WHAT DID THE VERIFIER VERIFY? AUDITING EVIDENCE ATTRIBUTION IN SCIENTIFIC AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific agents can reproduce numerical results while assigning them to unsupported claims. We audit their verification pipelines by separating execution-evidence representation, original-claim representation, and deterministic checking. Historical workflows motivate a controlled follow-up with 24 assertions arranged in 12 workflow-level pairs, two model configurations, three review interfaces, and two repetitions. All interfaces can access the same execution artifacts. Luna's direct review makes one false authorization per repetition, whereas full structured review makes none before deterministic checking. GPT-5.5 direct review is correct on every item in both repetitions. Ordinary and specialized implementations of the same checking contract agree on all 192 structured decisions. Replacing extracted evidence with executor records repairs one workflow-level error in one compact-extraction run, without a repeated advantage over full structured review. The follow-up does not reproduce the extraction-induced degradation observed in exploratory cases. Our component-audit framing is a retrospective synthesis of these results, not confirmation of that hypothesis. It makes performance attribution auditable through matched ordinary controls, intermediate judgments, and field-level traces. In this controlled suite, the observed safety improvement is already present upstream of deterministic checking: evaluating only final verdicts would obscure both its location and remaining representation errors.

1 INTRODUCTION

A scientific verifier can check a reported number without checking the scientific statement attached to it. In a historical data-to-paper workflow (Ifargan et al., 2025) examined here, a coefficient of 34.9 and a p-value of .0196 were accurately reproduced from a regression table. The table assigned them to the intercept; the generated prose assigned them to a policy-change association. The subsequent numerical-consistency check reported no mismatches among 17 values. These observations are compatible: the numerical tokens were traceable, but their attribution was wrong. This source-level example distinguishes numerical consistency from support for the original statement; it does not establish that every review stage failed or that a current model would repeat the error.

Adding a deterministic checker does not automatically connect these targets. A checker may compare a structured evidence record with a structured claim and correctly accept the pair, even when a model has misread the execution or changed the meaning of the original sentence. Conversely, a well-organized review interface may resolve an error before the checker is applied. In either case, the final pipeline score does not by itself identify the component responsible for success or failure.

We study this distinction through a component audit rather than a claim of a universally stronger verifier. Our evaluation separates a model's direct judgment, its judgment while producing structured representations, and the verdict obtained by passing those representations to deterministic rules. We compare a specialized RIEC scope checker against an ordinary checker implemented separately against the same finite contract. We additionally replace extracted evidence with executor-generated records while holding the extracted claim fixed. This intervention isolates what correcting the evidence side can change; it does not silently repair the original-claim representation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

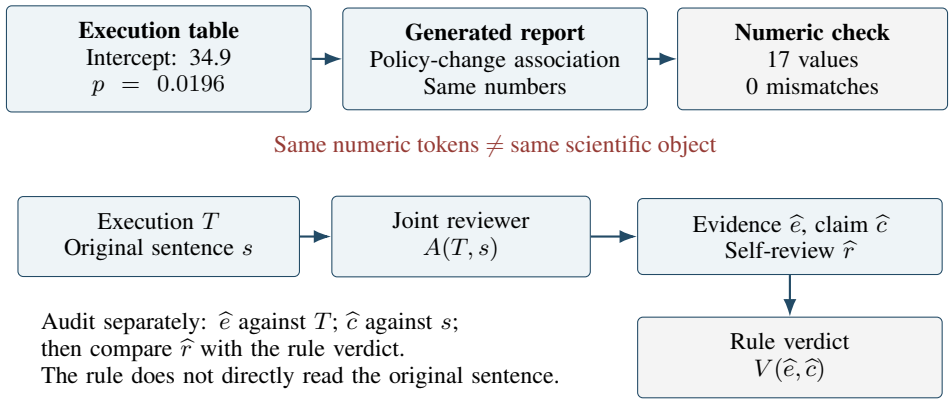


Figure 1: What was checked? The historical D2 example is a source-inspection finding, not an automatic detection result, and is excluded from the fixed first-claim denominator. The lower panel is the actual joint-review interface; auditing two representations does not imply two information-isolated parsers.

The results constrain an initially attractive hypothesis. Earlier exploratory cards exposed cases in which an extraction-based pipeline authorized more unsupported assertions than direct review. A frozen follow-up did not reproduce the prespecified degradation pattern. On its 24 assertions, GPT-5.5 direct review was correct in both repetitions. Luna direct review repeatedly missed one preprocessing-access issue, but its full structured review already avoided false authorization before rules were applied. The ordinary and specialized checkers produced identical outputs throughout the structured follow-up. A source-side intervention corrected one compact-extraction failure, without establishing an advantage over the complete ordinary workflow.

We make three contributions to the evaluation of verification pipelines. First, we operationalize an audit that separates source fidelity, claim fidelity, and checker behavior while reporting both unsupported authorization and supported-claim retention. Second, we provide a matched comparison retaining intermediate judgments, first submissions, and field-level traces. Third, we localize source-role errors with an evidence-only substitution and document representation errors hidden by correct final labels. The attribution-audit framing is a retrospective synthesis; the original escalation criteria were not met.

2 Scope and Related Work

Our task is post-execution attribution: given a designated computation and a sentence about it, does the sentence preserve the computation's object and conditions? It differs from choosing an appropriate statistical method, executing arbitrary research code, or certifying a scientific proposition about the world. We explicitly do not evaluate all assumptions needed for calibrated p-values, causal identification, or adaptive hypothesis selection.

Fisher-R1 and P-Bench (Miao et al., 2026) evaluate hypothesis-testing agents that select and execute analyses from questions and data. Our narrower unit is the relation between an already executed result and a candidate report: whether the sentence preserves the object and conditions of its designated computation.

CogEEGAgent (Hou et al., 2026) uses typed scientific contracts and execution-grounded reporting, and distinguishes semantic choices from deterministic scientific authority. Its internally consistent wrong-template examples expose a related trust boundary. We examine that boundary using intermediate model judgments and a matched ordinary implementation of a finite attribution contract.

ClaimReceipt (Zhu & Chang, 2026) formalizes claim-relative evidence sufficiency and separates semantic evidence from transport integrity and experiment coverage. Building on these distinctions, we audit the representations used in scientific attribution decisions and intervene on the evidence side while retaining the extracted claim.

Related extraction and verification studies identify analogous representation failures. Diagnosing Structural Failures in LLM-Based Evidence Extraction for Meta-Analysis (Tan & D'Souza, 2026) studies role reversals, numerical misattribution, and cross-analysis binding errors. The alignment bottleneck in decomposition-based claim verification (Akhter et al., 2026) examines how evidence alignment affects decomposed verification. We ask whether measured pipeline gains remain attributable to the checker once intermediate judgments and a matched ordinary implementation are exposed, and whether errors can be localized without correcting both sides of the decision using reference answers.

RIEC-Agent (Li, 2026) studies persistent evidence-identity accounting and global error control under adaptive multi-agent search. The present study instead concerns post-execution attribution and component-level evaluation. The RIEC scope checker used here is a finite compatibility implementation, not a test of the prior shared-budget architecture or its statistical guarantees.

3 Verification as a Composition

3.1 The object of a verdict

Let T denote retained execution artifacts and s the original sentence. A structured reviewer jointly produces an evidence representation, a claim representation, and a provisional review label:

$$(\widehat{e}, \widehat{c}, \widehat{r}) = A(T, s), \qquad \widehat{y} = V(\widehat{e}, \widehat{c}). \tag{1}$$

A direct reviewer instead returns $D(T, s)$. We audit $\widehat{e}$ against T and $\widehat{c}$ against s separately; this is a decomposition of output obligations, not independently executed or information-isolated extraction stages. Both representations can be influenced by both inputs. We compare the model and checker verdicts to a reference attribution label defined for the designated test and original assertion.

The evidence representation includes the outcome, coefficient or contrast, null and alternative where relevant, numerical result, adjustment set, and fitting/evaluation role. The claim representation records what the original sentence asserts about those fields. The claimed rejection decision is extracted from the sentence, not recomputed from the p-value: otherwise an incorrect assertion could be silently rewritten into a correct one before checking.

Three questions must consequently be measured separately. Does $\widehat{e}$ agree with the execution artifacts? Does $\widehat{c}$ preserve the sentence? Does V implement the declared compatibility rule on the representations it receives? A correct answer to the third question does not establish either of the first two. Nor does a correct final label imply that every intermediate field is correct: an unrelated mismatch can yield the expected refusal despite another erroneous field. These are operational distinctions, not a proposed new theorem.

3.2 Bounded labels

SUPPORTED denotes a match to the designated test's object, conditions, and rejection statement. CONTRADICTED denotes the opposite rejection statement for that same object and conditions. INSUFFICIENT denotes evidence about a different coefficient, adjustment specification, or evaluation role. UNKNOWN denotes unavailable required evidence or an unrepresentable case. A scope mismatch does not establish that the underlying scientific proposition is false.

For the follow-up's prediction workflows, held_out requires evaluation rows to be disjoint from both scaler-fitting and prediction-model-fitting rows. The complementary train label includes preprocessing access even if the prediction model did not fit those rows. This is the declared experimental boundary; it does not certify preprocessing performed before a packaged dataset was released. A supported statement about a training-role computation is not thereby valid evidence of out-of-sample performance.

3.3 Matched checking and source substitution

The ordinary checker uses explicit type, identity, role, adjustment, and decision comparisons. The RIEC scope checker implements the same declared finite contract through a separate implementation. Both receive identical representations. Agreement is expected when both correctly implement

Table 1: Evidence stages and units. Repeated judgments are not pooled as independent tasks.

Stage	Actual unit and size	Question it addresses
Historical audit	Four archived runs; ten selected claim-axis items and six empty positions	Can numerical consistency coexist with incorrect attribution?
Contract tests	114 constructed implementation checks	Do both checkers implement the finite contract?
Exploratory review	19 sessions, 18 configurations; 42 controlled cards and five natural items	Which failures motivate a follow-up?
Frozen follow-up	24 assertions in 12 workflow-level pairs; twelve model runs	Where do observed errors and improvements arise?

this contract; it is an implementation-conformance and attribution control, not an independently surprising discovery. The comparison establishes whether ordinary checking suffices for the measured outputs, without comparing every feature of the original RIEC research program or proving global algorithmic equivalence.

The source-side intervention evaluates $V(e_{\mathrm{exec}}, \widehat{c})$, where e_{exec} is produced by the finite executor before candidate inference. The candidate claim is unchanged. Executor records are derived from performed computations and membership sets, not from a hidden desired verdict. Each sentence names its designated output, so the intervention does not solve open-ended evidence retrieval. It also does not implement a general parser for arbitrary scientific code or validate every assumption of the recorded analysis.

4 Study Design and Evidence Tiers

Table 1 separates the stages of evidence. Repeated judgments, paired assertions, and implementation tests are not pooled into a single benchmark denominator.

4.1 Historical motivation and executable contract checks

We inspect archived data-to-paper trajectories as motivation, retaining positive and negative cases. A fixed four-run audit provides 16 run-by-axis positions: ten contain selected assertions, of which nine are locally supported and one insufficient; six contain no applicable assertion. The intercept example in the introduction is an additional later statement, not an extra failure inserted into that first-claim denominator. A second case describes a sex-group mean difference as statistically significant although the displayed inferential tests concern prediction residuals against zero. Neither case estimates present-day model error prevalence, and the annotations have not received independent domain-expert adjudication.

A separate executable prototype supplies 114 contract checks, comprising 96 main items and 18 controls. Both ordinary and RIEC implementations return the expected labels on all items. These checks validate implementation behavior for constructed cases; they are not 114 successful LLM scientific investigations.

4.2 Exploratory model comparison

The exploratory study contains 19 registered sessions covering 18 distinct model/effort configurations. One configuration was repeated, and one session is partial. All sessions use the same 42 controlled cards and five historical source-consistency items. Direct review is correct on all 42 controlled cards in 15 of the 19 sessions. The denominator includes the partial session; incomplete answers are not silently dropped.

The exploratory Luna low session is a useful warning case: direct review authorizes 4 of 21 unsupported controlled assertions, whereas extraction followed by either checker authorizes 13. Its structured self-review authorizes 15; the checker suppresses two through UNKNOWN rather than correct scope labels, and also loses one supported assertion, reducing retention from 15 to 14. The rule layer does not cause all deterioration or provide a lossless repair. For Luna medium the corresponding

direct and checked false-authorization counts are 0 and 2. These are exploratory configuration-level observations, not stable frequency estimates or a controlled causal effect of schema size. Natural source-consistency review and finite-rule labels are different tasks; we do not assign natural accuracy to the rule outputs.

4.3 Frozen follow-up

The follow-up uses three analysis families crossed with four datasets packaged with scikit-learn (Pedregosa et al., 2011): iris, wine, breast_cancer, and diabetes. The families concern OLS coefficient identity, OLS adjustment identity, and Ridge preprocessing/model/evaluation membership. Twelve workflow executions yield 24 assertions, with a faithful and a controlled altered assertion per workflow. Some executions reuse an OLS fit with a different attribution question. The breadth is four data sources and three template families, not twelve independent scientific domains.

Data choices, computation seeds, alterations, and presentation order were fixed before numerical outcomes were inspected for this gate; outcomes were not screened for significance. For even dataset indices the altered assertion changes an object or role; for odd indices it changes the rejection statement. The resulting reference labels contain twelve supported assertions, six insufficient assertions, and six contradicted assertions. The expected balance is not supplied to candidates.

We test gpt-5.6-luna/low and gpt-5.5/low in three interfaces, each repeated twice. D produces a direct label and reason using an explicit checklist. F produces full evidence and claim objects, a provisional review label, and a reason. C produces shorter objects retaining the task-relevant scientific fields, together with the same review outputs. Public constants supplied by the compact adapter are fixed for all items; the adapter does not correct model-produced fields using the answer key. Each run can inspect the same scripts, outputs, and membership records. The maximum budget is twelve fresh candidate contexts and 288 judgments, with no adaptive addition of tasks or inference retries.

4.4 Metrics, thresholds, and execution audit

We report exact-label accuracy, unsupported assertions authorized, supported assertions retained, missing/format errors, UNKNOWN, and field fidelity. Counts are kept at their actual units: paired statements and repeated reviews are not independent studies. We do not infer token cost, parameter count, or compute equivalence from a low-effort setting.

The predeclared degradation criterion required F to authorize at least two more unsupported assertions than D in each repetition, with implicated workflows spanning at least two families. The evidence-intervention criterion required at least two fewer false authorizations in each repetition of an affected extraction arm, losing at most one supported assertion. Direct review was considered sufficient if it made zero false authorizations and retained at least eleven of twelve supported assertions in both repetitions. If both D and C met this standard while only F failed, the protocol classified the result as a full-schema/interface burden rather than the proposed broad mechanism. These are research triage thresholds, not statistical-significance tests. Each run also required at least 22 technically valid entries.

All twelve runs produced 24 valid entries. Local database and turn-context metadata agree with requested model/effort configurations, but do not independently authenticate serving weights. Fresh-context dispatch excluded inherited conversation turns, yet the application injected prior project memory into every candidate context. Tool-access audits found no reading of the new private answer key or other candidates' answers; this is not an operating-system isolation guarantee.

One Luna D run modified a label after its first successful write without receiving a score. Primary analysis reconstructs the first successful submission from the trace and preserves the later file separately in provenance. Two attempted out-of-scope notifications were rejected. The original hash manifest covers scientific artifacts and evaluators but omitted prompt/request files; their hashes are available in a runtime snapshot taken after the first batch launched and remain unchanged afterward. These limitations are part of the evaluation, not claims of perfect preregistration or complete blinding.

Table 2: Frozen follow-up results from first submissions. Each cell gives repetitions 1 and 2. D: direct review; F: full extraction; C: compact extraction. "Rules" denotes identical ordinary/RIEC outputs. Model self-review is shown separately so that upstream gains are not credited to the checker.

Configuration	Decision source	Correct /24	False auth. /12	Retained /12
Luna low	D direct	22, 22	1, 1	12, 12
	F model self-review	23, 24	0, 0	12, 12
	F + rules	24, 24	0, 0	12, 12
	C model self-review	22, 24	1, 0	11, 12
	C + rules	22, 24	1, 0	11, 12
	C + executor evidence + rules	24, 24	0, 0	12, 12
GPT-5.5 low	D direct	24, 24	0, 0	12, 12
	F model self-review	24, 24	0, 0	12, 12
	F + rules	24, 24	0, 0	12, 12
	C model self-review	24, 24	0, 0	12, 12
	C + rules	24, 24	0, 0	12, 12

5 Results: Attribution Before Improvement

5.1 The safety gain is already present before deterministic checking

Table 2 reports the principal follow-up comparison. Each pair lists repetition 1 and repetition 2. Exact accuracy is out of 24, false authorization out of twelve unsupported assertions, and retention out of twelve supported assertions. F/C ordinary and RIEC verdicts are identical; model provisional labels are retained as separate rows.

Replacing evidence also leaves all F scores and GPT-5.5 C scores unchanged at 24. There are no missing items, UNKNOWN outputs, or format errors in this follow-up. Thus safety is not obtained by discarding all supported assertions.

Luna's F workflow reaches the same task-set score as GPT-5.5 direct review, but this is not general capability equivalence. More importantly, the F model's own provisional review already makes zero false authorizations. The rules add one exact-label correction in the first F repetition, changing CONTRADICTED to the protocol's INSUFFICIENT for a role mismatch. They do not supply a previously missing safety gain in that arm. Crediting the D-to-F improvement specifically to RIEC would therefore misattribute the observed difference.

5.2 No additional specialized-checker decision gain

Ordinary and RIEC outputs agree on all 192 structured follow-up decisions and on all 192 decisions after executor-evidence substitution. This agrees with their identical behavior on the 114 constructed contract checks. The result is conditional on matched representations and the declared support domain; it does not establish equivalence on all possible inputs or across untested RIEC functions.

As expected for two implementations of the shared contract, ordinary checking suffices for the measured rule outputs. The more informative component comparison is the retained provisional judgment: in F, the model's own review already avoids false authorization. A baseline restricted to numerical tokens would omit the object, role, and adjustment information that makes ordinary checking effective; it would not be a matched comparator for this attribution task.

5.3 The frozen degradation hypothesis did not recur

Luna F has zero false authorizations in each repetition, compared with one in D; GPT-5.5 has zero in both. The prespecified F-relative-to-D degradation condition is therefore not met. The observations also do not support blaming a full schema for the follow-up errors: the full interface works, while the compact interface fails once.

The source intervention removes one false authorization in the first Luna C run and none in the second. It does not meet the repeated intervention threshold. The frozen aggregate decisions stop escalation for Luna because the proposed mechanism did not recur, and for GPT-5.5 because ordinary direct review is already sufficient under the gate's criterion. This failed gate is not converted into a successful method result by the present paper's framing. The attribution audit is a retrospective synthesis of what the comparisons show and what they rule out.

5.4 Final labels can hide representation errors

On the six audited evidence fields, 189 of 192 structured entries match executor references within the frozen numeric tolerance; all 192 entries match on the five audited claim fields. Two evidence mismatches belong to the same workflow's paired assertions in Luna C repetition 1. A third mismatch occurs in repetition 2 despite a correct final label. These counts concern declared fields, not unrestricted prose fidelity, and expose why final accuracy alone is an incomplete representation audit.

6 Tracing Individual Failures

Preprocessing access missed during direct review. In wf11, evaluation rows are disjoint from prediction-model-fitting rows but 200 evaluation rows appear in the scaler-fitting set. Both Luna D repetitions support the assertion that observations were unused in either fitting stage. The first explanation misdescribes the membership facts; the second acknowledges overlap but describes the sentence as claiming used-in-fitting observations, despite its original unused wording. The same wrong authorization therefore need not reflect precisely the same intermediate error. F correctly handles the assertion in both repetitions.

Evidence-only substitution repairs one paired role error. In wf09, all 97 evaluation rows intersect both fitting sets. Luna C repetition 1 labels the evidence held_out. The resulting pipeline supports the unsupported held-out assertion and refuses the supported training-role assertion. Replacing only the evidence record corrects both labels: false authorization falls from one to zero and retention increases from eleven to twelve. This is one source-role error affecting two paired statements, not two independent repair successes. The second C repetition and both F repetitions have no corresponding error.

Correct verdict, imperfect fields. In Luna C repetition 2, item_012 adds intercept to the adjustment list for an intercept test. The designated coefficient already differs from the coefficient claimed in the sentence, so the checker returns the correct INSUFFICIENT label despite the extra field error. A 24/24 verdict score does not certify that the intermediate record is suitable for every downstream use.

7 Discussion and Limitations

7.1 Evaluate the claim actually checked

The experiments favor a reporting discipline rather than a universal architectural prescription. A pipeline report should identify the original sentence, the structured claim supplied to a checker, the evidence fields actually grounded in execution, and the decision produced before and after checking. A matched ordinary implementation and the model's provisional label help distinguish gains attributable to the overall interface from gains introduced by the final rule layer.

Native executor records are attractive where a finite interface can supply trustworthy facts without requiring the model to redescribe them. Our result demonstrates one local use, not a repeated gain over full structured review. Native evidence also leaves claim interpretation unresolved. Future systems should not present a source-bound record as proof that arbitrary natural language has been correctly formalized.

A. Direct-review error: preprocessing access (wf11)
Evaluation ∩ model fit: 0 rows. Evaluation ∩ scaler fit: 200 rows.
Luna D supports the "unused in either stage" assertion in both repeats.
Full structured self-review already refuses it in both repeats.

B. Evidence-only correction: role mismatch (wf09)
Evaluation overlaps both fitting sets in all 97 rows.
Extracted role: `held_out`. Executor role: `train`.
Luna C repeat 1: false authorization $1 \to 0$; retention $11 \to 12$.
One workflow, two paired assertions; no gain in repeat 2.

C. Right label, wrong field (item_012)
Luna C repeat 2 adds `intercept` to the adjustment list.
A coefficient mismatch already yields `INSUFFICIENT`.
The final label is correct, but the intermediate evidence is not fully faithful.

Figure 2: Source-level diagnostic traces. The panels locate observed errors and the reach of one evidence-only substitution. They do not establish error prevalence or a repeated repair advantage. Model self-review and rule verdicts are distinct measurements.

7.2 A SUPPORTED ANSWER TO A DIFFERENT QUESTION

A separate prototype audit illustrates another boundary. Canonical templates restate the test actually performed in 57 eligible cases; all restatements are supported under the ordinary checker, but only three preserve the original requested testing object and conditions. This does not make the remaining statements useless. It does prevent supported restatement from being counted as completion of the original scientific task. We include the example as a historical metric caution, not as another result from the 24-item follow-up or a newly evaluated repair algorithm.

7.3 WHAT THESE RESULTS DO NOT ESTABLISH

The follow-up is small and constructed from three template families and four data sources. Labels target designated-test attribution, not general scientific entailment. Paired assertions, repeated dataset use, and two repetitions per interface do not support broad population-level frequency or significance claims. Historical cases motivated construction and cannot simultaneously function as an untouched confirmation set. The exact-label reference audit was AI-assisted and is not independent domain-expert adjudication.

Both evaluated configurations already perform strongly under some interfaces. The results cannot establish that current powerful models systematically fail in natural scientific workflows, or that a weaker configuration attains their general ability at lower cost. Interface changes combine representation requirements and review behavior; the design does not separately identify all prompting, output-length, or deliberation effects. Common application-injected project memory further limits clean-context interpretation.

The ordinary comparator deliberately receives the same scientific fields as the specialized checker. Equality is evidence against an additional measured decision gain here, not proof that all verification implementations are interchangeable. Similarly, a recorded computation may violate assumptions required for scientific inference; accurately recording it does not correct those assumptions. We make no new family-wise error guarantee, causal-validity claim, or end-to-end autonomous-discovery claim.

8 CONCLUSION

Scientific verification should state both its verdict and the object to which that verdict applies. Our audit separates execution evidence, the original claim, and checking logic, then compares them

using intermediate judgments and a matched ordinary implementation. The follow-up finds local errors but no additional specialized-checker decision gain, and does not reproduce its proposed extraction-degradation mechanism. These limits are part of the result. They show how to keep numerical consistency, representation fidelity, rule compliance, and original-task completion from being reported as the same accomplishment.

REPRODUCIBILITY STATEMENT

The accompanying anonymous supplement contains execution scripts, retained outputs, review prompts, first-write candidate submissions, per-item scores, and separate ordinary and RIEC checker implementations. Offline score verification does not call a model. Source replay is provided separately and depends on the documented numerical environment. Nine workflow input projections are bundled; three diabetes projections are reconstructed locally from the installed scikit-learn loader and checked against their original hashes. Appendix G identifies the artifact structure, provenance transformations, and the distinction between reproducing scores and rerunning hosted models. Raw application transcripts are not redistributed; their exclusion and the limits of the sanitized provenance summaries are explicit. No anonymous public website is required to use the attached package.

AI USE STATEMENT

Generative AI tools assisted in research framing, hypothesis refinement, experimental design, implementation, construction of controlled evaluation assertions, source and literature inspection, result analysis, English drafting, and manuscript preparation. The underlying tabular data are public benchmark datasets, not AI-generated observations. Model outputs were also the subjects of the evaluation. Labels and source audits are AI-assisted and have not undergone independent domain-expert adjudication. Verification reported in this paper consists of the source inspections, deterministic tests, score recomputations, and artifact checks explicitly described here; it is not a claim of independent human validation of all AI-assisted work. No AI-assisted mathematical proof is presented. Responsibility for the submitted content remains with the human author.

ETHICS STATEMENT

This study uses packaged public datasets and previously public scientific-agent artifacts; it collects no new human-participant data. The examples are diagnostic tests, not clinical recommendations. Errors are attributed to specific retained outputs, not generalized to the competence or intentions of their authors. No natural error prevalence or autonomous scientific-discovery improvement is inferred. Historical source excerpts are identified through locators rather than redistributed wholesale. Private account information and raw application conversations are excluded from the anonymous supplement.

REFERENCES

Mahmud Akhter, Federico Ruggeri, Iman Munire Bilal, Rob Procter, and Maria Liakata. The alignment bottleneck in decomposition-based claim verification. In *IJCAI–ECAI 2026*, 2026. URL https://wrap.warwick.ac.uk/id/eprint/206302/. Accepted version; repository record lists in press.

data-to-paper contributors. Supplementary material for "Autonomous LLM-driven research from data to human-verifiable research papers". GitHub repository, n.d. URL https://github.com/rkishony/data-to-paper-supplementary. Snapshot dcd6ff7b6217faec6e3ade4c87000eb4a523ca5d; accessed September 18, 2026.

Dengzhe Hou, Lingyu Jiang, Fangzhou Lin, and Kazunori D. Yamada. CogEEGAgent: Toward autonomous cognitive EEG analysis with grounded execution and selection-aware verification, 2026. URL https://arxiv.org/abs/2607.25045v1. arXiv preprint, version 1.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous LLM-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1), 2025. doi: 10.1056/AIoa2400555. URL https://doi.org/10.1056/AIoa2400555.

Jincheng Li. RIEC-Agent: Local evidential compliance does not compose under adaptive multi-agent search. SSRN preprint, 2026. URL https://doi.org/10.2139/ssrn.7298738.

Jiacheng Miao, Jin Mu, Guanhua Chen, and James Zou. Fisher-R1: Training LLM agents for reliable hypothesis testing, 2026. URL https://arxiv.org/abs/2608.07437v1. arXiv preprint, version 1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12 (85):2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

scikit-learn developers. Toy datasets. scikit-learn documentation, n.d. URL https://scikit-learn.org/stable/datasets/toy_dataset.html. Dataset provenance documentation; accessed September 18, 2026.

Zhiyin Tan and Jennifer D'Souza. Diagnosing structural failures in LLM-based evidence extraction for meta-analysis, 2026. URL https://arxiv.org/abs/2602.10881v1. Accepted at IRCDL 2026; arXiv version 1.

Peiying Zhu and Sidi Chang. ClaimReceipt: Verifying evidence sufficiency and coverage in agent evaluations, 2026. URL https://arxiv.org/abs/2609.01992v1. arXiv preprint, version 1.

A FINITE ATTRIBUTION CONTRACT AND REVIEW INTERFACES

The evaluated contract concerns a designated executed computation, not unrestricted logical entailment. All follow-up assertions name a particular output. The evaluator therefore does not solve open-ended evidence retrieval. The ordinary baseline checks explicit fields; the specialized implementation checks the same finite contract. Constants and transport identifiers are supplied by deterministic adapters, not recovered from the answer key. The complete review instructions are supplied as `binding_gate/public/COMMON.md`, `DIRECT.md`, `FULL.md`, and `COMPACT.md` in the accompanying supplement.

Table 3: Output obligations in the frozen follow-up. Scientific fields shown here are audited separately from transport identifiers.

Object	Fields and interpretation
Evidence	`target`, `contrast`, `estimate`, `pvalue`, `adjustment`, `evaluation_role`: what the designated computation actually did.
Claim	`target`, `contrast`, `reject`, `adjustment`, `required_role`: what the original sentence actually says.
Full-only fields	Scope, null value, alternative, alpha, computational-integrity flags; claim kind. These fixed values are supplied by the adapter in C.
Review	A provisional four-way label and an artifact-linked reason; D supplies only a label and reason.
Representability	F/C may explicitly return unavailable representations. All retained follow-up entries were representable and technically valid.

For an OLS coefficient, the adjustment set contains the other fitted predictors, excluding the tested coefficient. For an intercept test it contains every fitted predictor. For the prediction-loss test it is empty. All tests in this follow-up use zero nulls, two-sided alternatives, and $\alpha = .05$. The stated rejection decision is extracted literally from the sentence; recomputing it from the p-value would change the assertion being evaluated. For an extracted numeric value a and reference b, fidelity requires $|a - b| \leq 10^{-9} \max(1, |b|)$, rather than exact decimal-string matching.

The full interface instructs the reviewer to extract the executed result and the original assertion separately, preserve an incorrect assertion rather than correct it, account for both fitting sets, answer all items, and avoid running a checker. The compact interface retains the same task-relevant fields while removing public constants. The direct interface asks for the original assertion's label and reason, using the same checklist and source access. Thus D/F/C differ in output requirements and review behavior; the design is not a token-matched causal experiment.

B HISTORICAL SOURCE AUDIT AND SELECTION

The retained snapshot of the data-to-paper supplementary repository is commit `dcd6ff7b6217faec6e3ade4c87000eb4a523ca5d` (data-to-paper contributors, n.d.). Four archived runs are included in the fixed audit: A3 (health indicators), B3 (social network), D2 (treatment policy), and Ea2 (treatment optimization). Four audit axes per run create sixteen positions. Ten positions contain selected assertions, with nine scoped supported labels and one insufficient label; six are empty. This is not a random sample of papers, models, domains, or failures. Selection, source locators, and hash metadata are retained in the supplement; raw historical logs are not redistributed.

The D2 opening example concerns a later sentence outside the fixed first-claim count. The relevant regression row is `const`, with estimate 34.9 and p-value .0196, while generated prose associates the pair with policy change. The native numerical check reports seventeen values and no mismatches. This does not establish that every scientific review stage approved the statement. It also does not invalidate the earlier PPV-policy assertion, which remains a scoped positive case.

For Ea2, descriptive male and female means (10.3 and 10.1) are followed by a significance statement. The displayed inferential code tests prediction residuals against zero rather than the sex-group mean difference. The native number check reports nine values and no mismatch. The annotation is insufficient support for that significance statement, not a finding of no sex difference. These source

readings were AI-assisted; a second source inspection is not independent expert adjudication. Neither example was automatically discovered by the evaluated rules.

C Executable prototype and supported restatements

The prototype contract tests comprise twenty-four groups with four main cases each and eighteen controls, for 114 cases. The families are nonzero-null identity, omnibus-versus-contrast identity, and prediction-evaluation role. There are 48 executed evidence records and six metadata-ablation records. The execution count is 48 statistical tests, eight fitted prediction pipelines, and sixteen prediction calls. A pipeline's component fits are not counted as additional independent scientific experiments. Both checking implementations match the expected labels on all 114 constructed cases.

A separate canonical-restatement exercise identifies 57 eligible cases. Templates state the test actually performed, and all 57 resulting statements are supported by the ordinary checker. Only three preserve the original request's testing object and conditions. The other statements are not necessarily useless, but cannot be counted as successful completion of the original task. This experiment evaluates templates, not model-generated repair or end-to-end autonomous research.

D Exploratory sessions

Table 4 retains every initial session, including the incomplete Terra low run and the duplicate Sol xhigh configuration. Each controlled set contains 15 supported, 21 unsupported, and six reference-UNKNOWN cards. The five natural items assess source consistency; natural rule outputs do not receive an accuracy score. Session identifiers in the release are neutral aliases. They preserve row identity without exposing application account identifiers.

Table 4: All nineteen exploratory initial sessions. Accuracy is out of 42 controlled cards. FA: false authorizations out of 21 unsupported cards. S: structured model self-review. O/R: identical ordinary and RIEC outputs.

Configuration	Effort	D	S	O/R	D FA	S FA	O/R FA
sol	high	42	42	42	0	0	0
sol	medium	42	42	42	0	0	0
sol	low	42	42	42	0	0	0
sol	xhigh	42	42	42	0	0	0
terra	xhigh	42	42	42	0	0	0
sol	xhigh	42	42	42	0	0	0
terra*	low	26	29	13	0	2	0
terra	medium	42	42	42	0	0	0
luna	low	30	27	24	4	15	13
terra	ultra	41	42	42	0	0	0
terra	high	42	42	42	0	0	0
luna	medium	42	37	36	0	2	2
luna	high	42	42	42	0	0	0
luna	xhigh	42	42	42	0	0	0
GPT-5.5	low	42	42	42	0	0	0
GPT-5.5	medium	42	42	42	0	0	0
GPT-5.5	xhigh	42	42	42	0	0	0
GPT-5.5	high	38	42	42	0	0	0
sol	ultra	42	42	42	0	0	0

*Terra low is partial; missing initial answers are retained as failures. Two Sol xhigh sessions remain separate.

The original scoring framework permits a mechanical structured-format repair when required. The release retains any repaired phase separately, and the table uses initial outputs only. The Luna low rules suppress two structured-self-review false authorizations by returning UNKNOWN and lose one supported statement; they do not turn those refusals into correct scope labels. Incomplete answers and conservative abstentions must therefore not be interpreted as equivalent mechanisms.

E FROZEN WORKFLOW CONSTRUCTION AND STOPPING RULES

Four datasets are loaded using scikit-learn: iris, wine, breast cancer, and diabetes (Pedregosa et al., 2011; scikit-learn developers, n.d.). Dataset choice, projection, seeds, and alteration rules were fixed before candidate inference. The computation seed is 16092026+ dataset index; assertion presentation uses seed 2026091601. Numerical outcomes were not screened for significance. For even dataset indices the altered assertion changes an object or role; for odd indices it changes the rejection statement. Each workflow contributes a faithful and an altered assertion. The reference labels are twelve SUPPORTED, six INSUFFICIENT, and six CONTRADICTED, without exposing the expected balance to candidates.

Table 5: Twelve executed workflow instances. Each yields two assertions. The four datasets recur across families; this is not twelve-domain coverage.

Workflow	Dataset	Family	Designated output / role
wf01	Iris	coefficient	intercept
wf02	Wine	coefficient	malic_acid
wf03	Breast cancer	coefficient	intercept
wf04	Diabetes	coefficient	bmi
wf05	Iris	adjustment	sepal_width_(cm); other predictors: none
wf06	Wine	adjustment	malic_acid; other predictors: ash
wf07	Breast cancer	adjustment	mean_texture; other predictors: none
wf08	Diabetes	adjustment	bmi; other predictors: bp
wf09	Iris	role	Ridge loss difference; train
wf10	Wine	role	Ridge loss difference; held_out
wf11	Breast cancer	role	Ridge loss difference; train
wf12	Diabetes	role	Ridge loss difference; held_out

The coefficient and adjustment questions sometimes reuse the same OLS fit: wf01/wf05 and wf03/wf07 are paired in this way. Their assertions concern different attribution questions; they are not independent data sources. The role family records both scaler-fitting and model-fitting row identities. In particular, wf11 has zero model/evaluation overlap but 200 scaler/evaluation overlaps; wf09 has 97 overlaps in both. Dataset preprocessing before its packaged release lies outside the recorded boundary.

The frozen triage rules were:

1. At least 22 of 24 technically valid unique entries per run; otherwise preserve the run and mark the affected behavioral comparison inconclusive.
2. Repeated degradation requires at least two more false authorizations in F than D in each repetition, with implicated workflows spanning at least two families.
3. If both D and C have zero false authorizations and retain at least eleven supported assertions in both repetitions while only F fails, classify the issue as full-schema/interface burden.
4. Evidence substitution must remove at least two false authorizations per repetition of an affected extraction arm while losing at most one supported claim. It must not correct the candidate claim using reference answers.
5. If direct review is already sufficient, improvement over F does not establish necessity over direct review. If ordinary and RIEC remain equal, do not claim RIEC superiority.
6. No reproduced mechanism, isolated effects only, or disappearance under reasonable ordinary review stops escalation of this hypothesis. Incomplete evidence does not automatically authorize sample expansion.

All twelve runs met technical adequacy. Neither configuration met repeated degradation or the intervention threshold. The retained aggregate decisions are `STOP_ESCALATION_NO_REPEATED_MECHANISM` for Luna and `STOP_ESCALATION_SIMPLE_REVIEW_SUFFICIENT` for GPT-5.5. These rules are research-allocation criteria, not significance tests. The present paper-level attribution framing is a post-hoc synthesis, not a redefinition of a failed original criterion as success.

F EXECUTION DEVIATIONS, CONFIGURATION, AND PROVENANCE

The two configurations are recorded as `gpt-5.6-luna/low` and `gpt-5.5/low`. Local database and turn metadata agree with the requested settings. This records the serving request, not an independent authentication of backend weights. The effort labels do not measure parameter count, tokens, or cost. Fresh candidate contexts excluded inherited conversation turns, but the application injected prior project memory in all twelve runs. Tool-access inspection found no read of the new private key or other candidate answers; it did not enforce an operating-system access boundary.

One Luna D second-repeat output was changed after its first successful write, with item_002 changed from INSUFFICIENT to CONTRADICTED and no score feedback. Primary analysis reconstructs the first successful submission from the retained trace: 22/24 correct, compared with 23/24 in the later file. Both versions have one false authorization. The later file is preserved separately and does not replace the primary result. Two out-of-scope notification attempts were rejected and did not create successful messages or additional experimental runs.

The original scientific manifest covers 62 scientific/evaluation files and external checker dependencies, but omits prompt/request files. A runtime snapshot records the latter hashes after the first batch launched; final hashes agree with that snapshot. This is a preparation coverage gap, not full pre-inference hashing of all materials. Sanitized provenance in the supplement preserves the type and outcome of these deviations without distributing raw private application transcripts. The included manifests distinguish original source hashes from anonymized derivative hashes; a derivative hash is not presented as the original hash.

G ARTIFACT ACCESS AND REPRODUCTION BOUNDARIES

The attached archive is self-contained for offline rescoring of the retained candidate answers. It does not call hosted models and does not need account credentials. The supplement README gives the exact commands, dependencies, and dataset reconstruction conditions. The main directories are:

- `binding_gate/`: public task definitions and workflows, private reference records, first-write answers, later answer versions, per-item scores, and ordinary/RIEC checker code.
- `prototype/`: finite-contract cases and retained prototype metrics.
- `exploratory/`: nineteen neutral session aliases, initial metrics, controlled-card material, and separately retained format-repair artifacts.

The privacy transformation replaces personal filesystem roots and thread identities with portable aliases while retaining scientific values and first/later-output distinctions. Full historical source logs and personal application transcripts are excluded; locators and integrity information identify the relevant original sources. Natural source-consistency scores are retained as historical records rather than advertised as fully replayable from the redistributed subset. Tabular data sources and third-party notices are documented. Nine input projections are bundled with source attribution; the three diabetes inputs are reconstructed from the installed data loader and checked against their original hashes rather than redistributed with unresolved dataset-specific terms.

Three reproduction levels must be distinguished. Offline score recomputation tests the saved answers against the fixed rules. Numerical replay tests execution outputs using the documented dependencies and deterministic reconstruction. Re-running a hosted model requires access to the same configuration and orchestration context, which this archive cannot guarantee; it is not required for the first two levels. Original raw application logs are not released, so third parties cannot independently re-establish every context-isolation observation from this archive alone.