Maintaining Supportable Credit Conclusions

Evidence sufficiency, bounded decisions and consistency across financial systems

EOJI Field Note v0.7 | 10 October 2026

This note presents a framework and evidence synthesis, not a report of completed EOJI or competitor experiments. Figure 1 is a version-scoped visualization of reported external-study results; Figure 2 is an explicitly synthetic calculation; Figure 3 is a documentary illustration rather than gold reference truth and remains NOT_GOLD. It does not establish legal compliance, borrower status, professional credit judgment or professional-user benefit, and it does not report an EOJI product benchmark. Comparative longitudinal testing is reserved for the extended research paper.

Suppose a credit memo contains a genuine quotation, a correct calculation and a reviewer's approval. Then someone notices that the quotation comes from the earlier agreement, the calculation uses an adjustment the amended definition excludes, and the approval belongs to a version that nobody is currently reading.

Nothing needed to be invented. The system could get into trouble by arranging authentic things incorrectly.

That is the problem with treating citations as the finish line for financial AI. A citation can help establish where a statement came from. It does not, by itself, establish that the statement answers this question, under these terms, at this time, for this use. Nor does putting a person between the answer and the action automatically settle those matters.

The objective is not a machine that always supplies a number, or one that earns a spotless record by declining to do any work. It is an analysis whose conclusion is proportionate to the evidence, and a process that preserves that relationship when something changes.

1. Specify the conclusion before asking whether it is supported

A leverage ratio, satisfaction of a numerical covenant, contractual compliance, permission to draw and a favorable investment recommendation are different conclusions. They are not successive confidence levels for one conclusion.

A borrower can satisfy a numerical test and still face a refinancing problem. A waiver can change contractual consequences without improving cash generation. A reviewed calculation might inform an approval without constituting that approval. A credit memo that compresses all of this into a green indicator has made the interface simpler. Whether it has made the decision simpler is another question.

Start with the entity, facility, period, operative interpretation and information cutoff. Keep reported facts, calculations, estimates, scenarios and authorized judgments distinct. If a statement depends on a particular interpretation, carry that condition into the answer rather than leave it behind in the working papers.

Dyreng, Ferracuti, Hills and Kubic compare aggregate apparent covenant-violation rates calculated from issuer-reported thresholds and realizations with rates estimated from commercial data and standardized definitions.[1] In the retained November 2024 preprint, the common Table 3 Panel B population contains 14,623 covenant observations. Using issuer-reported values for both components produces a reported 1.08% rate. Replacing only the

threshold with its proxy raises the aggregate rate to 6.45%; replacing only the realization raises it to 16.80%; proxying both produces 19.52%. The table does not provide an observation-level transition matrix, so these differences do not count individual classifications that changed.

The paper calls issuer-reported contractual quantities “true.” That is an operational study definition, not independent legal adjudication. The sample selects firms that disclose the relevant quantities, and the retained 2024 version is not the inaccessible 2026 journal PDF. Within those limits, the decomposition makes a material point: obtaining the agreement or current threshold is necessary but can remain insufficient when the contractual realization is reconstructed from generic accounting data.

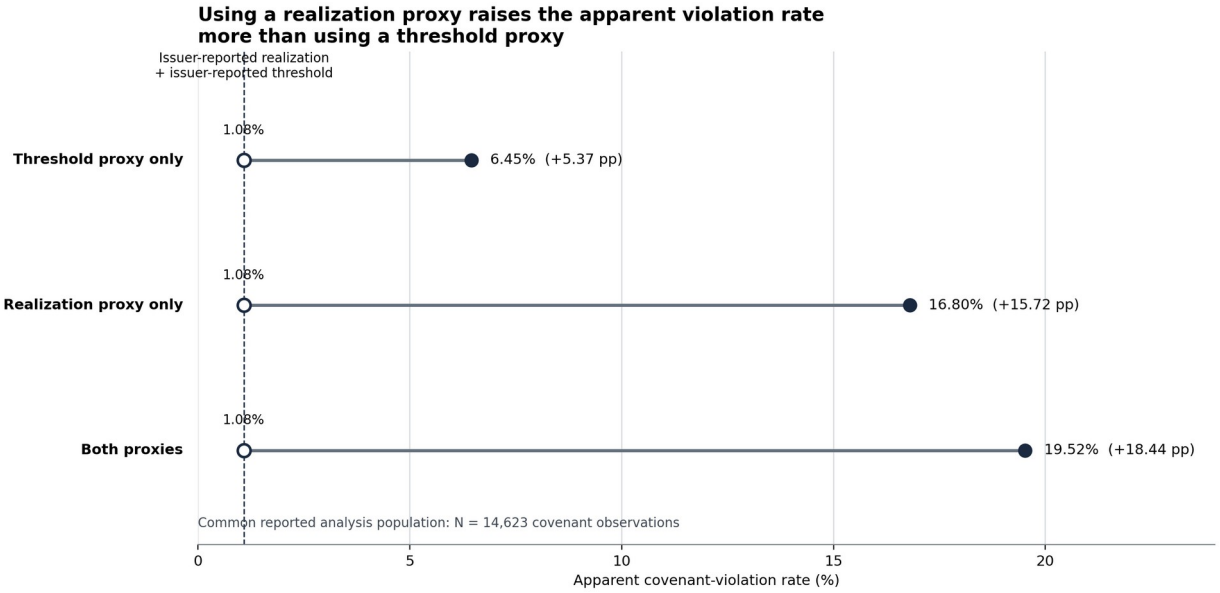


Figure 1. Using a realization proxy raises the apparent violation rate more than using a threshold proxy. Reported empirical aggregate rates for a common population of 14,623 covenant observations in Dyreng et al.’s retained November 2024 preprint, Table 3 Panel B. The source reports pairwise aggregate rates, not individual classification transitions. “Issuer-reported” is not independent legal ground truth; the selected sample is not a market-wide prevalence estimate, and this is not an EOJI benchmark.[1]

Better extraction cannot retrieve an omitted contract-specific input. Better provenance cannot make a standardized proxy contractually exact. “Not established here” is also not the same as “the borrower never supplied it”: a public filing, a lender reporting package and a custodian’s records are different information sets.

2. A missing number is not always a missing answer

Consider a deliberately simplified facility. Evidenced components of eligible EBITDA total €10 million. A further eligible adjustment, A, is not known exactly, but separate admissible evidence establishes $0 \leq A \leq 3$ million. Debt is known. The reviewed numerical test is $D/(10+A) \leq 5$.

Assume the definition and period are settled, the denominator is positive, the endpoints are included, and there are no other adjustments, caps, netting provisions or disputes.

Debt D	Possible ratio, approximately	Conclusion about this numerical test
€45 million	3.462 to 4.500	Satisfied throughout the evidenced range.
€60 million	4.615 to 6.000	Unresolved; satisfied precisely when $A \geq 2$.
€70 million	5.385 to 7.000	Exceeded throughout the evidenced range.

All three cases lack an exact ratio. Only the middle case lacks a determinate threshold conclusion. Figure 2 separates exact possible states from the status of an evidence interval. For D=60, an interval wholly above €2 million supports pass, one wholly below €2 million supports fail, and one crossing €2 million remains unresolved. No probability distribution is needed or implied.

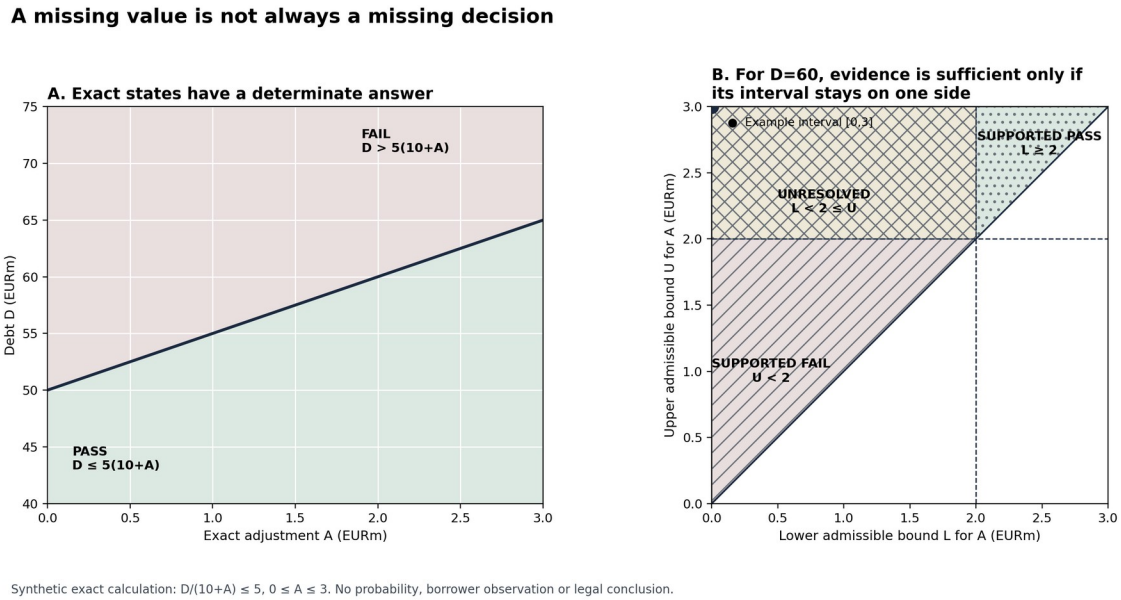


Figure 2. A missing value need not leave the decision unidentified. Transparent synthetic calculation for $D/(10+A) \leq 5$, with debt and EBITDA components in EUR million and $0 \leq A \leq 3$. The left panel shows exact states; the right shows admissible lower/upper bounds for A when D=60. This is not an observed facility, probability model, legal-compliance conclusion or product result.

For the middle case, the useful answer is not an invented estimate and not a blank screen. It is: “Under this interpretation, the numerical test is met if Adjustment A is at least €2 million. The package does not establish which side of that boundary applies.” That identifies what matters next. It does not declare overall compliance or authorize a transaction.

The assumptions are doing real work. A contractual cap can establish a permitted maximum without showing that the adjustment occurred. A management estimate is not an observation. Several uncertain inputs may share constraints, so their individual extremes cannot necessarily occur together. Inconsistent constraints leave no feasible answer. A zero denominator requires explicit treatment.

These are elementary bounds calculations, not a new EOJI algorithm. A spreadsheet, rules engine or conventional checklist can implement the same numerical criterion. The useful distinction is between a missing value and a missing decision.

3. Dependencies are useful only when their meaning is right

Imagine the same adjustment appearing in a management presentation, a lender spreadsheet and an AI-generated memo. Three files may contain one claim. Alternatively, another organization may have independently checked the amount and reached the same answer. Repeated wording, shared origin and independent verification are different things.

The cheapest useful safeguard may be less exotic than it sounds: group claims by upstream evidence before counting corroboration. A probabilistic dependence model can be valuable, but it should first beat a dated source-family table. Complexity has to earn its chair at the meeting too.

Enke and Zimmermann found correlation neglect in laboratory experiments and reduced it with an assigned package that made correlation and independent signals more salient.[11] That does not establish that a lineage display improves professional credit review. It does rule out presenting source double-counting as a recent financial-AI discovery.

Maintaining justifications also has a long history. Doyle records reasons for beliefs; Cała and Missier study provenance-based recomputation; ExplicitCase distinguishes potential change impact from effects requiring substantive revision.[2][3][4] Credit software should build on these foundations rather than rename them and claim to have invented memory.

The hard part is getting the relationships right. A calculation input, a cross-reference and a reviewer's authority are not interchangeable edges. Two passages mentioning the same borrower need not concern the same facility. Losing one source need not invalidate a conclusion with other sufficient support; keeping several related passages does not compensate for omitting the operative exception.

For the same deterministic specification and complete inputs, correct selective recomputation and correct full recomputation should agree. Incrementality earns its complexity through savings and controlled updates, not superior arithmetic. The cost of creating and maintaining the relationships belongs in that comparison.

4. Financial time and system time answer different questions

Suppose an amendment is signed on Monday, states an earlier effective date and arrives on Wednesday. Current knowledge about Monday is now different from the information available for Monday's actual review. Both histories matter.

Lamport formalizes causal ordering and compatible logical clocks.[5] An earlier scalar clock value does not, on its own, prove causal precedence. Nor can a message's arrival settle whether

contractual conditions for effectiveness were satisfied. The timestamp is not a substitute for the agreement.

Sorting by the latest upload can select the wrong rule. Sorting by the stated effective date can give an earlier reviewer knowledge they did not possess. Preserve the reporting period, claimed applicability, evidence availability and later corrections without pretending they are one date.

The public Marriott revolving-credit history illustrates why this separation matters. The April 2020 amendment waived a quarterly leverage test for a defined period, introduced a monthly liquidity floor and tightened other terms. Later draw, repayment and balance disclosures altered exposure evidence without independently answering compliance. January 2021 amendments extended relief, raised the liquidity floor, reset re-entry thresholds and separately changed rate floors. “Waiver” was never a global state in which every obligation disappeared.

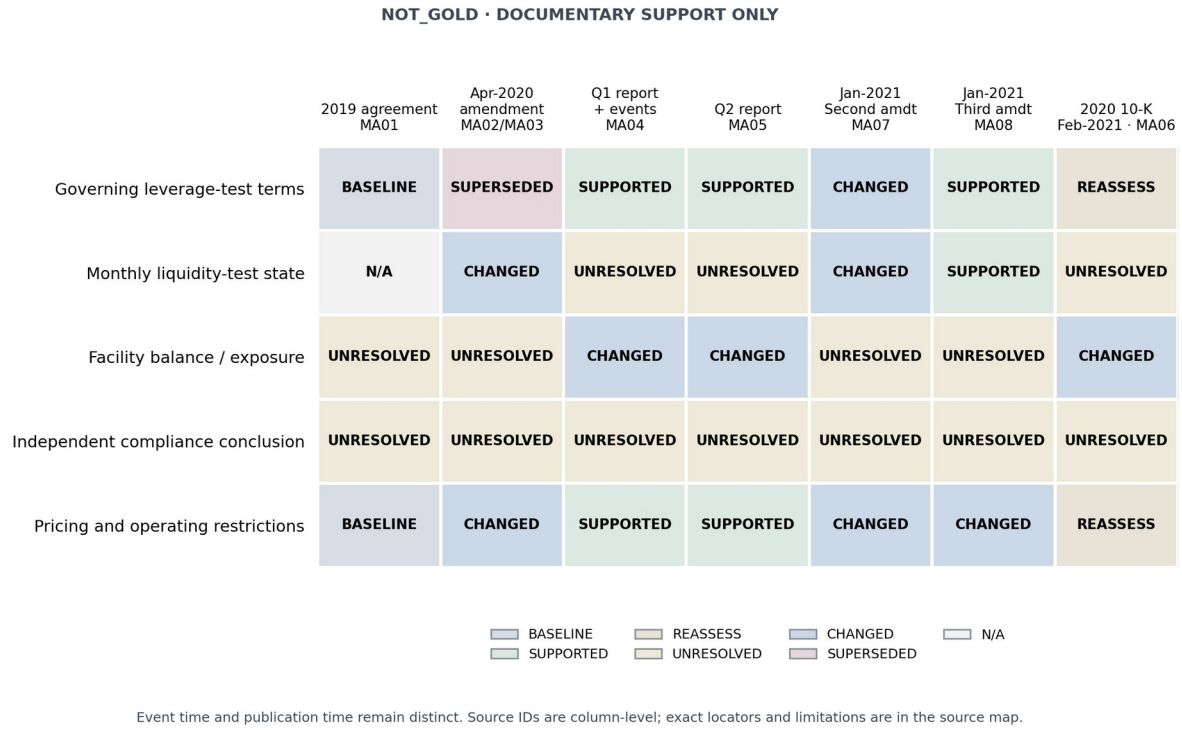


Figure 3. Selective documentary state change in one public Marriott facility history. Five claim rows are coordinator-coded across seven agreement, amendment and filing events from official SEC filing records reviewed through the 2026-10-08 cutoff. Baseline, supported, changed, reassessment and unresolved labels describe documentary support only and remain NOT_GOLD. The matrix does not establish legal compliance, default, remedies, draw authority, current exposure or credit merit; fresh SEC raw-byte capture was blocked and qualified adjudication remains unavailable.

The matrix is not evidence that a graph architecture is required. A disciplined covenant calendar, dated checklist and full rereview can produce the same observable state separation. A selective-reassessment system must demonstrate that it preserves this result while reducing work, not assume savings from a tidy diagram.

An annual report may contain a subsequent event. Supplying that event to a system answering an earlier question tests hindsight reconstruction, not contemporaneous judgment. The result can be impressively accurate because the answer arrived in the input packet.

5. Approval has to survive concurrency

Suppose a reviewer approves calculation version 12. While a background memo task is still running, an amendment creates version 13. The worker eventually returns a completely faithful account of version 12.

That is a successful computation. It is not necessarily the current approved conclusion.

Approval must remain attached to what was reviewed: the evidence, interpretation, calculation and intended use. An unchanged number need not imply unchanged authority, although an explicit institutional policy may permit some updates without renewed review. A changed number need not imply an adverse credit event either.

Compatibility matters as well as freshness. A debt balance, collateral value, eligibility rule and approval can each be authentic and recent yet belong to incompatible revisions. The assembled answer needs a justified common context, not merely plausible components.

Evidence of €10 million of available capacity can be read repeatedly. It cannot be independently consumed by two €7 million requests without accounting for both. A read-only memo need not become a settlement engine, but it must not describe a snapshot of headroom as a reservation or unconditional permission to spend it.

These are requirements, not disclosures of a synchronization algorithm. Not every update needs a global lock. Late work can remain accessible as history without acquiring current authority.

6. Sending a correction is not the same as enforcing it

A withdrawal has been accepted by one service. Somewhere else, a consumer still holds yesterday's assertion. The sender has a successful log entry. The consumer has a different state. Neither fact makes the disagreement disappear.

A timeout creates another ambiguity: did the remote action fail, or did it succeed while the acknowledgment was lost? AWS's discussion of idempotent APIs explains why careless retries can duplicate effects and why intent matters.[6] Repeating a request safely is not the same as establishing that it was authorized.

The interface needs explicit behavior for duplicates, stale versions, lost acknowledgments and unavailable status checks. "Current," "historical" and "could not establish current status" must remain different answers. A disconnected consumer cannot be promised instantaneous correction.

Measure receipt, impact detection, update, publication and consumer enforcement separately. An acknowledged notification is not proof that the revised policy was applied. Repeated messages must not manufacture approvals or revive old ones.

7. Integrity and performance do not establish financial truth

W3C's credential model distinguishes verification from the truth of encoded claims.[7] A cryptographically protected record can faithfully preserve an incorrect interpretation. A valid signature can belong to someone whose authority does not cover the intended action.

Blockchain publication can distribute a reviewed state without supplying an undisclosed accounting adjustment or deciding what the loan agreement means. Proving a calculation over a stated dataset is different from proving that every relevant obligation is in that dataset.

Performance deserves the same precision. More workers do not remove serial dependencies. Faster parsing can move the bottleneck to inference, storage or human review. Accepted requests are not completed analyses; a growing queue is not evidence of useful capacity.

A meaningful result specifies workload, hardware, memory limits, concurrency, failures and tail latency. It counts correct completions and tests recovery. A system that avoids publishing an error by never finishing anything has solved a different problem.

8. Evaluate the decisions, including the inconvenient ones

A demonstration in which the machine identifies one missing variable is useful. A demonstration in which the control is encouraged to guess while the product is told to be careful is useful mainly to the demonstrator.

Freeze the assigned unit before scoring: one facility, one test period, one decision cutoff and one requested conclusion. Report what entered the packet, what was admitted, what could be computed, what the reference settled, what the reviewer completed and whether the material decision was correct. Documents, passages and repeated configurations are not extra independent credit decisions merely because the spreadsheet has more rows.

The answer key also needs scrutiny. Distinguish a question the evidence settles, one it genuinely leaves unresolved, and one for which evaluators have not established an adequate reference. The last is our measurement problem, not the borrower's condition. An explicitly undefined ratio may still have a clear reference answer: the requested operation is undefined.

Correct output does not guarantee correct use. Buçinca and colleagues found that cognitive forcing reduced AI overreliance while effective designs received less favorable subjective ratings.[12] That was not a professional credit-review experiment. It is a reason to measure the tradeoff.

A useful candidate comparison would add explicit lineage to source-linked review, then compare a general reminder with a decision-specific explanation of what could change the conclusion. Randomized assignment can estimate the effect of that assistance. It does not automatically establish whether any benefit arose from understanding, attention, navigation or perceived credibility.[17] If simple source grouping and a reminder work as well, a more elaborate backend has not earned its expense.

Keep the errors in the denominator

For questions the evidence can settle, report supported-correct yield: correct, sufficiently supported conclusions divided by all assigned decidable targets. Alongside it, report materially wrong definitive answers, unsupported reasoning, unnecessary deferrals and non-completion. Answer coverage alone measures willingness to answer, not correctness.

For genuinely unresolved questions, count unjustified unconditional conclusions. Also assess useful qualifications or appropriate escalation. An empty response is not automatically a successful refusal, and a demand for unattainable evidence is not necessarily a useful next step.

For updates, count correct revisions, inappropriate use of superseded material and unnecessary changes to unaffected conclusions. Missing data, a known timeout and participant withdrawal are different events. Report them and test sensitivity to missing-outcome assumptions instead of deleting inconvenient observations.

Count the work, not the mouse movements

Measure decision quality, operational labor and perceived workload separately. Reading, thinking, rework, handoffs and necessary second review are work. Two people working for ten minutes use twenty person-minutes even if the calendar advances by ten.

NASA-TLX is one established option for subjective workload.[13] Select and disclose the version, scoring and time period. Feeling successful is not the same outcome as calculating the right result.

Ratios can conceal a tradeoff. In the prior hypothetical example, 100 assigned reviews used 1,000 minutes for 80 acceptable outcomes in one condition and 800 minutes for 70 in another. Minutes per acceptable outcome improved from 12.50 to 11.43 while acceptable yield fell from 80% to 70%. A sentence is more efficient than a chart for that arithmetic. The main-text figure is therefore removed. A genuine effort/supported-correctness frontier remains reserved until professional person-minutes, fixed denominators, material-error outcomes and statistical uncertainty actually exist.

A nonsignificant error difference is not proof of unchanged safety.[18] Research questionnaires are not product labor, but operational preparation is not free merely because an expert performed it before the session.

Do not grade only the people who understood the product

A quiz between initial review and amendment may teach participants what to inspect next. Research identifies reactivity risks, while other experiments found little evidence that factual-check placement changed treatment effects.[14][15] Pilot the actual questions rather than adopt an absolute rule.

The primary comparison should not exclude participants because they failed a post-treatment comprehension check. That selects on something the interface may itself have changed.[16] A late correct answer demonstrates understanding at that later point, not direct access to the reviewer's earlier mental state or a unique causal mechanism.[17]

The test is supposed to discover whether people can use the intervention. Removing people who could not use it would make the product easier to validate, but less relevant to buy.

9. Public evidence helps scrutiny, not automatic impartiality

Public data, public scoring and independent evaluation are different properties. Private evidence can be essential to a realistic task. Public data can still be selectively chosen, incorrectly labeled or familiar to the model.

Existing vendor disclosures illustrate the variety. Hebbia describes expert-informed synthetic cases without actual user data. ExactCov describes agreement-level testing and cited-field scoring while withholding model identities. Rogo publishes a BigFinanceBench subset and

reference harness.[8][9][10] These are descriptions of evaluation arrangements, not independently established rankings.

EOJI's proposed public-source work would expose task definitions, evidence cutoffs, source versions, annotation rules, scoring and corrections. A separately administered holdout and confidential client validation serve different purposes. A declared sample does not become representative merely by being declared.

Keep three questions separate: whether an intervention helps people, whether software can generate it from raw documents, and whether the complete service is worth its cost. Expert-prepared assistance can test the first without proving the second. A customer comparison can test a bundle without revealing which algorithm caused any benefit.

Sponsorship belongs in the record. EOJI has an interest in favorable findings. Reference reviewers should be separate from implementation, conflicts disclosed, and a justified challenge able to reduce EOJI's score. A public benchmark needs an answer key that can lose an argument.

What remains to be measured

Public legal datasets reviewed for this note test clause extraction, document-level entailment, provision typing or generated multi-clause compliance. None supplies the complete longitudinal credit-state task: versioned facility histories with event, publication and operative time; labeled reassessment sets; preserved ambiguity; unchanged controls; and independently reviewed change outcomes.

A benchmark should be inspectable

First-party evaluation is useful and often necessary. It is not independent evidence. Under official records inspected through 10 October 2026, CovenantAI reports a dataset and comparisons with earlier covenant methods, but no reusable public task-and-scoring package was established; that limits replay without erasing the prior art.[19] Hebbia explains expert-informed synthetic evaluation and model-judge controls, but no public finance task set, answer key or scoring code was established in the official locations inspected.[8][20] ExactCov reports a 60-agreement, 1,840-field test, while the official materials inspected did not expose the source agreements, field labels or a frozen model-level comparison; model names were explicitly withheld.[9]

Rogo is material counterevidence. BigFinanceBench releases a licensed 50-question subset with answers, rubrics and a harness while holding back 878 questions to reduce contamination.[10][21] Another commercial party replayed that subset and challenged some answers, demonstrating inspectability, although this is not neutral validation.[22]

"We wrote the exam, kept the answer key private, graded ourselves and passed" is a perfectly respectable internal QA process. It is not a public benchmark, and software sold into regulated financial workflows should not receive a looser evidentiary standard than the credit decisions it helps produce.

Confidentiality and contamination can justify restricted cases. Credible substitutes include a public subset, frozen protocol, scoring code, versioned outputs and controlled third-party replay. EOJI must meet the same standard: release the largest lawful subset, disclose

denominators and failures, preserve negative results and let a simple baseline win. This Field Note reports no EOJI product benchmark or superiority result.

A 40-case feasibility pilot therefore tested the evaluation schema rather than a system. It comprised 35 Marriott documentary cells that remain NOT_GOLD and five separately marked synthetic stress cases; no EOJI implementation ran. With affected-source mappings already supplied, a dated-rules and source-grouping baseline admitted 33 cases for review, compared with 40 under full recomputation. That difference is workload arithmetic, not accuracy or professional benefit. It shows that sophisticated architecture is not necessary in a structured setting and that future comparisons must charge every method for discovering and maintaining the mappings from equal raw evidence.

The extended study will compare full recomputation, dated rules, source-linked retrieval and a reproducible maintained-state procedure on independently reviewed facility histories. This note makes no product-superiority claim from the feasibility work.

10. A proportionate next step

The immediate proposal is a bounded comparison using authorized historical material, with ordinary cases as well as difficult ones. Establish questions, evidence stages and reference policy, then test whether assistance improves supported review at an acceptable total burden. Retain a strong conventional process as a comparator. No live lending or investment authority is required.

The smallest useful packet is not an entire confidential data room. It is one rights-cleared facility history: original terms, operative amendments or waivers, matching report and calculation certificate, dated corrections, and only the notices or conditions needed for the selected question. Qualified reviewers should decide each claim within scope and preserve disagreement. Until that packet exists, a public filing stress case can refine the task but cannot become independent legal or financial gold.

The scientific question is which intervention changes what reviewers do, under which conditions. The engineering question is whether it can be supplied reliably. The purchasing question is whether the result justifies its price and implementation cost. None becomes settled merely because another has a promising answer.

A simpler method may be sufficient. A useful explanation may be too costly to construct. A system may improve initial analysis and fail at the first correction. Those are possible results, not reasons to redesign the test until the product wins.

The case for a credit copilot is not that it can finish the memo. It is that the institution can still explain, after evidence, reviewers and connected systems have moved on, why it relied on the conclusion and whether it may continue to do so. That is what EOJI proposes to test. It is not a result this note claims to have earned.

Data, chart sources and reproduction

Figure 1 uses reported empirical results from one retained preprint version. Figure 2 is deterministic synthetic arithmetic. Figure 3 is a coordinator-coded documentary reconstruction and remains NOT_GOLD. Those evidence classes are not interchangeable.

Each figure is supplied with canonical CSV data, a data dictionary, source and assumption notes, a deterministic plotting script, PNG/SVG exports and a JSON validation record. The scripts run offline with the installed Python/matplotlib stack. No downloaded research code executes.

Figure	Data	Evidence	Reproduce
Figure 1	CSV · dictionary	source · assumptions	script · validation
Figure 2	CSV · dictionary	source · assumptions	script · validation
Figure 3	CSV · dictionary	source · assumptions	script · validation

The source paper, observation-level microdata and SEC documents are not redistributed. Figure 1 is a transcription/recalculation check, not a replication. Figure 3 remains NOT_GOLD and lacks professional adjudication. Source-use and disclosure controls are documented in the accompanying public source map and attribution record.

References and evidence boundary

This is a synthesis of established work and proposed tests. No new theorem, completed professional-user study, independent product validation or competitor failure is claimed. No cited author or institution endorses EOJI by appearing here. Reading a source does not imply its results were replicated or every downstream use is licensed.

[1] Dyreng, S., Ferracuti, E., Hills, R., and Kubic, M. *Measurement Error when Estimating Covenant Violations*. Retained November 2024 institutional preprint, 93 pages including Internet Appendix, SHA-256 a308c23caca817aacedcb099600992a9f49876047654f41abfe21e0e70acfe1f; later published as *Measurement error when estimating covenant slack and violations*, Journal of Accounting and Economics, article 101937. Figure 1 uses Table 3 Panel B of the retained preprint, PDF p.48. The 2026 journal PDF and microdata were not captured and are not assumed identical. https://doi.org/10.1016/j.jacceco.2026.101937

[2] Doyle, J. *A Truth Maintenance System*. MIT AI Memo 521, June 1979. https://dspace.mit.edu/entities/publication/5377b306-4ecc-4687-b1f5-78cbb4a0543a

[3] Cała, J., and Missier, P. *Provenance Annotation and Analysis to Support Process Re-Computation*. IPAW 2018, LNCS 11017, pp. 3–15. https://doi.org/10.1007/978-3-319-98379-0_1

[4] fortiss. *Assurance Case Maintenance*, ExplicitCase documentation. https://download.fortiss.org/public/projects/af3/help/assuranceCases/maintenance.html

[5] Lamport, L. *Time, Clocks, and the Ordering of Events in a Distributed System*. Communications of the ACM 21(7), 1978, pp. 558–565. https://lamport.azurewebsites.net/pubs/time-clocks.pdf

[6] Featonby, M. *Making retries safe with idempotent APIs*. Amazon Builders' Library. https://aws.amazon.com/builders-library/making-retries-safe-with-idempotent-APIs/

[7] W3C. *Verifiable Credentials Data Model v2.0*. Recommendation, 15 May 2025. https://www.w3.org/TR/2025/REC-vc-data-model-2.0-20250515/

[8] Renner, J. *Evaluating AI Agents: A Hybrid Deterministic and Rubric-Based Framework*. Hebbia, 26 March 2026. First-party methodology, not independent performance evidence. https://www.hebbia.com/blog/evaluating-ai-agents-a-hybrid-deterministic-and-rubric-based-framework

[9] ExactCov. *Open-weights models against the frontier on covenant extraction*. 8 September 2026. First-party test description; no independent reproduction. https://exactcov.com/research/open-weights-vs-frontier-on-covenant-extraction

[10] Rogo Technologies. *Big Finance public-release datasheet*. Commit d794a65fe583edc6852b44c817b0a2aef33ca831. https://github.com/Rogo-Technologies/big-finance-benchmark/blob/d794a65fe583edc6852b44c817b0a2aef33ca831/data/DATASHEET.md

[11] Enke, B., and Zimmermann, F. *Correlation Neglect in Belief Formation*. Review of Economic Studies 86(1), 2019, pp. 313–332. https://doi.org/10.1093/restud/rdx081

[12] Buçinca, Z., Malaya, M. B., and Gajos, K. Z. *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. PACM HCI 5(CSCW1), 2021. https://doi.org/10.1145/3449287

[13] NASA. *NASA Task Load Index (TLX)*. https://www.nasa.gov/human-systems-integration-division/nasa-task-load-index-tlx/

[14] Hauser, D. J., Ellsworth, P. C., and Gonzalez, R. *Are Manipulation Checks Necessary?* Frontiers in Psychology 9:998, 2018. https://doi.org/10.3389/fpsyg.2018.00998

[15] Kane, J. V., and Barabas, J. *No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments*. American Journal of Political Science 63(1), 2019, pp. 234–249. https://doi.org/10.1111/ajps.12396

[16] Montgomery, J. M., Nyhan, B., and Torres, M. *How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It*. American Journal of Political Science 62(3), 2018, pp. 760–775. https://doi.org/10.1111/ajps.12357

[17] Mathur, M. B. *On the Statistical Analysis of Experiments With Manipulation Checks*. Advances in Methods and Practices in Psychological Science 8(1), 2025. https://doi.org/10.1177/25152459241297537

[18] Lakens, D. *Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses*. Social Psychological and Personality Science 8(4), 2017, pp. 355–362. https://doi.org/10.1177/1948550617697177

[19] Saunders, A., Steffen, S., and Verhoff, P. *CovenantAI - New Insights into Covenant Violations*. CEPR Discussion Paper 20117, 9 April 2025; retained public conference-paper version dated 19 October 2025. No reusable dataset or scoring package was established in the official locations inspected through 10 October 2026. https://cepr.org/publications/dp20117

[20] Skinner, J., Li, D., and Ramanathan, A. *Reaching Autonomous Consensus on Agentic Outputs*. Hebbia, 19 September 2025. First-party methodology and finance-task results; not neutral independent product validation. https://www.hebbia.com/blog/who-evaluates-the-evaluator-reaching-autonomous-consensus-on-agentic-outputs

[21] Wang, A., Meinhardt, G., Katz, J., Kim, J. H., Chaudhary, P. K., Blagden, C., and Xu, E. *BigFinanceBench: A Workflow-Grounded Benchmark for Financial-Research Agents*. arXiv:2606.03829, submitted 2 June 2026; public repository inspected at the pinned commit recorded in the package. The public subset is partially replayable; vendor authorship and a held-back set remain separate from independent validation. https://arxiv.org/abs/2606.03829

[22] Smallwood, A. *Primer Tops BigFinanceBench*. Primer, 29 July 2026. A commercial competitor's reported replay and challenge of the public subset, not a neutral independent audit. https://www.primerapp.com/blog/primer-tops-bigfinancebench/