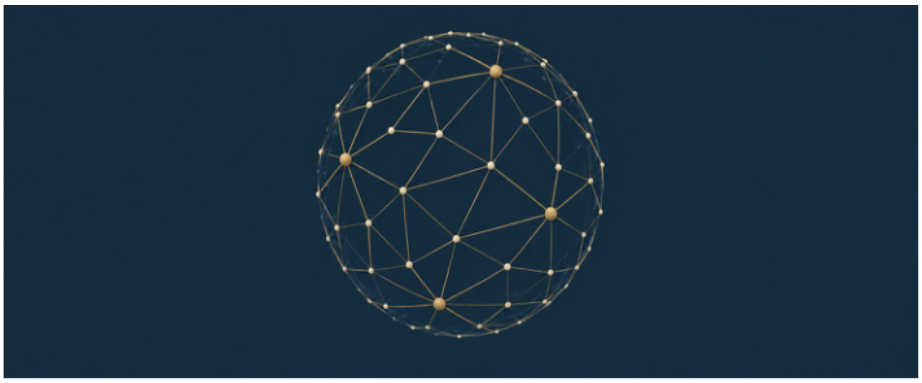
DeepFinance and the Standard for Verifiable Benchmarks

A constructive case for making financial-agent evaluation as auditable as the work it measures.

EDWIN GORDON / APERTURE / 8 NOVEMBER 2026 / REVISED 10 NOVEMBER 2026



THE QUESTION

Can we evaluate the judgment behind financial work as rigorously as its final answer?

EXECUTIVE SUMMARY

DeepFinance offers a serious, trajectory-aware benchmark for investment-banking and private-equity work. Julius Research reports 21 tasks, averaging 27.1 estimated human hours, 84 rubric criteria and 19 data-room files each.[1]

This essay distinguishes a correct answer, the evidence supporting it, and the reliability of the grade. Aperture brings a production-backed research perspective; its independent benchmark findings are still forthcoming.[3]

Scope: Julius's published methodology and quality-assurance account were reviewed. The Corliss prompt and four tab labels were inspected; the Data room, Ground truth and Rubric panel contents were not. No adverse inference follows.[2]

1 Julius Research, “DeepFinance: Can AI diligence a $100MM deal?,” Methodology and Sample task, 5 Nov 2026. Averages are reported by the authors.

2 Limited sample-panel inspection and uncached-origin freshness; see the full publication-scope note and source bibliography.

3 Aperture internal engineering records (unpublished). No external validation of comparative research quality or model-training gains is claimed.

A useful sample and a shared research question

Julius reports 21 finance tasks, averaging 27.1 estimated human hours, 84 rubric criteria and 19 data-room files each. Those are substantial design choices. The visible Corliss prompt asks for a defensible recommendation; its uninspected supporting materials are not judged here.[4]

Imagine reviewing a lending recommendation. The arithmetic is correct. The tables reconcile. The conclusion is persuasive. Yet the analyst has used management's preferred earnings subtotal where the credit agreement requires a different definition.

An arithmetic check can pass without making the recommendation dependable. The evaluator needs to examine what the number represents, why that interpretation is justified, and which conclusions depend on it.

DeepFinance takes this problem seriously. The report describes evaluation across output, workflow, and judgment, rather than numerical agreement alone. Its sample investment-diligence prompt gives readers a concrete example of the work the benchmark aims to assess.[5]

Julius also describes human-expert and agentic review addressing task clarity, evidence-supported reference solutions, acceptance of valid alternatives, and whether a failed criterion reflects the agent's work rather than a faulty reference, grader, or environment. It reports tracing disputed grades to source files and agent traces. These are relevant quality-assurance commitments that deserve explicit credit; this commentary does not independently establish their effectiveness or calibration.[6]

In my reading, the Corliss request is a useful design choice: it asks for a recommendation, payment structure, and justification rather than only a final number. Its conditional payment and timing context make the intended judgment problem tangible.

Three propositions must remain separate: the capability matters; the tested systems received particular scores; and the procedure measures the capability well enough to support the conclusions drawn from those scores.

These distinctions let us appreciate a benchmark's contribution while examining how its scores should be interpreted. Detailed criteria can make review more informative; their weighting, dependencies, and treatment of defensible alternatives remain important questions for any evaluator.

[4] Julius Research, "DeepFinance: Can AI Diligence," "Sample Task." The positive assessment concerns the visible prompt's analytical request, not the adequacy of uninspected materials.

[5] Julius Research, "DeepFinance: Can AI Diligence," "Scoring" and "Sample Task." The interpretation of the visible prompt is the author's assessment, not a measured result or a review of the other panels.

[6] Julius Research, "DeepFinance: Can AI Diligence," "Quality Assurance," accessed 8 Nov. 2026. This paragraph reports the authors' stated review procedure, not independently verified effectiveness or calibration. No claim is made that this wording was newly added.

What DeepFinance already measures

An accurate starting point for criticism is to recognize what the benchmark already does. The counts below are Julius Research's reported design characteristics, not results of an independent methodology audit.

21
TASKS
PE and investment-banking assignments

27.1
HOURS / TASK
Estimated human work time

84
CRITERIA / TASK
Average rubric requirements

19
FILES / ROOM
Average data-room size

FOUR TASK PILLARS

01 **VALUATION**
Determine value and defend the assumptions.

02 **DUE DILIGENCE**
Identify material risks in the records.

03 **UNDERWRITING**
Test economics, risks and financing terms.

04 **FINANCIAL MODELING**
Reconcile calculations and model outputs.

THREE GRADING DIMENSIONS

OUTPUT
Does the requested answer or artifact satisfy the criterion?

WORKFLOW
Was relevant evidence found and used?

JUDGMENT
Do decisions and assumptions follow from evidence?

HOW TO READ THE SCORE

Julius averages rubric-criterion satisfaction over attempts and then tasks, rerunning errored attempts to obtain ten scored executions per task. This is not the share of fully correct assignments or a direct estimate of first-attempt execution reliability.

Source: Julius Research, "DeepFinance: Can AI diligence a $100MM deal?", 5 Nov 2026, task construction, scoring, quality assurance and harness sections. This page paraphrases the authors' published account and does not certify individual graders or uninspected sample panels.

The visible Corliss prompt makes a professional analytical question concrete. Its Data room, Ground truth and Rubric panels were not examined in this review; their contents are neither presumed absent nor evaluated.[7]

The broader issue is the relationship between a benchmark's reported score and the capability it claims to measure. This essay proposes four complementary tests: inspect a consequential scored judgment; replay the grading decision; rerun the agent under specified conditions; and validate the evaluator itself against independent reference evidence. The framework and calibration example appear on pages 9–11.

Publication format is not the criterion. A technical article and repository can make grading decisions inspectable; an academic PDF can still omit decisive assumptions. What matters is what a qualified reviewer can actually examine.

> The visible sample deserves credit. Uninspected material warrants no adverse inference.

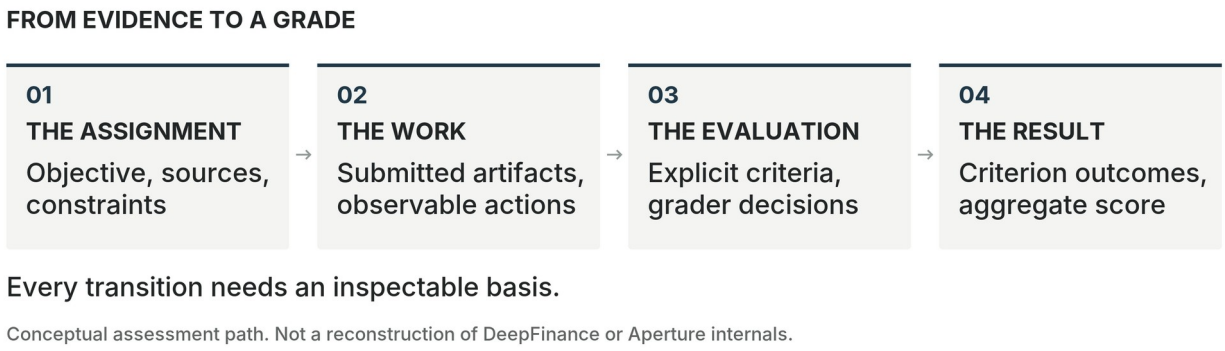


Figure 1. From evidence to a grade. A conceptual assessment path, not a reconstruction of either system or a finding that a particular component is missing.

What the research establishes

Correct answers, supported claims, and reliable grading are different achievements. The studies below examine those distinctions in different settings; they support the importance of verification without establishing one universal error rate or validating every proposed solution.[8]

[7] Limited recheck, 7 Nov. 2026: visible prompt and tab labels only; the other panels were not opened. Uncached-origin freshness and update chronology remain unverified. No absence or completeness finding follows.

[8] Synthesis of Wenqing Wang et al., “FinFIRST”; Zhu et al., “NumTrace”; Alfaqeeh et al., “Beyond Fabrication”; and Suh et al., “AgentSuite.” Their task populations and measurement procedures differ; no combined error rate is asserted.

Correct answers can lack support

FinFIRST, released in September 2026, separates answer correctness from supporting-evidence requirements. In its decomposable subset, 201 of 1,150 correct-answer attempts failed to satisfy all supporting-evidence criteria. This is a result for the evaluated configurations and subset, not a present-day error rate for every frontier model.[9]

The important methodological detail is that its judge receives the question, rubric, and **final response**, not the full tool trajectory. Its traceability score therefore reflects the evidence presented in the answer. Recorded trajectories are analyzed separately.[10]

This supports a narrower, useful conclusion: final-answer agreement can conceal missing evidentiary support. It also shows that a final deliverable can itself contain substantial evidence. **“Output-based” need not mean “check only the last number.”**

That distinction sharpens the question for a trajectory-aware benchmark: **what additional information does the execution history supply, and does using it produce better judgments?** The answer should be demonstrated rather than assumed.

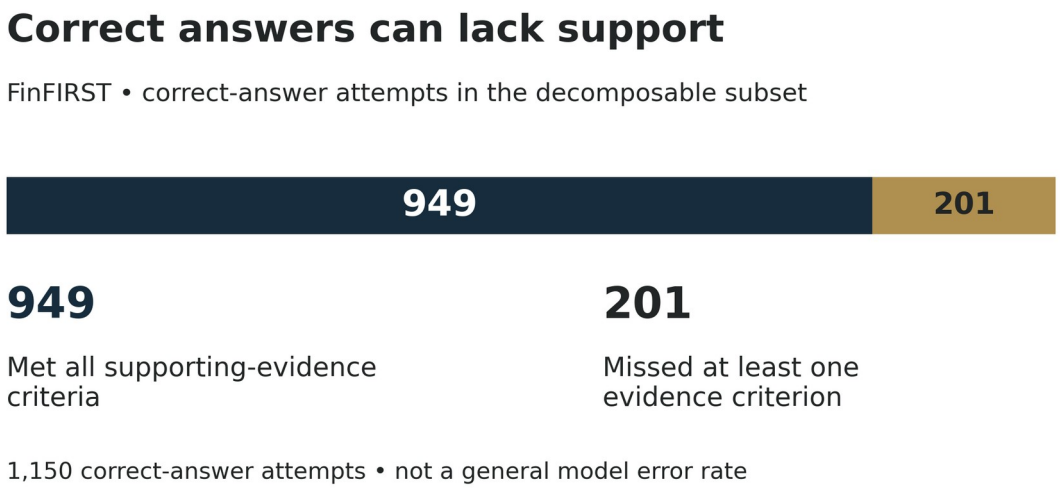


Figure 2. Correct answers, incomplete support. FinFIRST reports 201 of 1,150 correct-answer attempts missing at least one supporting-evidence criterion in its decomposable subset. The remaining 949 is calculated by subtraction. This is not a general model error rate.[11]

A number needs the right context

October’s NumTrace separates three checks: whether digits and arithmetic match the selected source; whether the number refers to the right entity, concept, period, unit, and scope; and whether authoritative evidence supports that relationship. Its deliberately narrow value check

[9] Wenqing Wang et al., “FinFIRST,” sec. 4.2. The 201/1,150 result is restricted to correct-answer attempts in the decomposable subset.

[10] Wenqing Wang et al., “FinFIRST,” sec. 3.4 and app. C. The judge scores the final response; trajectory analysis is separate.

[11] Wenqing Wang et al., “FinFIRST,” sec. 4.2. Chart reproduces the manuscript’s cited subset: 201 of 1,150 correct-answer attempts lacked complete support. The 949 remainder is calculated as 1,150 minus 201.

is not equivalent to full numeric correctness: scale and other semantic errors can be classified elsewhere.[12]

One finding is particularly relevant to evaluation transparency. After debugging its scorer, the authors report that Claude's binding pass rate increased from 70.8% to 93.8% **with model outputs unchanged**.

This was a correction to measurement, not an improvement in the agent. The debugging used inspected Claude failures, so the change is not an independent held-out validation of the repaired scorer.[13]

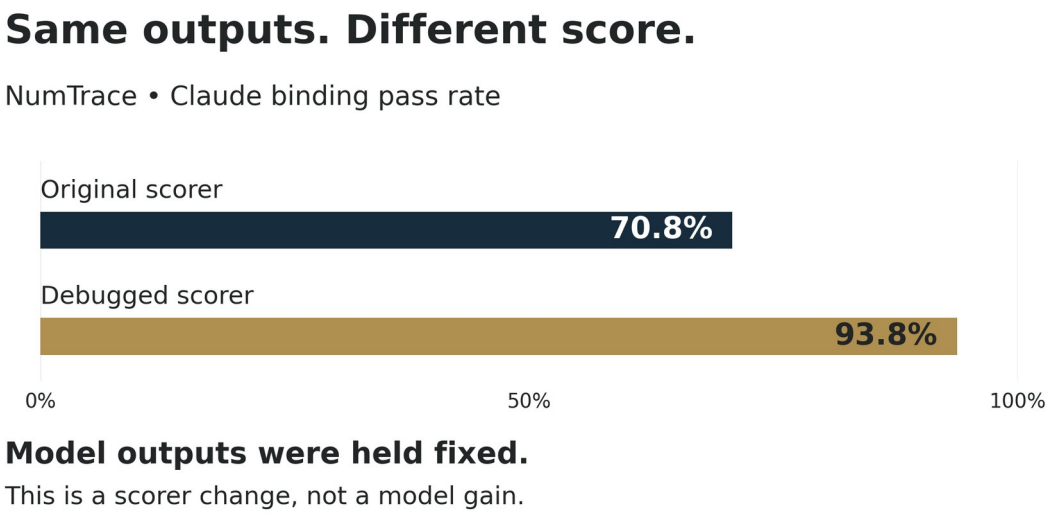


Figure 3. A scorer change, not a model gain. NumTrace reports Claude's binding pass rate at 70.8% before scorer debugging and 93.8% afterward, with the same model outputs. The debugging comparison is not independent held-out validation.[14]

The study uses a fixed retrieval-and-analysis pipeline rather than unrestricted organizational work. Its results should be interpreted within that scope. Nevertheless, the example makes the central point concrete: an evaluator is an instrument that can introduce errors, not an infallible observer.[15]

[12] Zhu et al., "NumTrace," secs. 3.2–3.4.

[13] Zhu et al., "NumTrace," sec. 3.4, app. G, and "Limitations." The measured change is scorer debugging with outputs held fixed.

[14] Zhu et al., "NumTrace," sec. 3.4, app. G, and "Limitations." Chart reproduces the cited scorer-debugging comparison, 70.8% versus 93.8%, with model outputs held fixed; it is not an independent held-out validation.

[15] Zhu et al., "NumTrace," sec. 2.4 and "Limitations."

Real numbers, wrong claims

The October revision of *Beyond fabrication* examines numerical claims in AI-written scientific papers. Of 53 unsupported claims in its manually labeled samples, 47 used values present in the retained artifacts but associated with the wrong experiment, metric, run, or statistic. These are sample counts from scientific research agents, not a financial-work failure rate.[16]

The study also distinguishes successful verification from the absence of verification. Its conservative checker misses some traceable claims. **Failure to certify a claim is therefore not, by itself, evidence that the claim is false.**[17]

For professional evaluation, both lessons matter. A source match must preserve the relationship between evidence and claim. At the same time, an evaluator should distinguish an unsupported assertion from a supported assertion its checking procedure could not resolve.

An honest uncertainty state is valuable. It should not be quietly converted into either a pass or an accusation of error.

Verification must do useful work

Agents Are Systems, Not Models, published in October, studies scientific tasks under different operating configurations. Its verification ablation supplies an oracle tool that scores a current submission against the reference solution. The authors report improved outcomes and calibration, accompanied by additional effort and cost. Simply strengthening a prompt's instruction to verify is a different intervention.[18]

The qualification is decisive: reference-aware oracle feedback is stronger than ordinary production verification. A research system usually does not possess the correct final answer in advance. These findings motivate testing useful checking capabilities; they do not establish that a generic reviewer, a particular financial workflow, or Aperture will achieve the same improvement.[19]

The experiment also reinforces the need to describe the system being evaluated. Tools, feedback, information, and budgets are parts of the treatment. Their benefits should not be silently attributed to the underlying model alone.

The benchmark and the agent can both fail

The ICML 2026 AgentSuite paper reports audits across six agent benchmarks. Its framework examines interactions among instructions, environments, reference behavior, and evaluation.

[16] Alfaqeeh et al., "Beyond Fabrication," sec. 5.2. These are raw labeled-sample counts, not a financial-work failure rate.

[17] Alfaqeeh et al., "Beyond Fabrication," sec. 5.3, "Non-verification is not evidence of falsehood."

[18] Wiedmann et al., "Agents Are Systems," sec. 4.3 and app. A.8. The intervention uses reference-aware oracle feedback.

[19] Wiedmann et al., "Agents Are Systems," app. A.8, "Oracle ablation." Transfer to ordinary production verification is the author's qualification, not a measured Aperture result.

The authors find defects that materially affect measured outcomes. This is evidence for auditing benchmarks, not an audit of DeepFinance itself; an automated auditing system also needs validation.[20]

A historical counterweight comes from *Why Do Multi-Agent LLM Systems Fail?* Its analysis identifies failure modes involving design, coordination, and verification. Those results concern the evaluated systems and model generations, not the current frontier's failure rate. They remain a useful warning against assuming that additional specialists and reviewers automatically produce reliability.[21]

Taken together, these studies support investigating evidence, interpretation, verification, and recovery. They do **not** establish that every extra step helps, that all work must follow one preferred trajectory, or that more recorded activity means more dependable reasoning.

The research question concerns which relationships are preserved, which checks are effective, and whether the resulting judgment can withstand independent examination.

Compare disclosed objects, not just ambitions

Benchmark contributions can be complementary rather than mutually exclusive. The comparison below recognizes different ways to make work inspectable, without ranking DeepFinance's overall disclosure on the basis of panels this review has not examined.[22]

BigFinanceBench describes weighted criteria, its judging setup, limitations, and an annotated execution example. Its public repository provides a 50-item subset with reference answers and rubrics, grading components, and example workbooks. The remainder is not an unrestricted public release. These materials let a researcher challenge a concrete grading choice without establishing that every historical score or fresh run is reproducible.[23]

Halluminate's Westworld Finance Diligence Bench illustrates why a blog can be methodologically useful. Its technical article includes a worked task, verifier weights, aggregation, and quality-assurance procedures. It distinguishes grading the final artifacts from diagnosing behavior through trajectories. The full private-deal corpus is not thereby made independently rerunnable.[24]

[20] Suh et al., "AgentSuite." The six-benchmark audit does not include DeepFinance.

[21] Cemri et al., "Why Do Multi-Agent LLM Systems Fail?" Version 3 is dated 26 Oct. 2025; the discussion does not estimate current-frontier failure rates.

[22] Comparative interpretation of Julius Research, "DeepFinance: Can AI Diligence"; Alex Wang et al., "BigFinanceBench"; Rogo Technologies, Big Finance Benchmark; Halluminate, "Westworld Finance Diligence Bench"; and Patronus AI, FinanceBench. No overall disclosure ranking is made.

[23] Alex Wang et al., "BigFinanceBench"; Rogo Technologies, *Big Finance Benchmark*, README and released subset. These disclose inspectable components, not unrestricted access to the full benchmark.

[24] Halluminate, "Westworld Finance Diligence Bench." The worked example, verifier design, and QA descriptions do not release the full private-deal corpus.

FinanceBench's public subset includes 150 annotated examples with answers, evidence, document information, and evaluated completions. Its question-answering scope is narrower than a transaction workflow, but necessary subskills remain legitimate research targets.[25]

The existing projects and the visible DeepFinance sample offer different entry points for scrutiny. Differences in task scope can be useful rather than disqualifying. Because this review has not inspected DeepFinance's other panels, it makes no overall ranking of disclosure quality.

A spreadsheet can expose formulas, assumptions, definitions, and source references. A memo can preserve uncertainty and rejected alternatives. Artifact-based evaluation is not necessarily evidence-blind, just as trajectory-aware evaluation is not automatically valid. The useful comparison asks which information changes the assessment.

> Realism and transparency should strengthen one another. Neither substitutes for the other.

Four tests of a trustworthy benchmark

Inspecting a score, replaying a grade, rerunning an agent, and validating the measurement answer different questions. This proposed framework separates the evidence each requires, without implying a certification ladder or demanding unrestricted release of confidential data.[26]

These questions are related, but none substitutes for all the others. A deterministic grader can consistently reward the wrong thing. A persuasive reference solution can depend on assumptions the candidate was never given. Repeating a fresh agent run need not reproduce the original output, while retained evidence may still permit inspection of the original grade.

Where grading is probabilistic, replay should report variability and disagreement handling. Where judgment is legitimately contested, the procedure should explain how defensible alternatives receive credit.

Bank-World supplies a complementary grader-validation example: no-action controls, deliberate action mutations, valid-alternative checks and deterministic replay. These are author-reported checks, not independent validation. The paper explicitly separates repeatability and judge agreement from correctness, and retains the complete environments, graders and trajectories privately.[27]

---

[25] Patronus AI, *FinanceBench*, README and annotated public subset.

[26] The four-part distinction is the author's proposed framework. See NeurIPS, "Paper Checklist Guidelines," item 4, "Experimental Result Reproducibility."

[27] Rupawalla, "Bank-World," app. C, sec. 2 and "Availability." Author-reported controls; not an independent audit.

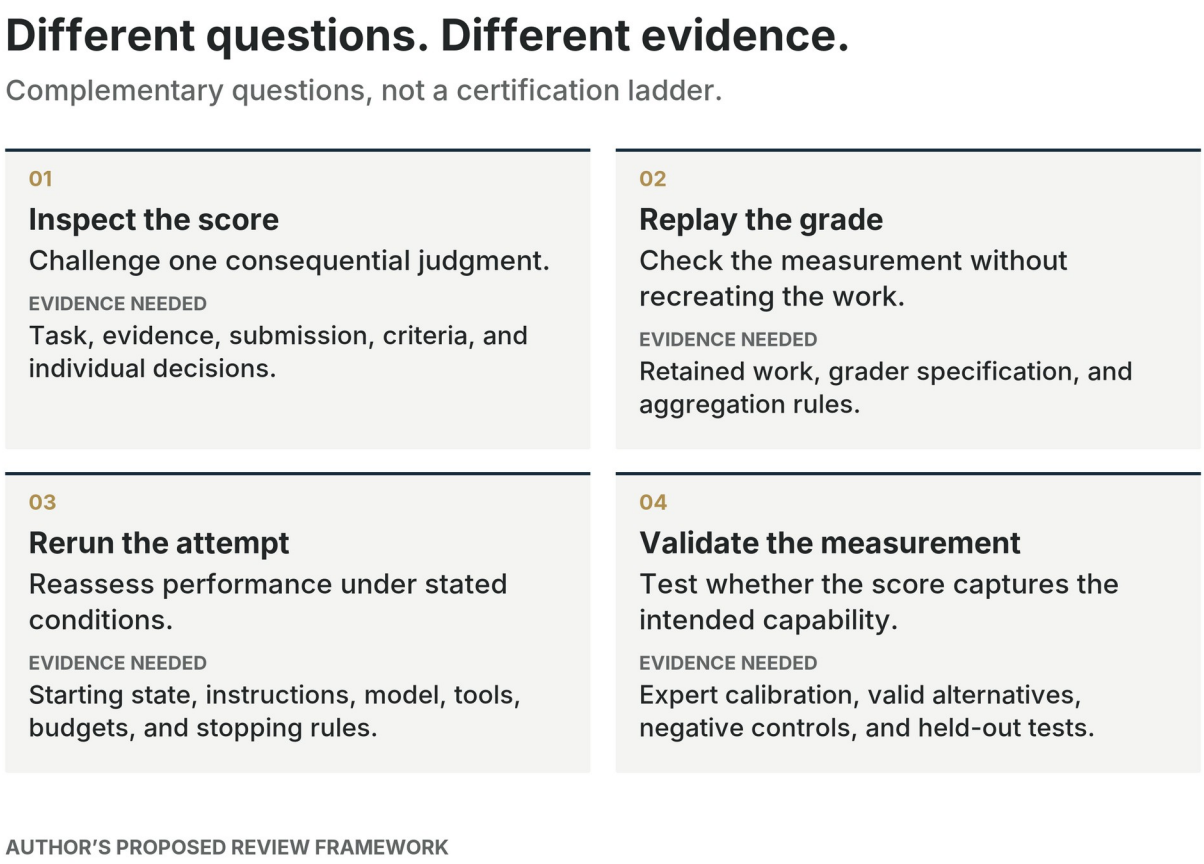


Figure 4. Four tests of a trustworthy benchmark. The author's proposed questions distinguish inspectability, grade replay, fresh execution, and measurement validity. They are complementary tests, not a certification ladder.[28]

For DeepFinance, these are possible questions for a fuller review, not failed tests established here. Published examples can make the intended standard easier to discuss. Assessing its application and validation requires examining the relevant materials, including any access arrangements the authors provide.

NeurIPS's checklist offers a useful general principle: provide a reasonable avenue to reproduce or verify a contribution, with access and experimental detail appropriate to the work. It does not make unrestricted release of all code and confidential data the only acceptable route. That is a research standard to learn from, not a claim that a company blog is governed by conference submission rules.[29]

[28] Author's proposed review framework, re-presenting the four-question table in this article. For the general reproducibility principle, see NeurIPS, "Paper Checklist Guidelines," items 4–8. The panels do not represent an accreditation sequence or a measured benchmark result.

[29] NeurIPS, "Paper Checklist Guidelines," items 4–8. This is a methodological reference, not a rule imposed on Julius Research.

How can a grader be wrong?

A useful evaluation separates the underlying truth about a claim from the checker's decision. Two kinds of mistakes can distort conclusions even when the score is perfectly reproducible.

INDEPENDENT ADJUDICATION

	GRADER ACCEPTS	GRADER REJECTS
SUPPORTED CLAIM	CORRECT DECISION **Correct acceptance** A supported claim receives appropriate credit.	GRADER ERROR **False rejection** Valid evidence or an acceptable method is rejected.
UNSUPPORTED CLAIM	GRADER ERROR **False acceptance** A claim passes despite incorrect context or support.	CORRECT DECISION **Correct rejection** A genuinely unsupported material claim is rejected.

A THIRD STATUS IS ESSENTIAL: UNRESOLVED

A checker that cannot link the evidence should not automatically treat the claim as false or true. Record abstention, missing evidence, unassessable cases and coverage separately.

HIGH-VALUE STRESS TESTS

- **Valid citation, wrong context**
  A real filing supports a different period or metric.
- **Correct result, invalid rejection**
  An equivalent source identifier or justified alternate method is rejected.
- **Local fix, stale memo**
  A corrected workbook leaves an outdated investment recommendation.
- **Missing evidence**
  The grader abstains rather than inventing support or declaring error.

Boundary: Conceptual only: no error rates for Julius or Aperture. Julius documents expert and agentic QA; these are proposed independent checks, not discovered shortcomings.

What the score actually measures

DeepFinance's rubric percentage measures requirements satisfied, not the fraction of assignments meeting every requirement. Its protocol uses native harnesses and reruns errored attempts. Those choices can serve a useful system-comparison objective; they should be interpreted within that scope rather than treated as evidence of improper practice.[30]

A useful extension for operational buyers would connect scored attempts to execution reliability, recovery and total effort. A benchmark may reasonably isolate some questions rather than answer all of them. Its chosen scope is not evidence of a measurement defect.

Repeated executions reveal variability; they do not turn one assignment into ten independent deals. Several rubric failures may descend from one mistaken premise. Report criteria, attempts and assignments separately. A confidence interval does not validate a rubric's meaning, a grader's accuracy or the representativeness of the work.

Bank-World complements this distinction in simulated banking operations. Across 50 tasks and 1,600 attempts, the study separates accepted decisions from completed required actions and examines consistency across repeated runs. An accepted main-answer judgment is not a full-task pass. These findings motivate, but do not answer, Aperture's questions about verification, downstream repair and learning from research trajectories.[31]

A correction is not a completed repair

Fixing an upstream number is insufficient when dependent conclusions remain stale. Return to the hypothetical lending assignment: the agent uses the wrong earnings definition, then calculates borrowing capacity correctly from that input. The arithmetic check should pass; the definition check should fail.

Now the error is found. The earnings adjustment changes and the model is recalculated, but the executive memo still contains the previous borrowing-capacity figure. That is a partial repair, not a fully reconciled assignment. The recommendation must be assessed against the material consequence of the error.

These are proposed evaluation states, not measured DeepFinance or Aperture results.[32]

---

[30] Julius Research, “DeepFinance: Can AI Diligence,” “Scoring” and “Harness.” Reported protocol, not a current-ranking claim or independent rerun.

[31] Rupawalla, “Bank-World,” secs. 3–4.3 and 5.2.1. Results concern banking tasks, not Aperture.

[32] Author's hypothetical lending example. No DeepFinance or Aperture performance result is asserted.

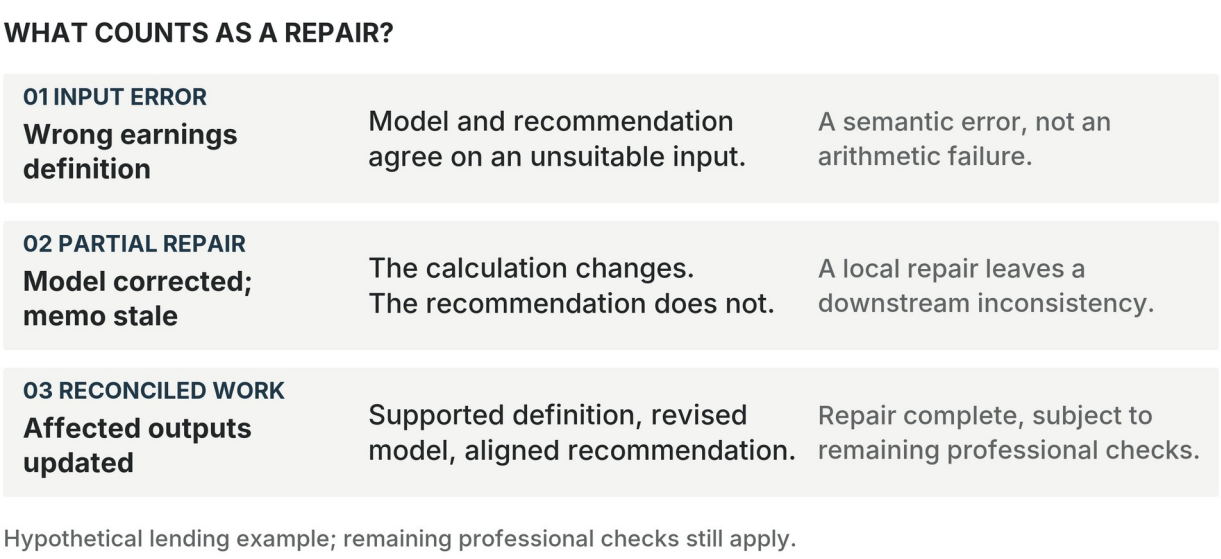


Figure 5. What counts as a repair? The article’s hypothetical lending example; no benchmark result or Aperture performance claim is represented.

A useful test would vary the evidence deliberately. Keep the number but remove its support. Replace a citation with a genuine but irrelevant source. Introduce a later authoritative correction. Supply a different valid method. Update affected outputs while leaving unrelated outputs intact.

The evaluator should respond to the relevant differences, not to how closely a submission resembles the author’s preferred sequence. Observed actions and artifacts can support these tests without exposing private model reasoning.

Recovery also needs its own accounting. Restarting after an execution failure, discovering an analytical mistake before submission, repairing after external feedback, and responding to changed evidence are different capabilities.

The question is which recovery modes a protocol measures, not whether it uses the word “trajectory.” Multiple attempts do not, by themselves, constitute feedback-guided repair. Conversely, a bounded assignment can contain meaningful self-correction without modeling an entire organization.

A study should preserve the pre-feedback result and specify what the feedback revealed. Repair after receiving the answer key is not independent diagnosis. Appropriate escalation is not successful completion, but neither should it be indistinguishable from confidently delivering unsupported work.

Aperture: from framework to evidence

Aperture reports an implemented framework and bounded internal tests, not independently established superiority. The distinction is between what its workflows are designed to preserve, what selected engineering checks document, and whether those records can improve another model's learning.[33]

Developer account · Implemented framework

Aperture was built first to deliver dependable financial research for institutional investors. Existing asset-management clients consume this work. Feedback on the research's depth and quality, including suggestions from some clients, led us to explore a second application: the verified execution records generated through those same assignments. That establishes an operating and commercial context, not independent proof of superior accuracy.[34]

The **organization** is simulated in its coordinated specialist roles and workflows, **not** in the professional purpose of the assignment. Specialist agents assemble sourced analysis, execute calculations, identify conflicts, review findings and correct errors as the research proceeds. The work product and its observable execution history emerge from the same process, rather than the history being written later to resemble a successful research trajectory.

This structure matters because a professional conclusion is downstream of several judgments. One specialist finding may alter an assumption, which changes a model and its scenarios, which in turn changes the investment recommendation. A locally correct computation can still yield a wrong conclusion if its economic premise is unsupported or dependent artifacts remain stale.

At the technical level, we distinguish three properties. **Traceability** means a reviewer can inspect the cited record. **Internal consistency** means outputs agree with declared inputs and calculations. **Substantive correctness** asks whether the source, premise and interpretation are appropriate. None proves the other two by itself. Aperture's bounded internal tests inform this engineering approach; independently evaluated professional quality remains a question for the forthcoming studies.

[33] Summary of the developer account and bounded internal records: Aperture, *Product positioning and messaging*; Aperture, *Combined supervised-learning export verification*; Aperture, *Executable task-environment proof of concept*; and Aperture, *Selected-case export validation and evidence holds*. Unpublished internal evidence, not independently established superiority or training improvement.

[34] Aperture, Product positioning and messaging. Unpublished developer account, 4 Nov. 2026; not an external implementation or performance audit.

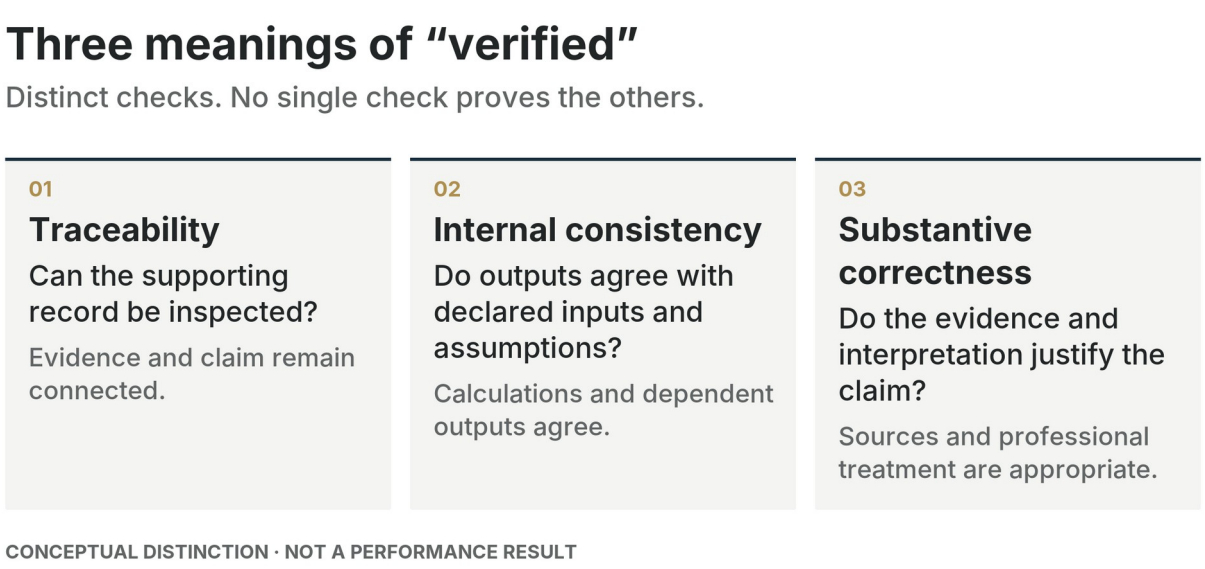


Figure 6. Three meanings of “verified.” Traceability, internal consistency, and substantive correctness are separate checks. This conceptual distinction depicts neither measured performance nor proprietary architecture.

Internal validation · Bounded tests

What the internal evidence establishes

Our development records contain more than a proposed architecture. They report bounded tests of selected evidence packages and an executable grading prototype. **These are internal engineering findings,** not a peer-reviewed performance study, an external audit, or a DeepFinance result.[35]

Reconstruction and receiving compatibility. Selected export checks tested whether retained records were preserved, whether supporting evidence resolved from the delivered package, and whether scoped tool exchanges matched the recorded execution. Receiving checks tested selected examples through two specified training-data loaders.

These findings support inspectability and compatibility for the tested selections. They do not establish complete-session trajectories, universal historical correctness, or improved model behavior.[36]

Some exported prompts and completions were study-authored representations of retained work. They should not be described as unmodified native transcripts. This is an important distinction between preserving source evidence and presenting a curated learning example.[37]

[35] Aperture, Combined supervised-learning export verification; Aperture, Executable task-environment proof of concept; Aperture, Selected-case export validation and evidence holds. Unpublished internal engineering records; no external replication is claimed.

[36] Aperture, Combined supervised-learning export verification, “Technical verification” and “Interpretation and permissions.” Unpublished internal record.

[37] Aperture, Combined supervised-learning export verification, “Interpretation and permissions”; Aperture, Selected-case export validation and evidence holds. Both are unpublished internal records; curated targets are not raw transcripts.

Controlled scoring and failure behavior. A separate prototype tested scripted correct and incorrect submissions against declared arithmetic expectations. It also used injected failure cases. The record describes a discovered defect, a regression that exposed it, a correction, and repeated checks.

That is concrete evidence of testing the evaluator itself. The demonstrated scope remains conditional arithmetic and execution-to-scoring behavior; the attempts were not generated by a learning model.[38]

Limits that remain visible. Other validation records withheld proposed tool-use views when retained exchanges did not establish the required task-specific connection. A reconstructed blocked outcome remained blocked; packaging it did not turn it into a successful remediation. These decisions support a bounded lesson about evidence discipline, not a claim that every incomplete record is automatically detected.[39]

The test details and underlying packets are not released with this article. These are author-reported internal findings, not evidence of superior public disclosure. Readers should not treat our description of an experiment as equivalent to access to it, or as a performance comparison with Julius.

What the findings do—and do not—show

The practical lesson is that evidence preservation, portable reconstruction, grading behavior, and professional correctness are separate engineering and research questions. A framework can make them separately testable. It cannot collapse them into a universal assurance.[40]

For Aperture, the externally meaningful demonstration is specific: a qualified receiving reviewer can understand a selected assignment, inspect its permitted evidence, reproduce its declared checks, and determine whether a recorded correction reached the relevant outputs. The demonstration must identify its tested scope.

Where that last property has not been demonstrated for a particular example, it should remain unestablished. Source references, intact files, and successful loading are necessary for some uses but do not prove the full analytical sequence.

This is also why native work units, retries, revisions, and alternative export formats should not be counted as independent tasks merely because each has a separate record. A useful dataset needs a clear unit of work and an account of shared evidence.

[38] Aperture, Executable task-environment proof of concept, “What was demonstrated” and “Validation and remaining work.” Unpublished internal record.

[39] Aperture, Selected-case export validation and evidence holds, “Results and holds.” Unpublished internal record; a preserved block is not successful remediation.

[40] Author’s interpretation of the bounded internal findings in Aperture, Combined supervised-learning export verification; Aperture, Executable task-environment proof of concept; and Aperture, Selected-case export validation and evidence holds. It is not an additional empirical result.

> A repeatable framework creates the conditions for verification. Its existence alone does not establish comparative performance.

Build on the release, then test the claims

DeepFinance's released methodology and sample prompt provide a substantial foundation for discussion. A constructive next step is to connect specific grading decisions to their evidence, acceptance of defensible alternatives and stated execution conditions. These questions extend the conversation; they are not an adverse finding about material outside this review.

Proposed research · No outcomes reported

Start with what is already released

A fuller review, beyond this article's limited inspection, could examine the sample's source files, reference reasoning, rubric, and any scored submissions. It should first establish what the released materials already enable and credit those contributions. No conclusion about the uninspected panels is supplied here.

Any follow-up request should address a specific unresolved question and consider an appropriate public or controlled-access route. A worked specimen can be useful without exposing the entire hidden test set. These are possible next steps for collaborative scrutiny, not a list of omissions established in this review.[41]

The assessment should distinguish what the scorer receives, who or what judges each requirement, how partial credit and aggregation work, and how missing records or disagreements are treated. Separately, identify execution conditions, failures, resource use, and the task, submission, and grader versions associated with each result.

Confidentiality and benchmark contamination can reasonably limit full public release. Useful alternatives include a rights-cleared specimen, a sealed evaluation set or qualified controlled-access review. The important distinction is which findings a reviewer can actually inspect, not whether every private input or implementation detail is unrestricted.

[41] Author's proposed next steps beyond this limited review. Start with what the released materials enable and request additional evidence only for a specific unresolved question. Compare NeurIPS, "Paper Checklist Guidelines," items 4–5. No omissions are inferred from uninspected panels.

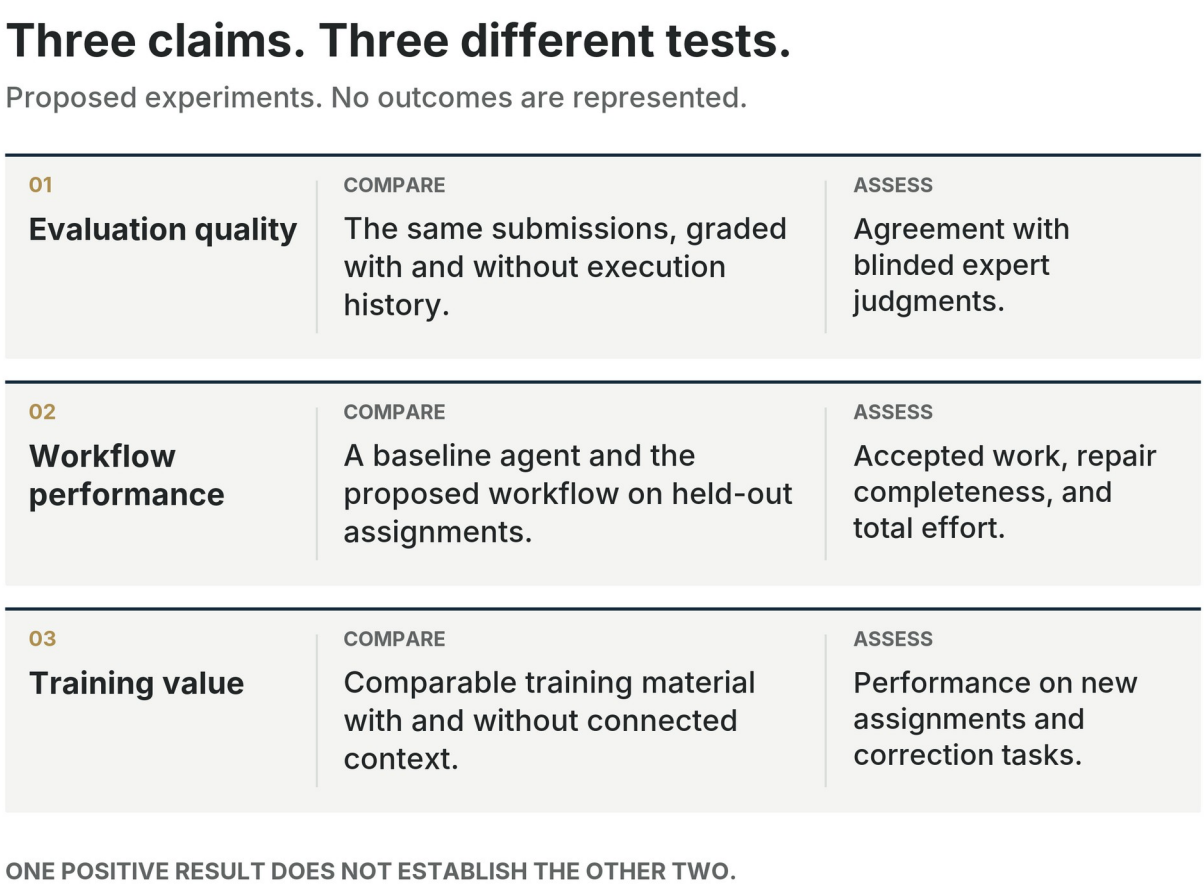


Figure 7. Three claims, three different tests. Evaluation quality, workflow performance, and training value need separate comparisons. Proposed experiments only; no outcomes are represented.[42]

Test what the execution history adds

A particularly useful experiment would assess **the same submissions** under two conditions. One evaluator sees the finished artifact, including its citations, calculations, and explanations. Another also receives the authorized observable execution history.[43]

Compare both with blinded expert judgments on consequential errors, valid alternatives, and correction completeness. Hold evaluator capability and review budget comparable where possible; otherwise report the additional cost and access as part of the treatment. Protect the expert reference decisions from the evaluated graders.

This would test whether trajectory access improves discrimination, where it helps, and whether it sometimes introduces irrelevant or misleading evidence. It would avoid assuming that longer records are inherently more informative.

[42] Author's proposed experimental comparisons. Final-response evidence and trajectory access are distinguished using Wenqing Wang et al., "FinFIRST," sec. 3.4, and Julius Research, "DeepFinance: Can AI Diligence," "Scoring." Neither the diagram nor the unpublished Aperture records reports outcomes for these three proposed tests.

[43] Proposed ablation, not a reported experiment. It follows the distinction between final-response evidence in Wenqing Wang et al., "FinFIRST," sec. 3.4 and execution-aware grading described by Julius Research, "DeepFinance: Can AI Diligence," "Scoring."

The grader should face negative controls: correct arithmetic with inappropriate inputs, an irrelevant source containing the right number, a local repair with a stale memo, and an acceptable alternative method. Report false acceptance, false rejection, disagreement, and coverage. A strict evaluator that rejects everything is not a useful professional standard.

Separate system gains from learning gains

A system-performance experiment would compare a native-agent baseline with the proposed workflow on held-out assignments. Use comparable information and a declared resource budget; disclose differences in tools, feedback, and human assistance. Preserve the first submission, then report detection, repair completeness, final consequential errors, accepted completion, and total effort.

A training experiment asks a different question. Does learning from connected work make another model more dependable on new assignments?

Aperture's research hypothesis is that preserving dependencies, observed errors, and corrections may provide useful learning signal beyond equivalent isolated examples.

A suitable comparison would hold the underlying cases, content exposure, training recipe, and token budget as constant as practicable while varying the connected context. An additional arm could test the contribution of correction history rather than bundling every change together.[44]

Hold out whole cases or source-run groups, not merely different renderings of the same work. Test unfamiliar document families and changes to upstream assumptions. Measure whether affected outputs are repaired and whether unaffected outputs remain valid.

These are proposed experiments, not reported Aperture training results. Success in receiving a dataset, executing a check, or operating a research workflow cannot substitute for them.

[44] Author's proposed training experiment. The unpublished records Aperture, Combined supervised-learning export verification and Aperture, Executable task-environment proof of concept report no model-training outcome.

A useful contribution to a shared research agenda

DeepFinance makes a consequential contribution by treating investment-banking and private-equity assignments as demanding, evidence-rich work rather than isolated numerical questions. Its published approach deserves serious engagement. The complementary question is how to test, independently, whether a scored result captures the correctness and recoverability that professional decisions require.[45]

Aperture's proposed research program approaches that question in stages. Study I will measure first-pass correctness and the effects and costs of verification and remediation. Study II will test fidelity across specialist handoffs and downstream corrections. A later human-comparator study will assess professional quality. These are distinct from operational banking-task evaluations; whether connected histories improve model training requires a separate experiment.

Publishing this essay ahead of the benchmarks is intentional. The research questions, measurement distinctions and failure cases should be visible **before** the empirical findings. We should be willing to report where a grader rejects valid work, where a verifier misses a consequential mistake, or where added coordination fails to improve the result. Commercial use provides an important reason to investigate these problems; it does not exempt Aperture from independent measurement.

Our aim is not to impose one architecture or preferred reasoning trace. It is to make professional judgments, evidence paths and corrective actions open to proportionate scrutiny—by benchmark authors, independent reviewers and practitioners alike. That standard applies as much to Aperture as to Julius or any other research group.

> When the capability being measured is auditability, the path from evidence to score is part of the result.

[45] Author's synthesis: Julius Research, "DeepFinance: Can AI Diligence," visible sample; Wenqing Wang et al., "FinFIRST"; Zhu et al., "NumTrace"; Alfaqeeh et al., "Beyond Fabrication"; and Suh et al., "AgentSuite." The assessment credits the premise and visible example without validating uninspected materials or asserting superiority.

Publication scope

This pre-benchmark essay reports no comparative Aperture performance. The Bank-World amendment and editorial/source-check pass are dated 10 November 2026. Julius's public methodology and quality-assurance text were rechecked; the Corliss inspection remains limited to its visible prompt and tab labels, not the other panels. No omission or inadequacy is inferred; update chronology remains unverified. An earlier internal suggestion that no sample existed is withdrawn. The cited numerical examples from FinFIRST, NumTrace and Beyond Fabrication, and selected methodological claims, were rechecked against primary sources. Other references retain their stated review scope. No benchmark, production session or training experiment was rerun. The author is Aperture's founder; internal findings remain developer-reported. No private implementation or evidence package is released. This editorial audit is not a review endorsed by Julius or Arcanum.

Works Cited

Alfaqeeh, Mosab, et al. “Beyond Fabrication: Misattributed Numbers in Papers Written by AI Research Agents.” *arXiv*, version 2, 31 Oct. 2026, https://arxiv.org/abs/2610.38691v2. Accessed 6 Nov. 2026. Preprint version consulted.

Aperture. Combined supervised-learning export verification. 5 Nov. 2026. Unpublished internal engineering record. Public descriptive title; underlying packet not released.

Aperture. Executable task-environment proof of concept. 1 Nov. 2026. Unpublished internal engineering record. Public descriptive title; scripted tests, not training results.

Aperture. Product positioning and messaging: canonical internal account. 4 Nov. 2026. Unpublished internal positioning record. Public descriptive title; not an independent audit.

Aperture. Selected-case export validation and evidence holds. 5 Nov. 2026. Unpublished internal engineering record. Public descriptive title; historical dispositions retained.

Cemri, Mert, et al. “Why Do Multi-Agent LLM Systems Fail?” *arXiv*, version 3, 26 Oct. 2025, https://arxiv.org/abs/2503.13657v3. Accessed 6 Nov. 2026. Preprint version consulted.

Halluminate. “Westworld Finance Diligence Bench: Evaluating AI Agents on a Complete Company Acquisition Due-Diligence Process.” *Halluminate*, 11 Aug. 2026, https://www.halluminate.ai/blog/due-diligence-bench. Accessed 6 Nov. 2026. Living technical article.

Julius Research. “DeepFinance.” *Julius Research*, https://www.juliusresearch.ai/deep-finance. Accessed 7 Nov. 2026. Living leaderboard; values retrieved 7 Nov. 2026.

Julius Research. “DeepFinance: Can AI Diligence a $100MM Deal?” *Julius Research*, 5 Nov. 2026, https://www.juliusresearch.ai/blog/deep-finance. Accessed 8 Nov. 2026.

Works Cited (continued)

NeurIPS. “NeurIPS Paper Checklist Guidelines.” *NeurIPS*, https://neurips.cc/public/guides/PaperChecklist. Accessed 6 Nov. 2026.

Patronus AI. *FinanceBench*. *GitHub*, https://github.com/patronus-ai/financebench. Accessed 6 Nov. 2026. Repository and README.

Rogo Technologies. *Big Finance Benchmark*. *GitHub*, https://github.com/Rogo-Technologies/big-finance-benchmark. Accessed 6 Nov. 2026. Repository and README.

Rupawalla, Daniel. “Bank-World: A Benchmark for Evaluating Agents on Banking Workflows.” *arXiv*, version 1, 29 Oct. 2026, https://arxiv.org/html/2610.39733v1. Accessed 10 Nov. 2026. Preprint.

Suh, Hyewon, et al. “AgentSuite: Toward More Reliable Agent Evaluation with a Component-Based Benchmark Auditing Pipeline.” *Proceedings of the 43rd International Conference on Machine Learning*, vol. 306, PMLR, 2026, pp. 116300–19, https://proceedings.mlr.press/v306/suh26a.html. Accessed 6 Nov. 2026.

Wang, Alex, et al. “BigFinanceBench: A Workflow-Grounded Benchmark for Financial-Research Agents.” *arXiv*, version 1, 2 June 2026, https://arxiv.org/abs/2606.03829v1. Accessed 6 Nov. 2026. Preprint.

Wang, Wenqing, et al. “FinFIRST: Benchmarking Search Agents for Financial Information Retrieval, Sourcing and Traceability.” *arXiv*, version 1, 21 Sept. 2026, https://arxiv.org/abs/2609.25192v1. Accessed 6 Nov. 2026. Preprint.

Wiedmann, Luis, et al. “Agents Are Systems, Not Models: Rethinking Agentic Evaluation.” *arXiv*, version 1, 1 Oct. 2026, https://arxiv.org/abs/2610.01618v1. Accessed 6 Nov. 2026. Preprint.

Zhu, Tianlei, et al. “NumTrace: Auditing the Number Provenance of Financial Analyst Agents.” *arXiv*, version 1, 26 Oct. 2026, https://arxiv.org/abs/2610.34887v1. Accessed 6 Nov. 2026. Preprint.