Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Valid Citations Are Not Enough: Auditing Scientific Claim Support Across Evidence Versions

Meshaal Obaid Mouawad

mm4922@msstate.edu

Mississippi State University

Research Article

Keywords:

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11587240/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When Valid Citations Are Not Enough: Auditing Scientific Claim Support Across Evidence Versions

Meshaal Obaid Mouawad

PhD Researcher

Department of Electrical and Computer Engineering, Mississippi State University

Mississippi State, MS 39762, USA

Correspondence: mm4922@msstate.edu

ORCID: 0000-0002-1152-8324

Abstract

Scientific claims can remain unchanged even as their supporting evidence changes, making their continued validity a question of both source provenance and interpretation. To examine how evidence availability and reference judgments shape that assessment, we audited a source-cited verification workflow that records source changes, claim-support states, and update consequences separately. The study combined six constructed numerical pairs with eight unchanged claims from four published papers evaluated against original and deliberately altered result tables. Paired-version inference was compared with separately executed single-version assessments, and frozen reference judgments were checked against the evidence actually supplied to each assessor. We then conducted a separate, post-exposure diagnostic using expanded published context and admissible source roles, with new references fixed before inference. In the numerical demonstration, both approaches detected the single loss of support without additional loss alerts. In the original published-table study, all 24 responses passed citation and input validation, yet some reference judgments relied on context absent from the inference packets. Against the original references, paired and single-version assessment identified two and one of four intended losses, respectively, with state agreement of 10/16 and 11/16. Rescoring the unchanged paired predictions against diagnostic references changed agreement to 12/16. In the expanded diagnostic, both approaches detected all three resolved losses and preserved all three supported controls; agreement with packet-specific AI references was 16/16 and 15/16. Causal and test/validation qualifications remained unresolved. These results show how evidence availability and reference selection affect the interpretation of version-specific claim support, even when source citations and input records are valid.

Keywords: scientific claim verification; evidence versions; research reporting; provenance; source context; reference judgments

1 Background

An unchanged scientific sentence can acquire a different evidential status after a result is corrected, a table is revised, or a source is supplemented. A file-change alert identifies an update; it does not establish whether the sentence remains justified. The researcher needs to know what the sentence asserts, what the evidence establishes at each version, and whether any unresolved qualification prevents a firm conclusion.

Consider a published comparison of memory-efficient training methods. In VeLoRA's main Table 1, SSF has a reported memory value of 4.13 and SSF with VeLoRA 3.46; their mean accuracies are 72.7 and 72.6%, respectively [1]. Those entries support the table-local statement of an approximately 16% memory reduction with a 0.1 percentage-point accuracy decrease. In our controlled altered table, replacing 3.46 with 4.50 contradicts that reduction, while an unchanged Hydra comparison retains its approximately 7% reduction and 0.1-point accuracy increase (Figure 1). Both claims depend on the changed table, but only one loses support. The altered value is an experimental evidence manipulation, not a new result reported by the original authors.

Table-local comparison	Original A	Controlled B
SSF memory: baseline / +VeLoRA	4.13 / 3.46	4.13 / 4.50
SSF accuracy: baseline / +VeLoRA (%)	72.7 / 72.6	72.7 / 72.6
Claim: 16% less memory; −0.1 points accuracy	Supported	Contradicted
Hydra memory: baseline / +VeLoRA	3.10 / 2.88	3.10 / 2.88
Hydra accuracy: baseline / +VeLoRA (%)	74.7 / 74.8	74.7 / 74.8
Claim: 7% less memory; +0.1 points accuracy	Supported	Supported

Figure 1: One changed table, two different claim consequences. Values derive from VeLoRA's main Table 1 [1]; only the SSF+VeLoRA memory entry changes in B. Relative reductions are $(4.13 - 3.46)/4.13 = 16.22\%$ and $(3.10 - 2.88)/3.10 = 7.10\%$. Memory values retain the original Table 1 presentation, labelled MB; Appendix Table 10 uses GB and different Hydra values. The source inconsistency is retained, and ratios are evaluated only within Table 1. The illustration represents actual case-study inputs.

The arithmetic is only part of the problem. A sentence may name a method through a symbol defined elsewhere, call a measurement "test performance" when a table says "validation accuracy," or attribute an observed difference to a mechanism. A verifier can quote every supplied number

correctly yet lack evidence for those qualifications. Reference reviewers can make the opposite mistake: they may use full-paper knowledge to label a sentence supported even though the model receives only an isolated table. Agreement with such references would not by itself validate the model's scientific inference.

Scientific claim verification already relates statements to supporting or refuting evidence. SciFact provides rationale sentences from abstracts [2]; SciFact-Open examines open-domain verification and specificity mismatches [3]. SciTab studies compositional reasoning over scientific tables [4], and SciVer combines multiple scientific evidence elements [5]. SciClaimEval retains scientific claims while modifying evidence and distinguishes table-only, short-context, and full-paper requirements [6].

Table-Text Alignment supplies supporting-cell annotations and an ambiguity taxonomy covering conversion errors, missing context, and unclear references; correct claim labels can coexist with poorly aligned evidence rationales [7]. SourceCheckup distinguishes valid source links from whether their content supports a statement [8]. Its URL-validity endpoint differs from the exact quotation/input checks used here. CLAIM-BENCH examines links to evidence dispersed across scientific papers [9]. These studies establish the general verification, context, and explanation problems that motivate our evaluation question.

Research-reporting guidance also treats verifiability as a distinct objective: the TOP 2025 framework distinguishes disclosure and sharing practices from independent checks of computational reproducibility and results transparency [10]. We examine how differences between an inference packet and the assumptions used to construct its reference judgments affect the interpretation of version-specific support decisions. LEAP implements the source-cited verification workflow examined here: it records evidence changes, support at each version, and consequences for an unchanged claim. Our contribution is a reproducible case audit of packet/reference information differences, including a change in agreement from 10/16 to 12/16 for the same saved paired predictions when the scoring reference changes. A separate same-case diagnostic examines which qualifications expanded packets resolve and which remain unresolved. We ask:

1. What support-loss and stable-support behavior is demonstrated by source-cited per-version assessment?

2. How do evidence completeness and reference construction affect the interpretation and scoring of that behavior?

3. What does paired-version assessment add relative to separately assessed single versions in the examined cases?

1.1 Position relative to prior work

Provenance and incremental computation address important parts of the update problem. Provenance semirings formalize how outputs depend on contributing inputs [11]. Adaptive functional programming tracks dependencies and propagates changes selectively [12]. Neither a lineage expression nor a recomputation guarantee establishes the meaning of a scientific sentence. These foundations motivate a record connecting source identity to semantic decisions and consequences across evidence states.

Table 1 compares the closest mechanisms and reporting standards by their actual units of analysis. The comparison motivates a case-specific audit of packet/reference differences and their consequences for support-state evaluation and composed update decisions.

Table 1: Closest work and the question examined by the present case audit. No external system was executed as an accuracy comparator.

Prior work	Established capability	Present examination
SciFact / SciFact-Open [2, 3]	Scientific support/refutation with rationales; evidence specificity	Explicit states at identified evidence versions
SciTab / SciVer [4, 5]	Table reasoning and multiple evidence elements	Complete literal qualifications versus supplied context
SciClaimEval [6]	Unchanged claims with altered evidence; context categories	Reference/packet disagreement despite exact citation checks
Table-Text Alignment [7]	Context/ambiguity taxonomy; verdict/evidence-rationale alignment	Fixed-prediction reference sensitivity and version-linked consequences
SourceCheckup / CLAIM-BENCH [8, 9]	Source support for attributed statements; dispersed claim–evidence alignment	Reference-packet sufficiency and version-linked decisions
TOP 2025 [10]	Reporting and verification practices for empirical claims	A reproducible record of evidence boundaries and decisions
Provenance semirings [11]	Composition of contributing-input lineage	Source identity separated from semantic support
Adaptive functional programming [12]	Dependency-driven selective recomputation	Update consequence separated from a change alert

2 Methods

The study comprised a constructed numerical demonstration, a table-only assessment and reference audit of authentic claims, and a separate diagnostic using expanded evidence packets (Figure 2). A populated study-specific reporting checklist is supplied separately alongside the three Additional

files.

2.1 Computational tools and AI assistance

Semantic assessments used OpenAI Codex with `gpt-6.1-sol` and high reasoning in separate paired and single-version sessions. AI systems prepared and separately checked reference judgments and audited source interpretations, as detailed below. Software execution and frozen validator replay produced the archived outputs; fixture responses were not substituted for model inference. Execution receipts retain model configurations, timestamps and accessible evidence. AI assistance also supported source recovery, analysis-code preparation, literature analysis, scientific auditing and manuscript preparation; a separate disclosure documents these tasks. Such records distinguish a statement of AI use from reviewable process documentation [13]. The author also used Elsevier LeapSpace for literature discovery, manuscript organization, language refinement and consistency checking, and takes responsibility for the article.

2.2 Verification record and decision rules

The inputs are an exact scientific claim, its original scope, a complete clause partition, identified evidence snapshots A and B, and admissible source roles. A snapshot is a preserved evidence state, not necessarily a historical publication revision. The default partition is the entire sentence as one clause; all eight authentic-source sentences retained that partition. The preparer cannot silently reduce a sentence to its easiest numerical fragment.

The workflow produces three separate outputs: observed source change; support at each version; and the consequence of the update (Figure 2). Source change includes original-byte differences, evidence additions/removals, and whitespace-normalized text differences. Text normalization is a change screen rather than a semantic equivalence test.

A separately executing language model returns exact source quotations, relations of support, contradiction, or non-addressing, scope fit, and reasons. For numerical comparisons, the assessor identifies the measurement, unit, comparison condition, and intended version and reasons over their relationship to the claim. The frozen semantic program validates citations and composes supplied relations; it does not itself measure scientific accuracy or establish numerical entailment. Source roles distinguish observations, analyses, methods, code, and other material. An inadmissible role or mismatched setting cannot establish support merely because its quotation is authentic. Role

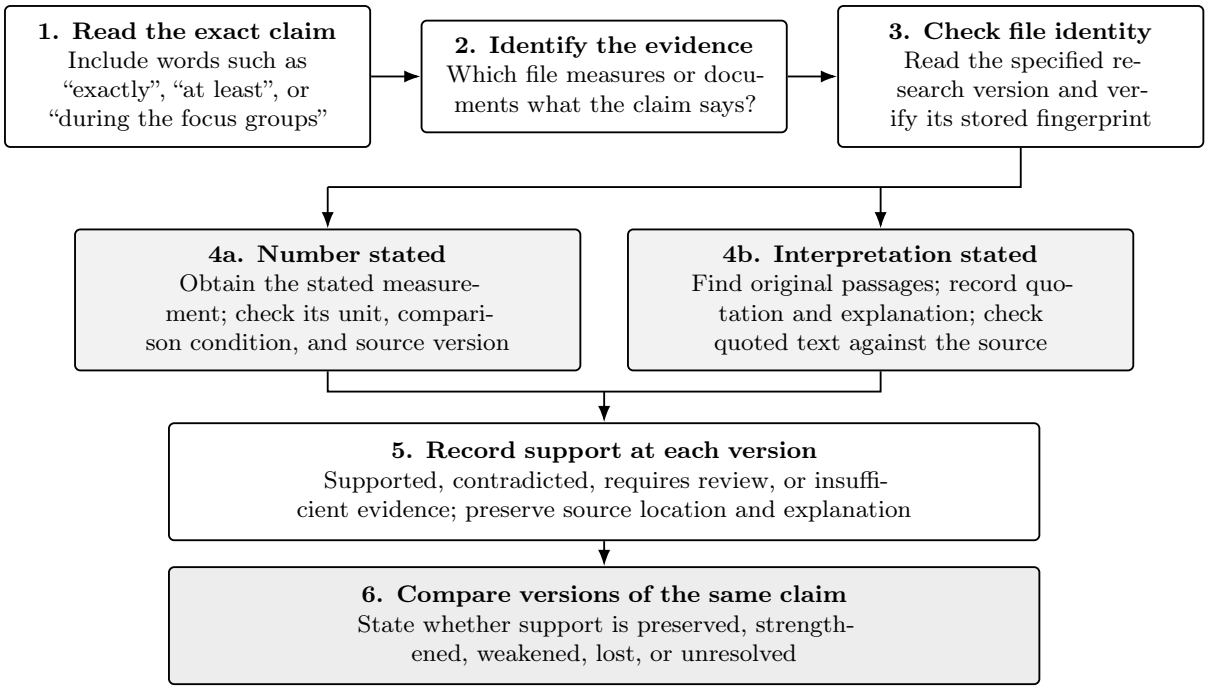


Figure 2: Version-specific verification procedure. Numerical claims require the asserted measurement and comparison; interpretive claims require relevant source passages and explanations. Source identity and quotation checks precede decision composition. The diagram describes workflow operations rather than measured accuracy.

bindings are preparer interpretations that require audit.

A separate program validates file availability, SHA-256 identity, exact line ranges and quoted text, request/method identity, source coverage, and clause coverage. Missing or altered required files block execution without a scientific label. When execution succeeds, admissible support without contradiction or uncertainty yields supported; admissible contradiction without support or uncertainty yields contradicted. Conflict or unresolved scope yields requires review; no admissible evidence yields insufficient evidence. An unread evidence item requires review. For an explicitly partitioned conjunction, a contradicted clause contradicts the conjunction; all clauses must be supported to support it. Because these authentic sentences were single clauses, their numerical and interpretive qualifications were assessed together.

Only supported-at-A followed by contradicted-at-B produces a proven-loss alert. Supported-at-A followed by review or insufficiency produces unresolved support, rather than proven invalidity. Supported states at both versions can be preserved, strengthened, or weakened. Strengthening requires a new supporting quotation; weakening requires removal of one. These are structural grounding rules, not a proof of semantic correctness. The program can establish which words were used and how supplied relations were composed; it cannot establish entailment from string

matching alone.

2.3 Materials and units of analysis

We retained two distinct experimental strata. A numerical demonstration used six constructed pairs from one source cluster: loss, preserved support, removed measurement, gained support, formatting-only change, and conflict. Arithmetic and missing/conflict rules fixed the references before inference. These were demonstrations of operational behavior, with no published-paper sample. The complete numerical experiment is retained in Additional file 1.

The authentic-source study used eight unchanged sentences from four published NeurIPS papers: GeoLIP [14], CLAP4CLIP [15], multiscale perturbed gradient descent (MPGD) [16], and VeLoRA [1]. Each paper supplied an intended affected claim and a comparison claim under the same table alteration. Six candidates were inspected and four selected purposively using recovered SciClaimEval development materials and original proceedings PDFs. Source PDFs, published wording, and table cells were checked before freezing; original transcriptions and any differences from benchmark wording were retained. No original research software was executed. Additional file 2 retains the complete frozen table-only study.

The experimental B records deliberately changed one GeoLIP sampling value from 175.24 to 50; CLAP4CLIP's CIFAR100 average for "Ours" from 86.13 to 80; all eight primary/variant MPGD validation accuracies to 60 while retaining gaps; and SSF+VeLoRA memory from 3.46 to 4.50. Other table bytes were preserved. These are controlled modifications of authentic result records, not historical revisions or measurements made by the original authors. The unit is one unchanged claim across two states; claims within a paper are correlated. Numerical demonstrations and authentic-paper cases were not pooled.

2.4 References, execution, and comparison

An AI preparer assembled arithmetic/source reference judgments, and a separately executing AI reviewer checked them before inference. Both had access to development records containing benchmark labels. References were frozen before subsequent LEAP predictions; no human expert adjudication was performed. Frozen references designated four losses, three supported stable controls, and one unresolved CLAP4CLIP control. A post-inference audit checked each literal qualification against the actual inference packet; original references and predictions remained unchanged.

One paired context received both versions. Two separate single-version contexts each received only its assigned version, duplicated into the validator's two snapshot slots for compatibility. Their duplicate transition labels were ignored; the separately produced A and B support states were composed into update consequences. The same method, prompt, role policy, citation rules, and recorded model configuration applied; no dated immutable backend snapshot was exposed. One first-pass prediction per condition/case was retained, with no semantic retries.

For the authentic-source runs, macOS Seatbelt restrictions denied access to references, previous experiments, source material outside the assigned packet, and other inference contexts. Failed sandbox setup attempts contained no assessments and were retained. Global Codex configuration/plugins remained loaded, so isolation was not hermetic. Numerical demonstrations used disjoint directories and instruction-based isolation. These differences prevent treating the strata as interchangeable replications.

We report support states, review/insufficiency, exact agreement with named references, loss alerts, and case explanations. For loss alerts, TP/FP/TN/FN denote true positives, false positives, true negatives and false negatives. Conditional confusion counts use the original seven-pair binary core and retain the unresolved eighth pair separately. A true negative means no loss alert on a reference non-loss; it need not mean correctly preserved supported states. Byte-change and normalized-text screens issue generic review alerts, not contradiction judgments. An always-supported comparator records the effect of predicting support everywhere. We made no population inference or significance claim from four purposively selected paper clusters.

2.5 Separate source-context diagnostic

A post-exposure diagnostic followed inspection of the original predictions and source audit (Additional file 3). It retained the same eight literal sentences, original tables, controlled B alterations, and one-clause scopes. Five original published views added CLAP4CLIP variant/comparator definitions, GeoLIP sampling and Lipschitz-function definitions, and MPGD parenthesized-variant and validation-gap mappings. Identical context was supplied at A and B. GeoLIP and CLAP4CLIP role bindings expanded from observation/analysis to include method evidence; the original MPGD caption supplied variant-identifying table metadata. Thus both context and admissible input roles changed, while the semantic implementation, assessor prompt, and composition rules remained unchanged. This was not a pure content-only intervention or a fresh holdout.

An AI source auditor, executing separately from inference and exposed to the earlier outcomes

but not the new diagnostic predictions, checked the complete literal claims, scope, context roles, and references against all actual requests before inference. Source/request hashes and references were then frozen. These new packet-specific references established three losses (GeoLIP, CLAP4CLIP, VeLoRA), three supported stable controls (GeoLIP, MPGD, VeLoRA), and two unresolved pairs: the CLAP4CLIP causal sentence and MPGD test-performance sentence. The latter remained unresolved because added validation definitions did not establish equivalence to test performance. The diagnostic used one new paired run and two new single-version runs with the recorded model configuration and file access restrictions.

Primary diagnostic counts used the six reference-established pairs; all eight pairs and sixteen unique states remained reported. A separate post hoc sensitivity analysis scored saved original table-only predictions against the new references. It did not replace original references or original reported scores. Neither this rescoring nor the new model calls isolate context from role bindings, reference changes, or inference variability. Raw first-pass outputs and source/explanation audits were retained with the diagnostic.

3 Results

3.1 Numerical behavior and the original table-only comparison

Both paired and single-version inference detected the one constructed loss and issued no loss alerts for the five other numerical pairs: TP1, FP0, TN5, FN0. Both matched 11/12 exact reference states; the paired record contained two review/insufficiency states. Missing/conflict controls are non-loss controls rather than supported scientific negatives. The two modes therefore demonstrated the same numerical loss behavior.

In the authentic-source study, all 24 first-pass response files passed the unchanged citation/input validator: eight paired files and eight files from each single-version context. Table 2 shows the paired states. Citation validity coexisted with unresolved semantic judgments in half of the paired states (8/16) and four of eight pairs.

Conditional on the frozen AI references, paired inference produced TP2, FP0, TN3, FN2; separately assessed single versions produced TP1, FP0, TN3, FN3 (Table 3). However, exact four-state agreement favored single-version inference, 11/16 versus 10/16. The modes agreed on 13/16 states. The apparent paired gain in loss alerts therefore coexisted with lower overall

Table 2: Original table-only paired results. A: original evidence; B: controlled altered evidence. "Review" abbreviates requires review. "Affected" identifies the intended altered-comparison claim, not an independently established loss. The last column reports its loss alert.

Paper	Affected claim A → B	Comparison claim A → B	Loss alert
GeoLIP	Review → review	Supported → supported	No
CLAP4CLIP	Supported → contradicted	Review → review	Yes
MPGD	Review → review	Review → review	No
VeLoRA	Supported → contradicted	Supported → supported	Yes

reference agreement. Paired non-alerts on intended losses were review outcomes, not affirmative supported-at-B conclusions. Of three intended supported controls, two remained supported; the MPGD comparison remained under review and still contributed a binary true negative.

Table 3: Original authentic-source comparison against unchanged frozen AI references. For loss alerts, TP: true positive; FP: false positive; TN: true negative; FN: false negative. A/B denote original/controlled altered evidence. The binary core contains seven pairs, excluding the unresolved CLAP4CLIP control; exact agreement includes all sixteen A/B states from eight pairs. Disputed supported-at-A references condition these counts. Change screens cover all eight pairs and issue generic review alerts.

Condition	TP	FP	TN	FN	Exact agreement (16 states)
Paired inference	2	0	3	2	10/16
Separate single-version composition	1	0	3	3	11/16
Always supported	0	0	3	4	10/16

Screening condition	Generic review alerts	Scientific contradiction output
Changed original bytes	8/8 pairs	None
Changed normalized text	8/8 pairs	None

3.2 What the source audit changed about interpretation

The audit found correct arithmetic and changed-cell identification alongside missing evidence for whole-sentence qualifications. For GeoLIP, the original ratios 185.18/175.24 and 425.04/306.98 are approximately 1.057 and 1.385; the altered first ratio is 3.704, exceeding 1.783. The paired assessor nevertheless retained review because the sentence also says Sampling provides a lower bound of the true Lipschitz constant. The table-only packet did not establish that property. Single-B treated the failed numerical qualification as sufficient for contradiction. This was a difference in whole-clause interpretation, not failure to notice the changed number.

For CLAP4CLIP, paired-A accepted the intended "Ours"/"Ours w/oVI" comparison, while single-A lacked a source mapping those rows to probabilistic and deterministic variants. The

published method paragraph supplies that mapping but was absent from the packet. Consequently, the extra paired loss alert did not isolate a benefit from reasoning over versions. The unchanged 1.46-point comparison did not establish the full comparison sentence's attribution to instance-conditioned prompt knowledge; review remained appropriate for that causal claim.

MPGD exposed two different issues. Its affected sentence says "test performance," whereas the table labels validation accuracy. Both modes identified that altered accuracies fell below the baselines, but the paired result retained the unresolved test/validation interpretation. Its comparison sentence names perturbations in equation (11); the PDF caption identifies the parenthesized variant, while the supplied transcription contains an unresolved `DS:sym` macro. Reference reviewers imported the original caption mapping although the model's packet omitted it. This was a packet/reference discrepancy, not an arithmetic error.

VeLoRA supplied the clearest within-table demonstration: both modes detected the affected SSF contradiction and preserved the Hydra claim. This finding is limited to main Table 1. Different Appendix Table 10 memory figures remain relevant to any unrestricted whole-paper claim. Finally, some unchanged headers/captions were tagged as contradicting links even though they only identified scope; the numerical rows supplied the counterexamples. Exact quotation checks did not establish every individual relation label.

3.3 Source-context diagnostic on the same cases

Both modes detected all three reference-established losses and preserved all three supported stable controls: TP3, FP0, TN3, FN0 in the six-pair diagnostic core. All 24 first-pass response files passed citation/input validation. Paired inference agreed with 16/16 packet-specific AI reference states; single-version inference agreed with 15/16 (Table 4). The modes agreed on 15/16 states. The one disagreement was MPGD affected/B: single-version inference returned contradiction while the reference and paired assessment retained review for the test/validation mismatch. Because A was already review, neither mode issued a loss alert for that sentence.

The paired context retained four review states and single-version inference three; each had review in two of eight pairs. Neither produced an insufficient-evidence state or invalid response. Generic byte/text screens still alerted on 8/8 pairs; always-supported agreed with 9/16 diagnostic reference states and detected none of the three established losses.

The changed decisions aligned with the supplied context: both modes resolved the Ge-

Table 4: Separate post-exposure context diagnostic. References were frozen against the expanded packets before inference. S: supported; C: contradicted; R: requires review. All state columns report A/B: original/controlled altered tables, each with the same added published context. "Affected" identifies the intended altered-comparison claim. The binary core contains three resolved losses and three supported controls; all eight claims/sixteen states from the same four paper clusters remain shown.

Paper	Claim	AI reference	Paired	Single-version
GeoLIP	Affected	S/C	S/C	S/C
GeoLIP	Stable control	S/S	S/S	S/S
CLAP4CLIP	Affected	S/C	S/C	S/C
CLAP4CLIP	Causal comparison	R/R	R/R	R/R
MPGD	Test-performance comparison	R/R	R/R	R/C
MPGD	Gap control	S/S	S/S	S/S
VeLoRA	Affected	S/C	S/C	S/C
VeLoRA	Stable control	S/S	S/S	S/S

oLIP sampling interpretation and MPGD parenthesized-variant mapping; single-A also resolved CLAP4CLIP's comparator identity. The diagnostic did not resolve the CLAP4CLIP causal claim or the MPGD test/validation distinction. The separate AI explanation audit identified the latter single-B result as a semantic boundary disagreement: its rationale reframed the sentence as a claim about the table's joint improvement while acknowledging validation-only evidence. Some single-A headers/captions received supporting relation labels as contextual members of multi-passage reasoning, rather than standalone entailments. Thus individual link granularity remained imperfect even when final states agreed. The diagnostic demonstrated feasible source-table-scoped decisions after a joint context-and-role intervention, rather than general whole-paper verification.

In the secondary sensitivity analysis, original saved paired predictions scored TP2, FP0, TN3, FN1 and 12/16 exact states under the new references; saved single-version predictions scored TP1, FP0, TN3, FN2 and 11/16. The difference between the original paired score of 10/16 and this 12/16 score resulted solely from changing the scoring reference, without changing predictions. The new diagnostic's 16/16 and 15/16 agreements therefore must not be presented as improvements over the original 10/16 and 11/16 under an unchanged evaluation contract.

4 Discussion

The examined workflow established selected controlled contradictions and preserved selected unaffected claims while retaining source-linked reasons. It also exposed a methodological problem in evaluating scientific verifiers: reference validity depends on the same evidence boundary as model assessment. A reviewer who sees the original paper may use a definition that an isolated

table does not contain. Scoring the model against that judgment conflates semantic reasoning with unequal information.

Three checks therefore answer distinct questions. Source validation asks whether the intended bytes and quoted words were used. Packet sufficiency asks whether those materials establish every necessary identity, definition, and qualification. Semantic review asks whether the cited evidence warrants the recorded relation. Our 24/24 citation-validation result answers the first question. The GeoLIP, CLAP4CLIP, and MPGD audits show why it cannot substitute for the latter two. Preserving these distinctions makes a disagreement inspectable. This contributes an operational record to the verifiability principles articulated in TOP 2025, using observed case-level disagreements rather than a general system-performance claim [10].

The missing mappings and imperfect link rationales accord with context and ambiguity findings in scientific verification, particularly Table-Text Alignment's distinction between a correct verdict and an aligned evidence rationale [3, 4, 6, 7]. Our additional observation concerns evaluation references: rescoring fixed paired predictions changed agreement from 10/16 to 12/16. That difference arose from the reference alone, without new model inference. The separate diagnostic showed that both modes could resolve the three supported-at-A losses after definitions, identity mappings, and admissible method roles were supplied. It did not establish that more context always helps: causal attribution and test/validation identity remained unresolved. A useful packet must preserve the particular context the claim requires, with explicit source roles, rather than merely increase text length. Reference review must inspect the actual packet and retain unresolved literal qualifications.

The paired-versus-single results do not establish an advantage. Numerical demonstrations yielded the same loss classifications and state agreement. In the original authentic-source experiment, paired inference raised one additional loss alert but agreed with fewer reference states. The CLAP4CLIP mapping discrepancy further confounded the apparent gain. In the context diagnostic, both modes detected the same three losses; their one-state difference concerned an unresolved claim. This is neither evidence of general superiority nor proof of equivalence; repeated matched inference on adequate references would be needed to distinguish those possibilities.

For research reporting, the practical contribution is a record that preserves exact claim wording, evidence version, admissible role, quotation, support state, update consequence, and the basis of any reference judgment. Such a record lets a later reader distinguish a numerical counterexample from missing context or an unresolved scientific qualification. Provenance and

selective recomputation provide foundations [11, 12]; the demonstrated addition is an auditable account of what the semantic decision did and did not establish in these cases.

5 Limitations

The authentic-source sample contains eight correlated claims from four purposively selected machine-learning papers. It is neither a random sample nor a fresh holdout. Controlled table alterations are not genuine historical author revisions. Numerical demonstrations contain one constructed source cluster and cannot enlarge the independent paper sample. Earlier conformance checks establish rule implementation rather than scientific accuracy.

Original references were AI-assisted, exposed to development labels, and later found to exceed the supplied context for several claims. Retaining them supports a transparent account of the original comparison, not a ground-truth accuracy claim. Diagnostic references were independently AI-checked against the expanded packets; they remain source-scoped AI judgments. The diagnostic was selected after exposure, reused every paper/claim, and changed content and role bindings jointly. It cannot isolate their effects or add independent cases. No independent human expert adjudication or matched external-verifier execution was performed. One inference per condition does not measure model variability; model aliases and loaded configuration limit exact inference replay. File access isolation does not remove the models' prior knowledge.

Single-clause treatment couples numerical and interpretive qualifications. Different defensible handling of a failed numerical conjunct and an unresolved qualifier can change the final state. Explicitly re-partitioning clauses could be studied, but would change this experiment's request contract. The workflow also starts from already identified claims and sources; it does not demonstrate exhaustive claim extraction or evidence retrieval. Broader resource or causal claims may require evidence beyond the scoped packets.

6 Conclusions

Source-cited per-version assessment detected selected controlled contradictions and distinguished them from unaffected claims. The audit traced packet/reference information differences and showed agreement changing from 10/16 to 12/16 for fixed paired predictions when the scoring reference changed. In a separate same-case diagnostic with added source context and admissible roles, both inference modes detected three established losses and preserved three supported

controls, while two claims retained unresolved qualifications. Scientific evaluation should preserve those distinctions and disagreements rather than treat citation validation as semantic validation. The evidence supports an inspectable workflow and bounded source-table-scoped feasibility; general loss-detection effectiveness and a paired-inference advantage remain unestablished.

Abbreviations

AI: artificial intelligence; CLI: command-line interface; MPGD: multiscale perturbed gradient descent; PDF: Portable Document Format; SHA-256: Secure Hash Algorithm, 256-bit; TP/FP/TN/FN: true positive/false positive/true negative/false negative. GeoLIP, CLAP4CLIP, SSF, Hydra, and VeLoRA are source-paper method labels retained in their original comparisons.

Declarations

Ethics approval and consent to participate

Not applicable. This study analyzed published scientific materials, software implementations and computational outputs. No human participants were recruited, and no human or animal participant data or tissue were studied. No institutional ethics approval or waiver is claimed.

Consent for publication

Not applicable. The manuscript reports no individual participant data, images or identifiable participant records.

Availability of data and materials

The materials supporting this article are included in the accompanying Additional files. They retain source PDFs and retrieval receipts, exact claims, A/B evidence and diffs, source/view hashes, frozen requests/references, first-pass responses, validation outputs, and source audits. Public original papers are identified in the references. The context diagnostic remains separate from the original experiment; Appendix A maps the preserved records.

Additional file 1. *Numerical experiment* (ZIP), `Additional_file_1_Numerical_Experiment.zip`: preserved numerical controls, frozen method, model outputs, and reproduction records.

Additional file 2. *Table-only study* (ZIP), `Additional_file_2_Table_Only_Study.zip`: unchanged authentic-source supplement, original references/predictions, and source-sufficiency audit.

Additional file 3. *Context diagnostic* (ZIP), `Additional_file_3_Context_Diagnostic.zip`: source-context packets, provenance, frozen packet-specific references, raw outputs, comparative analysis, independent audits, and replay records.

Additional file 4. *Reporting-applicability checklist* (PDF), `Additional_file_4_Reporting_Checklist.pdf`: completed study-specific checklist identifying applicable reporting items and their locations; no clinical-trial or systematic-review guideline is claimed.

Competing interests

The author declares no competing interests.

Funding

This research received no external funding.

Authors' contributions

Conceptualization, M.O.M.; methodology, M.O.M.; software, M.O.M.; validation, M.O.M.; formal analysis, M.O.M.; investigation, M.O.M.; data curation, M.O.M.; writing—original draft, M.O.M.; writing—review and editing, M.O.M.; visualization, M.O.M. M.O.M. is Meshaal Obaid Mouawad, the sole author. These roles include oversight and responsibility for the computational assistance described in Methods. No AI system, automated agent or software tool receives authorship or contributor status.

Acknowledgements

Not applicable.

References

[1] Miles R, Reddy P, Elezi I, Deng J. VeLoRA: Memory Efficient Training using Rank-1 Sub-Token Projections. In: Advances in Neural Information Processing Systems. vol. 37. Curran Associates, Inc.; 2024. p. 42292-

310. Available from: https://proceedings.neurips.cc/paper_files/paper/2024/hash/4a9eaf6dff3fdac9ab1aaf4c0fe2d563-Abstract-Conference.html.

[2] Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or Fiction: Verifying Scientific Claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2020. p. 7534-50. Available from: https://aclanthology.org/2020.emnlp-main.609/.

[3] Wadden D, Lo K, Kuehl B, Cohan A, Beltagy I, Wang LL, et al. SciFact-Open: Towards open-domain scientific claim verification. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics; 2022. p. 4719-34. Available from: https://aclanthology.org/2022.findings-emnlp.347/.

[4] Lu X, Pan L, Liu Q, Nakov P, Kan MY. SciTab: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2023. p. 7787-813. Available from: https://aclanthology.org/2023.emnlp-main.483/.

[5] Wang C, Shen Y, Kuang Z, Cohan A, Zhao Y. SciVer: Evaluating Foundation Models for Multimodal Scientific Claim Verification. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2025. p. 8562-79. Available from: https://aclanthology.org/2025.acl-long.420/.

[6] Ho X, Wu YA, Kumar S, Xia TC, Boudin F, Greiner-Petter A, et al. SciClaimEval: Cross-modal Claim Verification in Scientific Papers. In: Proceedings of the Fifteenth Language Resources and Evaluation Conference. Palma de Mallorca, Spain: ELRA Language Resource Association; 2026. p. 11060-71. Available from: https://aclanthology.org/2026.lrec-1.864/.

[7] Ho X, Kumar S, Wu YA, Boudin F, Takasu A, Aizawa A. Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers. In: Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics; 2025. p. 2509-17. Available from: https://aclanthology.org/2025.findings-emnlp.135/.

[8] Wu K, Wu E, Wei K, Zhang A, Casasola A, Nguyen T, et al. An automated framework for assessing how well LLMs cite relevant medical references. Nature Communications. 2025;16:3615. Available from: https://www.nature.com/articles/s41467-025-58551-6.

[9] Javaji SR, Cao Y, Li H, Yu Y, Muralidhar N, Zhu Z. Can AI Validate Science? Benchmarking LLMs on Claim →Evidence Reasoning in AI Papers. In: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. Mumbai, India: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics; 2025. p. 2355-79. Available from: https://aclanthology.org/2025.ijcnlp-long.127/.

[10] Grant S, Corker KS, Mellor D, Stewart SLK, Mayo-Wilson E, Cashin AG, et al. TOP 2025: An update to the Transparency and Openness Promotion Guidelines. Research Integrity and Peer Review. 2026;11:40. Comment. Available from: https://link.springer.com/article/10.1186/s41073-026-00223-0.

[11] Green TJ, Karvounarakis G, Tannen V. Provenance Semirings. In: Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Association for Computing Machinery; 2007. p. 31-40. Available from: https://doi.org/10.1145/1265530.1265535.

[12] Acar UA, Blelloch GE, Harper R. Adaptive Functional Programming. ACM Transactions on Programming Languages and Systems. 2006;28(6):990-1034. Available from: https://doi.org/10.1145/1186632.1186634.

[13] Cwik JC. Disclosure is not documentation: an open science framework for documenting generative AI use in scholarly research and publication workflows. Research Integrity and Peer Review. 2026;11:47. Methodology. Available from: https://link.springer.com/article/10.1186/s41073-026-00245-8.

[14] Wang Z, Prakriya G, Jha S. A Quantitative Geometric Approach to Neural-Network Smoothness. In: Advances in Neural Information Processing Systems. vol. 35. Curran Associates, Inc.; 2022. p. 34201-15. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/dd1322ce23cbbdd9d7ebb0ad1223c27a-Abstract-Conference.html.

[15] Jha S, Gong D, Yao L. CLAP4CLIP: Continual Learning with Probabilistic Finetuning for Vision-Language Models. In: Advances in Neural Infor-

mation Processing Systems. vol. 37. Curran Associates, Inc.; 2024. p. 129146-86. Available from: https://proceedings.neurips.cc/paper_files/paper/2024/hash/e94c57064dd5740c117b453bde8404c9-Abstract-Conference.html.

[16] Lim SH, Wan Y, Şimşekli U. Chaotic Regularization and Heavy-Tailed Limits for Deterministic Gradient Descent. In: Advances in Neural Information Processing Systems. vol. 35. Curran Associates, Inc.; 2022. p. 26590-602. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/hash/a9a238a00f7e885389fc02d65fb00994-Abstract-Conference.html.

A Provenance map

The unchanged semantic method is identified in the preserved records as R17, with eight source/protocol files and aggregate SHA-256:
10df807ba15a6999b5885cfcc29a43f11fe9cd814a8f6e73da59e42538b0ab87.

The following locators identify the distinct scientific records in Additional files 1–3.

The numerical demonstration is the resumed R29 Stage B record, described in R29_Scientific_Validation_Report.md; its six pairs form one constructed source cluster. The original authentic-paper study is the frozen LEAP_Paper4_R29_Authentic_Source_Supplement, whose SCIENTIFIC_REPORT.md preserves the reference-sufficiency disagreement. Original references, predictions, and interpretations remain unchanged. The separately identified post-exposure context diagnostic is stored under experiment/context_diagnostic; it neither replaces those records nor introduces new paper clusters.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile4ReportingChecklist.pdf
- Additionalfile1NumericalExperiment.zip
- Additionalfile3ContextDiagnostic.zip
- Additionalfile2TableOnlyStudy.zip