition recommendation remains blocked until its own evidence requirements are met.

15 Operate the audit and implement the record

A repeatable transaction procedure

1. Freeze the profile and evidence boundary before scoring. Record the proposed action, source universe, material questions, reviewers, and permitted uses.
2. Capture the source manifest and actual workflow states. Reconcile required inputs and extraction exceptions before evaluating generated claims.
3. Identify material claims and create evidence obligations. Check source support, applicability, conflicts, and negative assertions.
4. Compare the relevant handoffs and inspect the human decision. Record defects and unresolved questions with evidence-linked reasons.
5. Assign vector scores and gate states. Adjudicate required disagreements and calculate the disposition using the approved profile.
6. Route the outcome to the authorized owner. Preserve the original record, verify remediation, and issue a separate reassessment where needed.

Separate the evidence record from the scorer

The full ledger contains source objects, claim records, obligation transitions, review findings, and decision dependencies. The reference scorer consumes a compact assessment summary: transaction and assessment identifiers, profile version, audit boundary, five vector ratings, and five gate determinations. Each known rating or gate result includes a rationale and evidence identifiers.

The scorer validates structure, calculates EIS, preserves blockers and unknowns, and applies the release-routing rules. It cannot verify that a cited object exists, that a reviewer read it, or that the rating is justified. That work belongs to the evidence store, assessors, and controls outside the calculator.

Minimum audit output

Every report states the assessed action and boundary, unresolved limitations, five scores, gate findings, EIS or why it is unavailable, disposition, affected decisions, remediation owners, and the authorizing role for the next step. Include profile and assessment versions so different rules are not compared as though they were the same.

Start in shadow mode: audit completed or staged transactions without allowing the experimental score to control consequential release. Establish a reliable assessment process before connecting a gate to production. Once authorized, the release control should prevent bypass and log alternative actions without concealing failed findings.

Implementation files: scorer.py, audit_input.schema.json, test_scorer.py, case_before.json, and case_after.json. The README explains the command line and the distinction between deterministic rule tests and real-world validation.

16 Validate the method before relying on its thresholds

A convincing evaluation must test whether the method finds decision-relevant failures that a capable existing process misses, at an acceptable review cost. It must also measure incorrect blocks, missed failures, and assessor disagreement. A successful arithmetic test does not establish that the audit improves decisions.

Study design

Choose one defined workflow and register its materiality rules, source requirements, rating anchors, and success criteria. Separate development cases from a held-out evaluation set. Use naturally occurring examples where permitted, together with controlled fault cases whose construction is documented. Report the two groups separately.

Include omitted qualifiers, incorrect source attribution, silent extraction loss, conflicting versions, bounded negative searches, and invalid review records. Also include harmless paraphrases, legitimate inferences, resolved conditions, correct use of older operative records, and careful reviews completed quickly. Those controls test whether the method overreacts to superficial signals.

Have qualified assessors label the decision-relevant defects and permitted actions independently of the method's results. Preserve initial disagreement and adjudication. Compare the same held-out transactions against existing review, a source-grounded claim/citation evaluator, and an appropriate RAG evaluation baseline. A coherence-only judge is an insufficient main comparator. [1–4]

Measure	Required interpretation
Critical-failure recall	Share of independently labeled critical failures detected; report counts and uncertainty.
False-block rate	Share of adequately supported permitted actions incorrectly blocked; inspect the causes.
Incremental findings	Defects detected beyond each baseline, with evidence of decision relevance.
Review effectiveness	Correct exception handling and final actions, not just clicks or review duration.
Assessment agreement	Gate agreement and ordinal-rating agreement before adjudication, with disputed categories.
Operational cost	Reviewer effort, latency, rework, and burden by case type and consequence.

Analyze errors by transaction as well as by claim. Claims within a document and transactions from the same matter may be correlated. Keep triggered incident samples separate from a representative sample when estimating prevalence. Report the sampling frame, exclusions, sample size, and denominator.

Test sensitivity to plausible weights, rating uncertainty, and thresholds. A transaction whose disposition changes under a reasonable one-point disagreement needs examination. Revise the profile transparently, then evaluate on a fresh holdout set. Publish only the conclusions supported by the evaluated workflow and sample.

17 Governance crosswalk and scope limits

This crosswalk identifies related control objectives. It does not establish compliance, certification, or endorsement. Applicability depends on the system, actor role, jurisdiction, effective requirements, and the organization's complete control environment.

Reference	Contribution and remaining limits
NIST AI RMF 1.0	Evidence records can support MEASURE 2.1, 2.3, 2.8, and 2.9; routing can inform MANAGE 1.1 and 2.4. The voluntary framework covers substantially more than this method. [8]
ISO/IEC 42001:2023	The ledger may supply records for an organizational AI management system. This is a conceptual relationship, not a clause-level conformity assessment or certification. [9]
EU AI Act Articles 9 and 12	Risk findings and retained decision records may inform relevant risk-management and record-keeping work for an applicable high-risk system. [10]
EU AI Act Articles 13 and 14	Preserved limitations and evidence of authorized review may inform transparency and human oversight. Click counts do not establish compliance. [10]
EU AI Act Article 15	Accuracy-related findings can contribute evidence. The method does not replace robustness, cybersecurity, or system-level accuracy testing. [10]

Current banking references require care

SR 26-2, issued on 17 April 2026, superseded SR 11-7. Its attachment expressly excludes generative and agentic AI models from its scope while recognizing the role of broader bank governance for uncovered tools. The ledger therefore does not claim direct compliance with that guidance for an LLM workflow. [11]

CFPB Circular 2023-03 was included in the Bureau's May 2025 guidance withdrawal. It is not used here as a current operative authority. Withdrawal of guidance is distinct from repeal of underlying statutes or regulations; any financial-services profile needs a separate current applicability assessment. [12]

What a successful ledger still cannot establish

It cannot guarantee that the source universe contains every relevant fact, that an authoritative source is correct, or that a human reviewer understood every implication. It cannot turn an ordinal score into a probability of safety. It does not replace professional judgment or the domain's substantive requirements.

Its practical test is narrower and demanding: can an authorized reviewer reconstruct the decision, identify its material premises and limits, inspect how they changed, and see why the action was permitted or withheld? The method succeeds only to the extent that evidence and validation support that answer.

References

Sources checked 3 November 2026. Bibliographic references support the background and crosswalk. The proposed rules, weights, and examples are this specification's design choices, not requirements imported from these sources.

[1] Min, S., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP. https://aclanthology.org/2023.emnlp-main.741/

[2] Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. EACL System Demonstrations. https://aclanthology.org/2024.eacl-demo.16/

[3] Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. EMNLP. https://aclanthology.org/2023.emnlp-main.398/

[4] Murugaraj, K., Lamsiyah, S., and Theobald, M. (2026). RAGVUE: A Diagnostic View for Explainable and Automated Evaluation of Retrieval-Augmented Generation. EACL System Demonstrations. https://aclanthology.org/2026.eacl-demo.35/

[5] Peters, U., and Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. Royal Society Open Science 12, 241776. https://doi.org/10.1098/rsos.241776 ; author manuscript: https://arxiv.org/abs/2504.00025

[6] Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1. https://doi.org/10.1145/3449287 ; https://arxiv.org/abs/2102.09692

[7] W3C (2013). PROV-Overview: An Overview of the PROV Family of Documents. https://www.w3.org/TR/prov-overview/

[8] NIST (2023). Artificial Intelligence Risk Management Framework 1.0. NIST AI 100-1. Relevant tables: MEASURE, printed pp. 29–30; MANAGE, printed p. 32. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[9] ISO (2023). ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system. Public scope description only; the full licensed standard was not used for a conformity analysis. https://www.iso.org/standard/42001

[10] European Union (2024, as amended). Regulation (EU) 2024/1689, Articles 9, 12, 13, 14, and 15. Official legal text: https://eur-lex.europa.eu/eli/reg/2024/1689 ; Commission article texts: https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-9 (substitute the relevant article number). This crosswalk does not assess classification, exceptions, or commencement for a particular deployment.

[11] Federal Reserve (2026). SR 26-2, Revised Guidance on Model Risk Management, 17 April 2026. Cover letter and attachment, especially attachment p. 3, footnote 3. https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm ; https://www.federalreserve.gov/supervisionreg/srletters/SR2602a1.pdf

[12] CFPB (2025). Withdrawal of Certain Policy Statements, Interpretive Rules, and Other Guidance. 90 Federal Register 20084, 12 May 2025; Circular 2023-03 listed at 20086. https://www.govinfo.gov/content/pkg/FR-2025-05-12/pdf/2025-08286.pdf

Version and development record

Version 1.0 expands the supplied three-page concept into a complete method. It adds evidence obligations, handoff comparisons, bounded negatives, full anchors, unknown states, independent gate routing, a reproducible illustrative case, and validation requirements. It corrects the example arithmetic and removes unsupported novelty, deployment, and regulatory claims.
No registration, public publication, or external certification is represented.