Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Scientific Claim Verification With Figures: A Visual Audit

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-31 · scientific claim verification · multimodal claim verification · cross-modal claim verification · scientific figure claim verification · chart claim refutation · sciver · sciclaimeval · scifact

Original article: https://reliablecontext.com/articles/can-a-figure-refute-a-scientific-claim



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes: What Counts as Evidence
4. Key Changes: What the Reader Sees and Where It Fails
5. Benchmark Comparison: SciFact, SciVer and SciClaimEval
6. Designing a Visual-Evidence Test
7. Data Analysis and Evidence
8. Implementation Considerations and Process Changes
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

A text-only test is adequate for scientific claim verification only when the decisive evidence is literally stated in prose. When a claim turns on an axis, a legend, a plotted value, an error bar, a table cell or a denominator, a text-only test cannot tell a reader that reads the evidence from one that guesses. The 2020 text baseline, **SciFact**, pairs about **1.4 thousand** expert-written claims with evidence-containing abstracts (huggingface.co), withholds its test labels (github.com), and verifies claims against a corpus of abstracts (arxiv.org). Figures and tables sit outside its scope by construction.

Two recent benchmarks move the decisive evidence into visuals. **SciVer** (Wang et al., 2025) reports **3,000** expert-annotated examples over **1,113** papers (arxiv.org), requires that claims not be verifiable from text alone, and reports that GPT-4.1 reaches **73.2%** against **93.8%** for human experts (arxiv.org). **SciClaimEval** (Ho et al., 2026) reports **1,664** annotated samples from **180** papers (aclanthology.org) and creates refuted claims by modifying the evidence, so that one claim is paired with two pieces of evidence (github.com). The authors state that, in the models tabulated in their results table, results on table-based evidence are higher than on figure-based evidence (arxiv.org).

These are the authors' reported findings in distinct settings. Raw scores must not be compared across the two datasets: they differ in labels, evidence supply, metrics and human protocols, and no independent reproduction of either headline result was found. SciVer's own error analysis, a **100-case** sample on one model, attributes **32%** of errors to failing to retrieve relevant information (arxiv.org) and 21% to visual element misinterpretation. Whether visual reading, evidence selection or claim reasoning dominates failure in other systems is an open research question.

The recommended design is a paired, synthetic fixture. Each item keeps one claim fixed and changes one controlled element of the figure or table: axis scale, plotted series, denominator or caption. The evaluator must name the exact evidence element (panel, axis, cell) and reproduce the arithmetic, and unreadable or ambiguous visuals are scored as unresolved rather than forced into a verdict. Scoring separates localization, interpretation, verdict, justification and abstention, with Wilson intervals (projecteuclid.org) and an **exact McNemar test** for paired comparisons (direct.mit.edu). In a hypothetical fixture of 60 pairs, a text-only reader

would score **40.0%** (Wilson 95% interval 28.6 to 52.6) against **76.7%** (64.6 to 85.6) for a visual reader, the kind of gap that decides whether a text-only test suffices. Figure-groundedness never proves that the source is true: it shows only that a verdict agrees with one rendering of the evidence.

Introduction and Background

Scientific claim verification asks whether a statement about research findings is supported or refuted by the evidence in a paper. Anyone who needs to verify claims in research papers faces a practical question before any modeling choice: is the decisive evidence in the prose, or in a figure or table? A text-only test cannot answer that question, and treating the two cases as interchangeable inflates confidence in systems that never read a plot.

This report defines the terms used throughout, compares three benchmark families, and proposes a test design that isolates visual reading. The terms are:

- **Claim:** a declarative statement about a result, such as "Method A improves F1 over Method B by at least two points on Dataset X".
- **Visual evidence element:** the smallest region of a figure or table that carries a quantity the claim depends on: a bar, a point, an error bar, a legend entry, an axis tick, or a table cell.
- **Caption:** the author-supplied text attached to a figure or table. It can state units, sample sizes and series definitions, and it can disagree with the plot.
- **Legend and axis:** the mapping from marks to series (legend) and from positions to values (axis, including scale and starting value).
- **Unit and denominator:** what a number counts, and what it is divided by. A percentage without a stated denominator is not a verifiable quantity.
- **Verdict:** supported, refuted, or unresolved. The third state is part of the design, not an error condition.

The question is live because the literature has begun to separate the text and visual cases. SciFact verifies claims against abstracts (arxiv.org). The SciTab authors criticize text-based benchmarks for "an over-reliance on text-based evidence" and release 1.2 thousand expert-verified claims paired with tables (arxiv.org). Human readers also misread visuals: in a survey of **473** respondents, many leading researchers had severe misconceptions about how error bars relate to statistical significance (pubmed.ncbi.nlm.nih.gov).

This report does not rely on keyword search-volume estimates, which are approximate figures from Google Ads data (docs.dataforseo.com).

The article describes published methods and reported results; it does not advertise a model or a use case, and it states where evidence is thin.

Key Changes: What Counts as Evidence

From abstract sentences to panels, axes and cells

SciFact derives its claims from citation sentences in the literature, and its refuted claims were written by an NLP expert who negated existing claims. The label set is Supports, Refutes or NoInfo, and the corpus is abstracts (arxiv.org).

SciVer and SciClaimEval change the object that carries the decision. SciVer's four reasoning subsets cover direct extraction of a single piece of information, integration across distinct sources, step-by-step inference across modalities, and application of scientific principles. The released repository lists tables and charts as images (github.com). SciClaimEval supplies an evidence file, either figure or table, along with its caption (arxiv.org).

The consequence for test design is concrete. A claim such as "the improvement exceeds two points" is decided by reading two bar heights against an axis, or two cells in a row. No sentence in the prose may state either number. A test that supplies only the paper's text therefore measures whether the prose repeats the quantity, not whether the system can read it.

From rewritten claims to edited evidence

Many verification datasets build refuted claims by editing the claim. The SciClaimEval authors observe that "many refuted claims simply rely on inserting negation words such as not" (arxiv.org). Their alternative keeps the authentic claim and edits the evidence. For tables the permitted edits include changing cell values; for figures they include flipping a graph or swapping legend text. The same claim text therefore appears twice, once with the original evidence (supported) and once with altered evidence (refuted), and the paper defines a pair as correct only if both samples are correct (arxiv.org). Because both twins must be right, the chance level for binary labels is 0.25 (**0.5 squared**).

SciVer takes a different route. Annotators write entailed claims, then generate refuted claims by "introducing factual errors that contradict the supporting evidence" (arxiv.org). The edit is applied to the claim, but the validation rule is about evidence: a claim should not be verifiable using textual context alone (arxiv.org). MuSciClaims combines the two ideas for figures in biology, chemistry and physics: supported claims are extracted automatically and manually perturbed to produce contradicted ones.

Editing evidence has a footprint. A fixture designer should expect that not every supported claim admits a meaningful, non-superficial edit, and should record which items could not be paired.

Key Changes: What the Reader Sees and Where It Fails

Evidence supplied versus evidence retrieved

SciClaimEval's input is a claim, an evidence file with its caption, and optional context. Claims are paired with evidence by regular expressions that find paragraphs mentioning a table or figure, and appendices are excluded. Evidence is supplied, not retrieved. SciVer's main benchmark is also context-grounded: the authors choose the setting that "removes variability introduced by retriever performance" (arxiv.org). Its separate

retrieval experiment exists because open-source retrievers do not natively support scientific tables and charts, and its Table 4 reports validation-set accuracies for original, retrieved, filtered and oracle evidence on two models.

A fixture must state which regime it tests. A system that is handed the right panel is measured on visual reading and reasoning. A system that must find the panel is also measured on selection. Mixing the two hides the source of failure.

Three separable error sources

SciVer's human error evaluation covers one model, **Qwen2.5-VL-72B**, on **25** randomly selected instances from each of four subsets (arxiv.org). The reported categories are failure to retrieve relevant information (**32%**), visual element misinterpretation (**21%**), failure in multi-step reasoning (17%), heavy reliance on the text modality (12%), and domain-specific misconceptions (10%). The five listed shares sum to **92%**.

Outside claim verification, CharXiv separates descriptive from reasoning questions on real arXiv charts and reports a "substantial, previously underestimated gap between the reasoning skills" of models and their descriptive skills (arxiv.org). ChartQA's error analysis attributes weaker non-bar-chart results to higher data extraction errors, especially for pie charts (aclanthology.org). The two sets of results point the same way: reading and reasoning fail separately, and a single verdict accuracy cannot tell them apart.

Verdicts must include "unresolved"

SciVer and SciClaimEval are binary: entailed versus refuted, supported versus refuted. SciVer's prompt instructs models to treat inconsistency, missing evidence or ambiguity as grounds to consider the claim incorrect (arxiv.org). That is a reasonable prompt for a benchmark with no unanswerable class, but it merges "cannot read the figure" and "claim is refuted" into one output. SQuAD 2.0 added more than **50,000** adversarial unanswerable questions so that systems must determine when a paragraph supports no answer and abstain (aclanthology.org); a strong system fell from **86%** to **66%** F1 (aclanthology.org). Abstention research frames the tradeoff as risk against coverage: abstain when in doubt while keeping coverage as high as possible, accepting that risk can be traded off for coverage.

The design rule that follows is simple: an unreadable or ambiguous visual must be an acceptable, scored outcome, and a confident verdict on an unreadable visual must be penalized.

Benchmark Comparison: SciFact, SciVer and SciClaimEval

Table 1 below summarizes how the three primary benchmarks, and two close neighbors, differ on the dimensions that matter for a visual-evidence test. Entries are the authors' reports, and the comparison is deliberately about design, not scores.

Benchmark	Evidence modality	Size	Labels	Claim and refutation construction	Evidence supply and humans
SciFact (Wadden et al., 2020)	Abstracts (text) (arxiv.org)	1.4 thousand claims (huggingface.co)	Supports, Refutes, NoInfo (arxiv.org)	Claims written from citation sentences; refutations by negating claims	Claims verified against a corpus of 5,183 abstracts (arxiv.org); test labels hidden (github.com)
SciTab (Lu et al., 2023)	Scientific tables	1.2 thousand claims (arxiv.org)	Not itemized here	Expert-verified claims from authentic publications	Models other than GPT-4 barely above random guessing (arxiv.org)
SciVer (Wang et al., 2025)	Text, charts, tables as images (github.com)	3,000 examples, 1,113 papers (arxiv.org)	Entailed, refuted	Entailed claims by annotators; refuted by perturbing them	Paper context supplied; 1,000 validation and 2,000 test (arxiv.org); 40 claims for humans (arxiv.org)
MuSciClaims (2025)	Figures from biology, chemistry, physics	1,515 data points, balanced over three labels (arxiv.org)	Support, neutral, contradict	Supported claims extracted, then manually perturbed	Best model **0.72 F1** (arxiv.org)
SciClaimEval (Ho et al., 2026)	Figures and tables, with captions	1,664 samples, 180 papers (aclanthology.org)	Supported, refuted	Authentic claim kept; evidence edited by annotators (github.com)	Evidence supplied; **80-sample** human baseline (aclanthology.org)

Three comparisons follow:

- **Modality:** only SciVer, MuSciClaims and SciClaimEval make non-text evidence decisive by design, and SciVer enforces it with a text-insufficiency rule.
- **Label space:** SciFact has an explicit no-evidence class, whereas SciVer and SciClaimEval are binary, so a binary benchmark cannot measure abstention without an added unresolved option.
- **Permitted context:** SciClaimEval distinguishes items that need no context from items for which the full text may be needed, while SciVer supplies a trimmed paper context.

Why raw scores cannot be ranked across datasets

The reported numbers use different metrics and different human protocols. SciVer reports accuracy: GPT-4.1 reaches 73.2% and humans 93.8%, the latter from 40 claims (arxiv.org). Its best model row, o4-mini, is listed at 77.7% on the test average in the paper's Table 3. SciClaimEval reports macro-F1 and pair accuracy: its final-version Table 5 lists o4-mini at test macro-F1 **75.4** and pair accuracy **53.3** on figures (aclanthology.org), against a human macro-F1 of **89.6** with **91.4%** agreement (aclanthology.org). The datasets draw on different fields: SciVer is primarily computer science, and SciClaimEval spans NLP, ML and medicine. No row of Table 1 can be turned into a ranking.

Version and release notes

The final published version of the SciClaimEval paper is the source for any quoted score in this report. Full data release is planned after the NTCIR conference (sciclaimeval.github.io).

Designing a Visual-Evidence Test

Build a synthetic fixture with one controlled change

Original synthetic drawings avoid image-rights questions (see Implementation Considerations) and give exact ground truth. Each base item has one claim and one figure or table. Each perturbed twin changes exactly one element:

- **Axis scale:** the same data plotted with a different starting value or a log scale, so the visual impression reverses while tick labels still disclose the truth. A bar chart whose y axis does not begin at zero can visually exaggerate effect sizes, and four deceptive visualizations built from distortion techniques were tested in a crowdsourced study (dl.acm.org).
- **Plotted series:** a legend swap or series swap, so that "Method A" labels the other bars. SciClaimEval's figure operations include swapping legend text.
- **Denominator:** a table whose percentage column is computed over a different n, so the same cell value implies a different count.
- **Caption:** a changed unit or sample size in the caption, with the plot untouched.

Pairing removes the claim as a confound. The pair is scored correct only if both twins are correct, with a chance level of 0.25 for binary labels. The structure has a known weakness: when pairs share identical claim text, a method can group items by claim text and assign the higher-scoring evidence to supported. A shared-task team describes a method that recovers the contrastive pairs (sciclaimeval.github.io), and the shared-task site notes that approach descriptions were given by the teams (sciclaimeval.github.io). A fixture should therefore vary claim wording between twins where meaning allows, shuffle presentation order, and include unpaired singletons.

Require an evidence pointer and a transparent calculation

Every answer must cite the exact evidence element and reproduce the comparison. The minimum record is an **evidence-element ledger entry**:

- **Source and version:** URL or file hash, version, observed date.
- **Locator:** figure or table number, panel, series name, axis, or row and column.
- **Literal claim:** the claim string as evaluated.
- **Extracted quantity:** the value read, with unit and denominator.
- **Calculation:** the arithmetic that links the extracted quantities to the claim, written as an expression.
- **Verdict:** supported, refuted, or unresolved, with a reason when unresolved.

(Hypothetical Example) A claim says "Method A reaches at least 2.0 points higher F1 than Method B in panel (b)." The ledger entry reads: locator "Figure 3, panel b, series A and series B, y axis labeled F1 (%)"; extracted A = 81.4 from the data label, B = 80.1 by reading the tick scale (gridlines every 0.5); calculation 81.4 minus 80.1 = 1.3; verdict refuted because 1.3 is below 2.0. If the data label were occluded and the gridlines absent, the extracted quantity would be unreadable and the verdict would be unresolved.

Reading values from plots has a quantified error. A study of digitizing a continuous-outcome figure reported percent differences from **0.23% to 30.35%** (pubmed.ncbi.nlm.nih.gov), and the authors recommend transparency when estimating data from figures (pubmed.ncbi.nlm.nih.gov). The Cochrane Handbook acknowledges that some needed numerical data are presented only in figures (training.cochrane.org) and describes a "cannot tell" response that can prompt contacting study authors for clarification (training.cochrane.org). A tolerance on extracted values, set per element before the evaluation, belongs in the protocol.

Score five things separately

A single accuracy conflates failures. The fixture should report:

1. **Visual localization:** did the answer point to the correct panel, axis, series or cell?
2. **Interpretation:** is the extracted quantity within tolerance, with the right unit and denominator?
3. **Verdict:** supported, refuted or unresolved, compared to the adjudicated label.
4. **Justification:** does the arithmetic reproduce, and does it use the cited element?
5. **Abstention:** on items built to be unreadable, did the system return unresolved, and on readable items, did it avoid needless abstention? Report risk and coverage together.

Localization and interpretation are the two scores that text-only tests cannot produce. SciVer's categories "failure to retrieve relevant information" and "visual element misinterpretation" are the closest published analogs to scores 1 and 2, but the published analysis is a 100-case sample on one model.

Human adjudication and disagreement

SciVer added a third expert to adjudicate disagreements between two annotators. SciClaimEval's authors report a human macro-F1 of 87.9 on tables and 89.6 on figures, with agreement rates of 86.7% and 91.4% (aclanthology.org).

For a new fixture, adjudicate every item with at least two readers, keep disagreements as data, and report agreement with a chance-corrected statistic. Percent agreement and kappa should both be reported (pmc.ncbi.nlm.nih.gov); kappa needs enough comparisons, with at least **30** as a heuristic (pmc.ncbi.nlm.nih.gov), and a small sample can yield a confidence interval wide enough to include almost any interpretation (pmc.ncbi.nlm.nih.gov). For multiple raters and missing data, Krippendorff's alpha works with any number of observers (asc.upenn.edu). An item on which adjudicators cannot agree about the figure is itself evidence that the figure is ambiguous, and belongs in the unresolved class.

Data Analysis and Evidence

The reported quantities fall into three groups: error decomposition, human-model gaps, and statistical uncertainty. *Table 2* below collects the decompositions that exist in the literature. They differ in model, sample and categories, so the table is a reading aid, not a pooled estimate.

Source and setting	Sample	Reported breakdown
SciVer error analysis, Qwen2.5-VL-72B	100 instances, 25 per subset	Evidence retrieval failure 32%; visual element misinterpretation 21%; multi-step reasoning 17%; text over-reliance 12%; domain misconceptions 10% (shares as reported)
CharXiv, descriptive versus reasoning	Real arXiv charts	Reasoning gap larger than descriptive gap
SciClaimEval, table versus figure evidence	1,664 samples	Authors state that tabulated models score higher on tables than on figures
ChartQA, chart type	**9.6 thousand** human-written questions (aclanthology.org)	Weaker non-bar-chart results attributed to data extraction errors (aclanthology.org)

No published analysis decomposes errors by modality across both visual-evidence benchmarks. SciVer's 100-case analysis is the only itemized breakdown for claim verification, and it covers one model. SciClaimEval's authors analyze by operation: Others, Graph Swap and Category Swap are harder than Graph Flip, and pair accuracy drops notably for tables even where macro-F1 is close to human. CharXiv supplies a perception-versus-reasoning split but on charts without claims. The question of which visual-reading, evidence-selection and claim-reasoning errors explain failure is not answered by these tables. A fixture of the kind described above is how a team would answer it for its own systems.

Human-model gaps and their uncertainty

The human baselines are small. SciVer's 93.8% comes from 40 claims; if two experts' answers are pooled as 80 judgments, 93.8% corresponds to 75 of 80 (an inference from the reported average, not a figure the authors give). SciClaimEval's figure baseline uses 35 samples. The **Wilson score interval** is preferable to the Wald interval for such proportions: its lower limit cannot be negative (itl.nist.gov), the Wald interval has erratic coverage (projecteuclid.org), and Wilson is recommended for small n (projecteuclid.org). *Table 3* below applies it to derived and hypothetical counts.

Quantity	Numerator / denominator	Estimate	Wilson 95% interval	Status
SciVer human experts, pooled judgments	75 / 80	93.8 %	86.2 to 97.3	Derived from the reported 93.8% on 40 claims
SciVer GPT-4.1, test	1,464 / 2,000	73.2 %	71.2 to 75.1	Derived from the reported 73.2% on a 2,000-example test set
Illustration only: a proportion of 88.6% on 35 items	31 / 35	88.6 %	74.0 to 95.5	Arithmetic illustration of sample size; the reported baseline is a macro-F1, not an accuracy
(Hypothetical) text-only reader, axis-scale pairs	4 / 15	26.7 %	10.9 to 52.0	**Hypothetical fixture**
(Hypothetical) visual reader, axis-scale pairs	12 / 15	80.0 %	54.8 to 93.0	Hypothetical fixture

Quantity	Numerator / denominator	Estimate	Wilson 95% interval	Status
(Hypothetical) text-only reader, series pairs	3 / 15	20.0 %	7.0 to 45.2	Hypothetical fixture
(Hypothetical) visual reader, series pairs	10 / 15	66.7 %	41.7 to 84.8	Hypothetical fixture
(Hypothetical) text-only reader, denominator pairs	5 / 15	33.3 %	15.2 to 58.3	Hypothetical fixture
(Hypothetical) visual reader, denominator pairs	11 / 15	73.3 %	48.0 to 89.1	Hypothetical fixture
(Hypothetical) text-only reader, caption pairs	12 / 15	80.0 %	54.8 to 93.0	Hypothetical fixture
(Hypothetical) visual reader, caption pairs	13 / 15	86.7 %	62.1 to 96.3	Hypothetical fixture
(Hypothetical) text-only reader, all 60 pairs	24 / 60	40.0 %	28.6 to 52.6	Hypothetical fixture
(Hypothetical) visual reader, all 60 pairs	46 / 60	76.7 %	64.6 to 85.6	Hypothetical fixture

The first two rows show how little a 40-claim human baseline pins down: an interval about **11 points** wide. The 2,000-example test interval is about **4 points** wide, so differences of one or two points between models on a single benchmark sit within sampling noise, before prompt or version differences. The hypothetical rows show the fixture's decision logic. A caption-only change is detectable from text alone (80.0% for the text-only reader) because the caption is text. Axis-scale, series and denominator changes are not: the text-only intervals at 15 pairs each are wide and overlap chance (25%), while the visual-reader intervals sit higher. The axis-scale intervals barely separate (text-only upper limit 52.0 against visual lower limit 54.8), while the series and denominator intervals overlap; only the pooled 60-pair contrast separates clearly.

Paired comparison on the same items

Both readers answer the same 60 pairs, so the comparison is paired. (Hypothetical Example) Suppose 20 pairs are correct for both readers, 26 only for the visual reader, 4 only for the text-only reader, and 10 for neither; the totals of 46 and 24 in Table 3 follow. McNemar's test uses only the 30 discordant pairs: an exact two-sided binomial test of 26 versus 4 gives p of about **0.00006**. McNemar's test was found to have low type I error when algorithms can be run once (direct.mit.edu), and significance testing in NLP is often ignored or misused (aclanthology.org). The decision rule is explicit: if the paired difference is indistinguishable from zero on the claim classes that matter, a text-only test is adequate for those classes; if it is distinguishable, it is not.

Implementation Considerations and Process Changes

Leakage, shortcuts and what is not yet reproduced

A visual-evidence benchmark can be solved without the visual. Schuster et al. found that claim-only classifiers perform competitively with top evidence-aware models on FEVER (aclanthology.org), and that performance of FEVER-trained models significantly drops on a debiased test set (aclanthology.org). The FEVER authors themselves reported **50.91%** when evidence was ignored (aclanthology.org). Hypothesis-only baselines are a recognized diagnostic for dataset artifacts (aclanthology.org). Every fixture should therefore include several arms:

- **Claim-only arm:** the model sees the claim and no evidence.
- **Text-only arm:** the model sees the claim, prose and caption, but not the figure or table.
- **Visual arm:** the model sees the rendered figure or table with its caption.
- **Matched table arm:** the same data as a table, where the visual arm is a chart.

Contamination is the second concern. Its extent is hard to measure (aclanthology.org), and it can overestimate the performance of a contaminated model on a target benchmark (aclanthology.org). SciVer's source papers were posted to arXiv between **1 September and 15 November 2024**, and SciClaimEval draws on public ACL, arXiv and PeerJ papers; public originals do not settle whether a given model saw them. A keyword search of the SciVer and SciClaimEval texts read for this report found no contamination analysis. Synthetic fixtures sidestep the issue because their figures never existed publicly.

What is not independently reproduced: both headline results come from the benchmark authors. SciClaimEval's repository offers scripts for the authors' own model predictions (github.com), and the shared-task site lists team-submitted results on the evaluation set (sciclaimeval.github.io), with accuracy on claim pairs as the primary metric (sciclaimeval.github.io). Those are not independent reproductions of the paper's 11-model test. The organizers released human-verified JSON tables after the paper (sciclaimeval.github.io). Both papers state their own limits: SciVer notes that annotation is labor-intensive and that some fields rely on other modalities; SciClaimEval has more table than figure samples and uses only PNG images for table evidence.

Rendering conditions

Rendering is an uncontrolled variable in the published designs. SciClaimEval renders tables to PNG. The papers read here do not state a figure resolution or cropping policy. A fixture should fix and publish four rendering conditions:

- **Resolution:** pixel dimensions and any downscaling before the model sees the image.
- **Color and contrast:** palette, and whether series are distinguishable without color.
- **Font size:** tick labels and legends at the delivered resolution.
- **Caption display:** whether captions are shown, hidden or altered.

Image rights

Published figures cannot be reused freely, and the licensing layers differ:

- **Dataset level:** SciClaimEval's dataset page states **CC BY 4.0** with the caveat that individual samples may have their own licenses (huggingface.co).
- **License terms:** under CC BY 4.0, attribution is required and the license may not give all the permissions needed (creativecommons.org); material may carry rights held by others (creativecommons.org), and modifications must be indicated (creativecommons.org).
- **ACL Anthology:** the CC BY 4.0 policy covers commercial and non-commercial use (aclanthology.org) but not third-party materials (aclanthology.org).
- **arXiv:** the common non-exclusive license grants no further reuse permissions directly (info.arxiv.org), and a reuser of a figure must contact the submitter or copyright holder (info.arxiv.org).

This report reproduces no published figure and recommends original synthetic drawings or figures with an explicit open license. This is not legal advice.

Accessibility, provenance and the ledger

A table alternative for a chart is also an evaluation aid: accessibility guidance suggests supplying the actual data in a table where possible (w3.org), which lets a fixture include a matched table arm. For the ledger itself, provenance is information about the entities, activities and people involved in producing a piece of data (w3.org), usable to assess its quality, reliability or trustworthiness (w3.org); the FAIR principles call for detailed provenance (nature.com). A plot digitizer exists for extracting values from chart images (automeris.io), but its error should be recorded, not hidden.

Implications and Future Directions

For researchers choosing a test, the evidence supports a conditional answer. If the claims to be verified are all decided by prose, a SciFact-style text test is adequate and cheaper. If any claim turns on a plotted quantity, a table cell or a denominator, a text-only test measures something else, and the paired fixture is the smallest design that shows the difference. Two independent construction strategies exist, editing evidence as SciClaimEval does (github.com) and requiring text insufficiency as SciVer does, and either is a defensible starting point.

Figure-groundedness is not source truth. A verdict that matches a rendered panel shows agreement with that rendering. The paper's plot may be wrong, the supplementary data may disagree, and the claim may concern something the figure does not show. The ledger therefore records what was read from which version on which date, and the unresolved class exists so that missing support is not reported as refutation. This is the provenance problem in miniature: information about the entities and activities involved in producing a value is what allows an assessment of its reliability (w3.org).

Three open questions remain:

- **Error dominance:** whether visual-reading errors dominate evidence-selection errors across models (SciVer's sample says evidence selection was the largest category for one model).
- **Pair structure:** whether pair-structured benchmarks overstate figure reading.

- **Transfer:** how well results carry over from curated paper figures to messy organizational evidence such as scanned reports and dashboards. No transfer study was found, so statements about transfer are hypotheses. One such hypothesis is that the evidence-element ledger above would need, for organizational evidence, the same fields it uses here: the value read, its source version, the observation date, and an explicit unresolved state when a visual cannot be read.

Frequently Asked Questions (FAQs)

What is multimodal scientific claim verification?

It is the task of judging a claim about a paper when the decisive evidence is a figure, table or chart rather than prose. SciVer is released with tables and charts as images (github.com); SciClaimEval supplies a figure or table with its caption.

What does cross-modal claim verification mean?

It means the claim is text and the evidence is another modality. SciClaimEval's title uses the term, and its refuted cases come from edited evidence (github.com).

Is there a benchmark for scientific figure claim verification?

Yes, several. SciClaimEval, SciVer and MuSciClaims are examples; ChartCheck covers **1.7 thousand** charts with **10.5 thousand** claims and explanations (aclanthology.org). They are not directly comparable.

Can AI check charts against paper claims?

To a limited, measured extent. On SciVer, GPT-4.1 reaches 73.2% against 93.8% for human experts; on MuSciClaims, the best model reached 0.72 F1. Results depend on the benchmark and cannot be generalized to a new corpus.

How do I verify claims against figures and tables?

Identify the element, extract the quantity with its unit and denominator, reproduce the comparison, and record the version and date. If the element is unreadable, record the claim as unresolved. Estimates from plots can differ materially from the underlying data (pubmed.ncbi.nlm.nih.gov), so record what the article reports separately from what was digitized (training.cochrane.org), and note the digitizer used (automeris.io).

How is multimodal fact checking of scientific papers different from text fact checking?

Text benchmarks such as FEVER draw evidence from Wikipedia sentences (aclanthology.org) and SciFact from abstracts (arxiv.org), and TabFact adds tables for **118 thousand** statements (arxiv.org). Visual benchmarks require reading marks, not sentences, and SciFact's own test labels are withheld (github.com).

What is chart-based claim refutation?

It is refuting a claim because a chart contradicts it. In SciClaimEval the contradiction is manufactured by modifying the evidence (github.com); in a real paper it must be discovered, which is harder to benchmark.

How should researchers verify claims in research papers in practice?

Use the ledger described above, at least two readers, a claim-only arm and a text-only arm, percent agreement alongside kappa (pmc.ncbi.nlm.nih.gov), and Wilson intervals on small samples (projecteuclid.org) (itl.nist.gov).

Conclusion

A figure can refute a scientific claim, and a text-only test cannot tell whether a reader noticed. The literature now contains text-oriented, table-oriented and figure-and-table-oriented benchmarks, each reporting a gap between models and humans in its own setting. What is missing is an independent decomposition of that gap into visual reading, evidence selection and claim reasoning.

The practical resolution is a paired, synthetic, ledger-based fixture with one controlled change per item, an evidence pointer for every answer, five separately scored behaviors, an explicit unresolved verdict, and interval estimates on every proportion. Such a test says whether a text-only evaluation suffices for a given claim class. It does not say whether the paper is right.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://huggingface.co/datasets/allenai/scifact
2. https://github.com/allenai/scifact
3. https://arxiv.org/abs/2004.14974
4. https://arxiv.org/html/2506.15569v1
5. https://aclanthology.org/2026.lrec-1.864
6. https://github.com/SciClaimEval/sciclaimeval-shared-task
7. https://arxiv.org/html/2602.07621v1
8. https://projecteuclid.org/journals/statistical-science/volume-16/issue-2/Interval-Estimation-for-a-Binomial-Proportion/10.1214/ss/1009213286.full
9. https://direct.mit.edu/neco/article/10/7/1895/6224/Approximate-Statistical-Tests-for-Comparing
10. https://arxiv.org/pdf/2004.14974
11. https://arxiv.org/abs/2305.13186
12. https://pubmed.ncbi.nlm.nih.gov/16392994
13. https://docs.dataforseo.com/v3/keywords_data/google_ads/search_volume/live
14. https://github.com/QDRhhhh/SciVer
15. https://arxiv.org/pdf/2406.18521
16. https://aclanthology.org/2022.findings-acl.177.pdf
17. https://aclanthology.org/P18-2124
18. https://arxiv.org/abs/2506.15569
19. https://arxiv.org/html/2506.04585v2

20. https://arxiv.org/abs/2506.04585
21. https://aclanthology.org/2026.lrec-1.864.pdf
22. https://sciclaimeval.github.io
23. https://dl.acm.org/doi/10.1145/2702123.2702608
24. https://pubmed.ncbi.nlm.nih.gov/28268241
25. https://training.cochrane.org/handbook/current/chapter-05
26. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052
27. https://www.asc.upenn.edu/sites/default/files/2021-03/Answering%20the%20Call%20for%20a%20Standard%20Reliability%20Measure%20for%20Coding%20Data.pdf
28. https://aclanthology.org/2022.findings-acl.177
29. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
30. https://aclanthology.org/P18-1128
31. https://aclanthology.org/D19-1341
32. https://aclanthology.org/N18-1074
33. https://aclanthology.org/S18-2023
34. https://aclanthology.org/2023.findings-emnlp.722
35. https://huggingface.co/datasets/alabnii/sciclaimeval-shared-task
36. https://creativecommons.org/licenses/by/4.0
37. https://creativecommons.org/licenses/by/4.0/legalcode.en
38. https://aclanthology.org/faq
39. https://info.arxiv.org/help/license/reuse.html
40. https://www.w3.org/WAI/WCAG22/Understanding/non-text-content.html
41. https://www.w3.org/TR/prov-overview
42. https://www.nature.com/articles/sdata201618
43. https://automeris.io/WebPlotDigitizer
44. https://aclanthology.org/2024.findings-acl.828
45. https://arxiv.org/abs/1909.02164

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.