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.

Citation language changes little after findings fail to replicate: a longitudinal analysis

Dongdong Guo
Shanxi College of Applied Science and Technology
Jiaxuan Li
xuanxuani@outlook.com

Shanxi College of Applied Science and Technology

Research Article

Keywords: replication, citation practices, research integrity, scholarly communication, epistemic certainty, large language models

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10681663/v1

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

Additional Declarations: No competing interests reported.

Citation language changes little after findings fail to replicate: a longitudinal analysis

Dongdong Guo[1] and Jiaxuan Li[1*]

[1]Shanxi College of Applied Science and Technology, Taiyuan, China.

*Corresponding author(s). E-mail(s): xuanxuani@outlook.com;

Abstract

Background: Later papers may keep citing a finding after a large replication has failed. We asked whether later papers respond by using more cautious language or mentioning the failure.

Methods: We examined 110 findings from three registered replication projects; 68 failed to replicate and 42 replicated. We collected 15,467 sentences from papers citing the original studies. A language model, given only sentence text, selected sentences that restated each finding and scored wording certainty. We compared the groups overall, over time, and around publication of the replication results. Inference used a claim-level bootstrap. Three other models and human annotators checked the measurements.

Results: Later papers used nearly the same certainty for findings that failed or succeeded replication. Mean certainty differed by 0.004; the 95% confidence interval was −0.011 to 0.019. Wording for failed findings did not become less certain relative to successful findings during the next eight to eleven years. The difference in before–after change was 0.0056; the 95% confidence interval was −0.0205 to 0.0303. Seven of 150 post-failure restatements mentioned the failure or conflicting evidence. Human annotators identified five. All four models gave the same main result.

Conclusions: In this mostly psychological sample, citation wording gave readers little indication of whether a finding had replicated. Failed replications were rarely mentioned, even years later. Linking original studies to replication outcomes may make contrary evidence easier to find for authors and editors.

Keywords: replication, citation practices, research integrity, scholarly communication, epistemic certainty, large language models

1 Background

Most readers encounter an earlier finding through a brief restatement in a later paper, not through a fresh reading of the original study. A registered replication failure should therefore become visible in later citation contexts, either as more cautious wording or as an explicit mention of the contrary evidence. Large replication projects allow this expectation to be tested across many findings evaluated under preregistered protocols [1–4]. These projects arose from concerns about false positives and analytic flexibility [5–7], selective publication [8, 9], and questionable research practices [10–12]. So far, their results have had limited effects on citation behavior. Findings that failed replication received more citations than findings that replicated, few later papers acknowledged the failure [13], and citation counts declined only modestly after the replication results became public [14]. Whether the wording of those citations became less certain has not been measured.

Researchers can identify some fragile findings before a replication is run. Prediction markets and expert surveys anticipate replication outcomes [15, 16]; text-based classifiers recover replicability signals at discipline scale [17, 18]; and the wording of original papers alone contains predictive information [19]. It is less clear whether researchers update the public description of a finding after direct replication evidence becomes available.

Earlier work shows how unreliable claims can remain in use. Errors have been traced through citation networks [20]; only 5.4% of post-retraction citations in one PubMed study acknowledged the retraction [21]; and quotation inaccuracies remain common, with little evidence of improvement over time [22]. Proposed remedies include citation-level annotation systems and interoperable retraction metadata [21, 23]. Failed replications present a harder case because they do not alter the status of the original article. Recent work describes irreproducible articles that continue to receive affirmative citations [24], while case studies find that qualifications are lost and contrary

findings ignored as claims are retold [25, 26]. Follow-up studies of the large replication projects have mainly counted citations [13, 14, 27]. We instead examine the words used to restate the findings, both before and after publication of the replication results.

These restatements matter because they are reused in reviews, teaching materials, and automated evidence summaries. If their wording does not respond to replication outcomes, replicated and nonreplicated findings will appear similarly well established to readers and language models.

For 110 findings from three registered replication projects, we retrieved the sentences in which later papers cite them and scored each sentence's stated certainty on a continuous scale. We retained sentences that restate the claim rather than cite the work as a method (Fig. 1; Methods). Throughout, we call a claim *refuted* when its registered replication failed and *robust* when it succeeded; these labels describe the projects' outcomes and do not imply that the underlying effects are absent (Methods). Because the projects, rather than the present authors, determined replication outcomes, group assignment did not depend on our assessment. We tested whether refuted claims were stated less confidently than replicated claims, whether publication of the replication result changed subsequent language, and how often post-refutation restatements mentioned the failure.

2 Methods

2.1 Claim set and replication status

The claim set comprises the original studies of three registered replication projects: the Reproducibility Project: Psychology [1], the Experimental Economics Replication Project [2], and the Social Sciences Replication Project [3], as compiled with replication statistics in the projects' public data [13]. A claim is labeled *refuted* when the project's replication yielded $p \geq 0.05$ or an effect opposite in sign to the original, and *robust* when the replication yielded $p < 0.05$ with the original sign; multi-laboratory

rows were aggregated to study level by median before labeling, and claims with genuinely ambiguous outcomes (a significant replication of an effect opposite to the original) were excluded rather than forced into a group. Labels therefore derive entirely from the projects' outcomes, not from any judgment of ours about the underlying science. We use *refuted* throughout as operational shorthand for a claim whose registered replication failed; the term denotes the replication outcome, not a demonstration that the original effect is absent, which a single failed replication cannot establish. A failure may reflect low power, moderators, or context sensitivity as well as a false original, and our normative expectation is correspondingly modest: that a published failure should register in the citing language at all, not that it should overturn the claim. Of 135 study-level claims, 110 resolved unambiguously to a Semantic Scholar record (title match with year tolerance ±2, disambiguated to the most-cited candidate, discarded when the best candidate had fewer than 20 citations); resolution rates were similar for refuted (80%) and robust (84%) claims, so resolution does not select on replication status.

2.2 Retrieving citing sentences

For each resolved claim we retrieved citing sentences with the publication year of each citing paper from the Semantic Scholar Academic Graph [28]. Because the interface returns the most recent citations first, records were sampled across the full available offset range so retrieved sentences span the claim's citation history; for the most heavily cited claims the interface's offset ceiling bounds coverage to the most recent ten thousand citing papers. Retrieval was conducted in July 2026; duplicate records were removed by year and sentence prefix, and sentences shorter than forty characters were discarded. Retrieval totaled 15,624 dated sentences, of which 15,467 were scored (median 127 per claim). Sentence counts per claim, retrieval windows, and citation-count distributions by group are reported in Supplementary Figs. 4, 5, and 10.

2.3 Scoring stated certainty

Each sentence was scored by a large language model (Qwen2.5-72B-Instruct, temperature 0) in a single two-part judgment. An *assertion* label identifies whether the sentence restates the claim itself, as opposed to citing the work as a method, paradigm, dataset, or generic background; only assertion sentences enter the analyses, which removes the dilution of paradigm citations. Assertion rates were 58% for refuted and 52% for robust claims (Supplementary Fig. 4); all certainty comparisons are made within assertion sentences of both groups. For assertion sentences, a continuous *certainty* score $s \in [0, 1]$ rates how definitively the claim is stated (anchors: 0 fully hedged, 0.5 neutral report, 1 definitive or causal), following the linguistics of hedging in scientific and medical prose [29, 30] and its computational detection [31, 32], and judging rhetoric rather than truth; the five anchors define an ordinal scale that we treat as interval, a standard simplification whose validity for our purposes rests on the human validation rather than on the numeric spacing of anchors. A stratified, blinded validation sample of 368 sentences, spanning the certainty range across both groups and including non-assertion items to test the filter and 12 blind duplicates, was independently rated by trained annotators under a written codebook; two raters completed it and met the pre-specified quality criterion, and their agreement statistics and the model–human correlation are reported in the Supplementary Information (assertion $\kappa = 0.65$, certainty $\alpha = 0.60$, model–human $\rho = 0.42$, filter accuracy 86%). The scorer receives only the sentence text, with no information about the cited claim's identity, its replication status, or the citing paper's year, so it is blind to group membership by construction; the full prompt is reproduced in the Supplementary Information. Common blinded scoring prevents direct differential scoring by known replication status, although imperfect measurement may attenuate a real contrast. As a robustness check every sentence was independently rescored with three models of unrelated architecture (DeepSeek-V3, DeepSeek-V4-Flash, and LongCat-2.0); each agrees with the

primary scorer on the assertion label for 85 to 88% of sentences, and each reproduces the between-group level null within its own scale (Supplementary Fig. 15).

2.4 Certainty drift

For each claim we grouped assertion sentences into adaptive year bins of at least eight sentences and estimated the yearly drift rate β of stated certainty from a weighted regression of $\ln(1 - \bar{s})$ on binned year anchored at the claim's publication year, $\beta = 1 - e^{\text{slope}}$; this parameterization follows a saturating drift model in which each retelling moves certainty a fixed fraction toward its ceiling. Claims entered the estimate only with at least three bins spanning at least four years and three distinct bin years (85 of 110). Group means were compared by cluster bootstrap over claims (1,999 resamples).

2.5 Event analysis

Each claim's refutation event is the publication of its project's results (2015, 2016, and 2018 for the psychology, economics, and social-science projects respectively). Claims with at least ten assertion sentences on each side of the event (73) entered a claim-level difference-in-differences: the change in mean certainty from pre- to post-event among refuted claims minus the same change among robust claims, equally weighting claims; inference by cluster bootstrap over claims. Event-study traces use claim-demeaned sentence scores in yearly bins from eight years before to eight after the event, requiring thirty sentences per bin. The social-sciences project (SSRP) contributes only four claims (two refuted, two robust) and carries no independent statistical weight; it is retained for completeness of the cohort, and the cross-project robustness claim rests on the psychology and economics projects, which have adequate claim counts.

2.6 Statistics

All group comparisons weight claims equally and use nonparametric cluster bootstrap over claims (1,999 resamples) for confidence intervals and two-sided p-values,

which respects the dependence of sentences within a claim and makes no distributional assumption about the scores. We interpret the principal results as bounded nulls rather than mere failures to reject: each two-sided 95% interval excludes a between-group certainty-level difference larger than 0.019 on the unit scale, a differential drift larger than 0.0103 per year, and an event response larger than 0.0303. These are precision bounds, not proof of exact equivalence, because the smallest effect of interest was not fixed before analysis. No multiplicity adjustment was applied: the intervals describe three distinct estimands and are not combined into a single familywise significance claim. Two properties of the scorer bear on how the bounds should be read. Applying the same blinded instrument to both groups prevents differential scoring by replication status, but it does not make a null immune to measurement error. Random error and compression at the neutral anchor can attenuate a true between-group contrast. Treating the correlation between the two independent model scorers (the primary Qwen2.5-72B scorer and DeepSeek-V3; Spearman $\rho = 0.54$, Supplementary Information S5 and Supplementary Fig. 15) as a rough reliability proxy widens the corresponding true-scale bounds to approximately 0.03–0.04. This calculation is a sensitivity analysis rather than a correction validated against ground truth; it leaves the observed contrasts small but makes the instrument's limited model–human correlation ($\rho = 0.42$; Supplementary Information S1) explicit in their interpretation.

2.7 Acknowledgment rate

From post-event assertion sentences of refuted claims we drew a stratified random sample of 150 and labeled whether the sentence mentions the replication failure, mixed or inconsistent evidence, or any controversy. The prespecified rate reported in Fig. 5 is the four-model majority label (7 of 150, 4.7%; Supplementary Information S8). Three independent human annotators subsequently labeled the same sentences while blind to the model labels and to 12 concealed duplicate identities. Majority vote classified

5 of 150 sentences (3.3%) as acknowledging the failure and agreed with the model on 146 of 150 (97.3%; Cohen's $\kappa = 0.65$). Cautious wording without reference to contrary evidence does not count as acknowledgment.

2.8 Use of generative AI in manuscript preparation

OpenAI ChatGPT and Codex were used to run language-model analyses, polish the manuscript's language, and modify code. The authors reviewed the outputs and take responsibility for the manuscript.

3 Results

We compare certainty levels and multi-year trajectories between refuted and replicated claims (Figs. 2 and 3), treat publication of each replication project as a discrete correction event (Fig. 4), and examine whether post-refutation sentences mention the failure (Fig. 5).

3.1 Measuring stated certainty at scale

We matched 110 of 135 original studies from the Reproducibility Project: Psychology [1], the Experimental Economics Replication Project [2], and the Social Sciences Replication Project [3] to unambiguous records in the Semantic Scholar Academic Graph [28]. We retrieved 15,467 dated citing sentences and scored each with a large language model in two steps: an assertion filter retained sentences that restate the claim itself, excluding method, paradigm, and background citations, and a continuous certainty score $s \in [0, 1]$ ranged from fully hedged (0), through neutral (0.5), to definitive or causal (1) phrasing (Methods). Figure 1 summarizes the design and instrument, and Table 1 describes the cohort. The median claim contributes 127 dated sentences, and post-refutation observation windows span eight to eleven years (mean 10.8). Of

the scored sentences, 8,598 passed the assertion filter and entered the certainty analyses. Among them, 31.6% restate the claim definitively ($s \geq 0.75$), 64.2% neutrally, and 4.3% cautiously ($s \leq 0.25$).

Three sentences from the corpus illustrate the scale. At $s = 0$: "People might feel farther away from future events with potential negative than positive implications." At $s = 0.5$: "recalling failures versus achievements has been shown to alter subjective distance to earlier selves." At $s = 1$: "People who expect to own a good in the future exhibit signs of psychological ownership for it." The score reads the rhetoric of the sentence, not the claim's identity or replication outcome. Adequacy rules based on sentence counts and time coverage leave 85 claims (34 replicated, "robust"; 51 failed, "refuted") for trend estimation and 73 for the event analysis. These rules require enough years to estimate a trend and observations on both sides of the event, avoiding estimates extrapolated from a single period.

Human validation of the scorer on a stratified blinded sample is reported in the Supplementary Information: two independent raters agree on the assertion label at Cohen's $\kappa = 0.65$ and on the five-level certainty scale at Krippendorff's $\alpha = 0.60$, both meeting the preregistered threshold. The model–human certainty correlation is only moderate ($\rho = 0.42$) and falls below the preregistered 0.6 threshold. Concentration of model scores at the neutral anchor partly explains the shortfall: excluding $s = 0.5$ ties raises ρ to 0.64. We therefore treat the score as a common, noisy instrument for within-instrument contrasts rather than as a calibrated substitute for human judgment; common measurement error may attenuate latent differences, a limitation incorporated into the bounds below (Methods).

3.2 Similar certainty for refuted and robust claims

If citation language tracked replication outcomes, refuted claims should be restated more cautiously. The data show no such difference (Fig. 2). Mean stated certainty is

0.568 for refuted claims and 0.564 for robust ones (difference +0.004, 95% CI −0.011 to +0.019), and the full sentence-level distributions are nearly coincident (Fig. 2B). The upper tails are also similar: 32.0% of sentences restating a refuted claim state it definitively ($s \geq 0.75$), compared with 30.6% for claims that replicated. The groups do not drift apart over time. Fitting each claim's certainty trajectory with a yearly drift rate β (Methods) yields group means of +0.0033 (robust) and −0.0001 (refuted); the bootstrap 95% confidence interval for the difference, −0.0036 to +0.0103, excludes any differential drift larger than one percentage point of the scale per year (Fig. 3A). Certainty is also unrelated to the strength of the replication evidence: across claims, drift rates show no dose-response relation to the replication p-value (Fig. 3B).

The null is not an artifact of analytic choices (Fig. 6). Group means are unchanged when the minimum bin size is varied (Fig. 6B) and when estimation is restricted to claims with well-determined trajectories (Fig. 6C); no individual replication project shows the pattern the pooled data lack (Fig. 6A); assertion rates are similar across groups (Fig. 6D); and the same absence of separation holds within each discipline separately (Supplementary Fig. 2). Individual claim trajectories tell the same story as the aggregates: the most heavily cited refuted claims track their robust counterparts year by year, with the refutation year invisible in both (Fig. 7).

3.3 No detectable change after publication of replication results

Each project's results were published at a known date, allowing us to compare the certainty of restatements before and after that event. Robust claims passed through the same publication date without being refuted and serve as the control. Across all 68 refuted claims, the dataset contains 3,476 restating sentences after the event and 2,123 before it, with post-event windows of eight to eleven years. At the claim level, every refuted claim contributes at least one retrieved post-event citing sentence and

at least one post-event restatement; none was observed only before its refutation. Post-event restatements have a median of 39 per refuted claim (interquartile range 15–75.5), although 12 claims contribute fewer than ten and therefore do not meet the event analysis's post-event adequacy rule. The difference-in-differences is estimated on the 45 refuted claims with at least ten restatements on each side of the event. Its estimate is +0.0056 (95% CI −0.0205 to +0.0303, cluster bootstrap over claims, $p = 0.65$), providing no evidence of the predicted decrease. The interval excludes an event response larger than 0.03 on the unit scale, one-tenth of the distance between a neutral and a definitive restatement. Figure 8 makes the bound concrete: a correction that moved one post-refutation restatement in five down a single anchor step, a shift of 0.05, would fall far outside the observed trajectory.

The event-study traces (Fig. 4A) show both groups moving together through the refutation year with no break at the event and no post-event divergence; the formal differential pre-trend is −0.003 per year, indistinguishable from zero, so the design's parallel-trends premise holds. The null is also consistent across projects: the difference-in-differences is +0.002 within the psychology cohort (39 refuted, 20 robust) and +0.002 within the much smaller economics cohort (4 refuted, 6 robust), whose interval is correspondingly wide, so the within-project agreement should be read as consistency, not as an independently well-powered replication (Fig. 4B).

The event-study sample requirement also does not select against claims that fell silent after refutation. Of the 68 refuted claims, the 23 that lack enough assertion sentences to enter the analysis are lost more often to a sparse *pre*-event record (11 claims) than to a sparse post-event one (6). As a group, they carry a higher share of their citing sentences after the event than the retained claims (0.65 versus 0.55), the opposite of the survivorship pattern in which a genuine citation collapse would remove the most-corrected claims (Supplementary Information, S6). Exclusion from the event analysis therefore does not appear to conceal claims whose citations collapsed after refutation.

3.4 Few citations acknowledge failed replication

The absence of a certainty response could reflect balanced reporting if citing authors flagged the failure while retaining the original phrasing. Only 4.7% of post-refutation sentences that restate a refuted claim mention the replication failure or any controversy (Fig. 5A); all four language models returned similarly low rates (Supplementary Information, S8). The remaining 95.3% do not mention the failure, and their certainty is indistinguishable from that accorded to claims that replicated. Independent human annotation of the same 150-sentence sample corroborates the estimate: the three-annotator majority labels 5 sentences (3.3%) as acknowledging the failure and agrees with the model label on 146 of 150 sentences (97.3%; Cohen's $\kappa = 0.65$; Supplementary Information, S1). This sentence-level estimate is lower than the earlier finding that 12% of post-refutation *citations* acknowledge the failure [13]. Among the 148 sampled sentences with matched certainty scores, the six that engage with the failure have a mean stated certainty of 0.42, compared with 0.58 for the other 142 sentences (Fig. 5B). This contrast provides a within-instrument positive control: on the same scale and corpus, the scorer registers the lower certainty of sentences that mention contrary evidence. The aggregate null therefore cannot be attributed solely to an instrument unable to detect cautious wording.

The result was not specific to the primary scoring model. Three additional models (DeepSeek-V3, DeepSeek-V4-Flash, and LongCat-2.0, from two other vendors) rescored all 15,467 sentences. They agreed with the primary model on the assertion label for 85 to 88% of sentences. Their absolute scales differed, with group means ranging from 0.60 to 0.70, but the within-model gaps between refuted and robust claims were again small: +0.0003, −0.002, and +0.004 (Supplementary Fig. 15).

4 Discussion

Across 110 claims and 8,598 restating sentences, expressed certainty was nearly the same for findings that did and did not replicate. The trajectories did not diverge, and the confidence interval excluded differential drift larger than about one percentage point per year. Publication of the replication results was followed by no detectable differential change. Previous studies found that nonreplicated claims continue to accumulate citations, with at most modest declines [13, 14]. The present results add that the wording of those citations changes little as well.

Unawareness is one possible explanation. The replication projects were published in prominent venues, but individual citing authors may not have seen them. Awareness could also remain low if each year's authors encounter the original claim without encountering the replication. The design cannot distinguish persistent unawareness from a decision not to discuss known contrary evidence. It does show that awareness did not diffuse enough to produce a gradual decline in certainty over eight or more post-event years.

The assertion filter is unlikely to explain the result because it retains hedged restatements and excludes only citations that do not restate the target claim. Scale compression is a more serious concern, especially given the concentration of scores at the neutral anchor. The binary acknowledgment analysis does not depend on fine differences in certainty, however, and it also found little response. Sentences that did mention contrary evidence were less certain on the same scale. Scientific prose can express the relevant caution; most of the sampled restatements simply did not engage with the failed replication.

One plausible mechanism is inherited phrasing. Authors may reuse the established description of a finding without reassessing all of its evidence. Earlier work describes cited papers as concept symbols rather than texts that are reread [33]; recurring misprints suggest that references are sometimes copied from other papers [34]; status

affects how papers are read and cited [35]; and distinctive wording can pass between articles [36]. If a replication is published separately and is not linked to the original record, familiar wording may persist even among authors who would have written differently had the contrary evidence been placed before them.

If citation language is copied rather than composed, its inertia may be a structural feature of scholarly communication rather than evidence about any author's private assessment. That distinction does not make citation language irrelevant. Citing sentences enter the public record that textbooks summarize, reviews aggregate, and language models learn from. Scientific self-correction does not require every author to re-evaluate every claim, but the public record should eventually incorporate strong contrary evidence. Our measurements show that this correction is rarely carried in subsequent restatements.

5 Limitations

The main limitation is measurement. The model–human correlation for certainty was $\rho = 0.42$, below the prespecified threshold of 0.6. Blinded use of the same instrument in both groups prevents scoring from responding directly to known replication status, but random error can still attenuate a real contrast. A sensitivity calculation based on inter-model reliability widens the estimated true-scale bounds to roughly 0.03–0.04 (Methods). The acknowledgment result is less dependent on fine-grained scoring: four models estimated 4.7%, and the three-annotator human majority estimated 3.3%. Chance-corrected agreement among the human raters was modest because positive labels were rare, so the human result is best read as corroboration rather than a replacement for the prespecified model estimate.

Citation generation is proxied by publication year, a coarsening that would blur a real drift but cannot manufacture a decade-scale null. It can, however, soften a sharp

break in the year or two after an event, because a paper published just after a replication failure may have been written before it; our event inference therefore rests on the absence of any gradual decline across the full eight-year post-event window rather than on a discontinuity at the event year itself. We measure the language of the citing literature, not of the original papers; whether refuted findings were already worded more confidently at the source is a separate question our design does not address. Our scale reads how confidently a claim is stated, not whether new boundary conditions are attached to it: an author who responds to a failed replication by qualifying the claim ("under condition X") rather than softening it would register as unchanged certainty, so the language of *refinement* is a dimension distinct from the language of confidence that our instrument does not capture.

The cohort is heavily weighted toward psychology (91 of 110 resolved claims); experimental economics contributes 15 claims and the Social Sciences Replication Project only four. The pooled estimate is therefore principally an estimate for psychology, while the economics and social-science splits are descriptive consistency checks rather than independent inferential replications. Whether the pattern extends to fields with different bias pressures remains open [37]. Low replicability is also documented in preclinical biomedicine [38–40], at estimated annual costs of tens of billions of dollars [41], but the present data do not establish that citation language behaves similarly there. Biomedicine's stronger retraction culture and stricter gatekeeping could produce a different pattern. The fields sampled here may also be a best case because awareness of the replication crisis is relatively mature. The refuted/robust grouping is itself a $p < 0.05$ dichotomy of the kind criticized in the replication literature; for this reason, the continuous dose-response analysis against replication p-values (Fig. 3B) is a more general test, and it also finds no gradient. Finally, the post-event windows span eight to eleven years (mean 10.8), so a change that takes more than a decade would fall outside the observation period.

6 Conclusions

Formal retraction and failed replication are communicated differently. Retraction notices are attached to article records, and retracted papers tend to lose citations and attention [42, 43]. A failed replication normally appears as a separate article. Unless an author searches for it or a database links the two records, the original article can be cited without any visible warning. The low acknowledgment rate in this study is consistent with earlier evidence that retraction and correction metadata often fail to reach citing authors [21]; the problem may be greater for failed replications, for which no standard status marker exists.

The confidence of a citing sentence is therefore a poor indicator of whether the underlying finding replicated. Evidence reviews and automated summaries should not treat confident restatement as evidence of independent confirmation. Language models trained on the published literature face the same problem: if replicated and nonreplicated findings are described in similar terms, the prose alone provides little basis for separating them.

Open-science reforms have concentrated on how new findings are produced and reported. Registered reports, for example, substantially reduce the proportion of positive results [44]. Those reforms do not by themselves update descriptions of findings already in circulation. Calls for better calibrated scientific prose [45] therefore need to be paired with systems that make later replication evidence visible at the point of citation.

Persistent, machine-readable links between original articles and replication outcomes are one practical response. Similar proposals have been made for retractions and inaccurate citations [21, 23]. A reference manager could display linked replication evidence when an author inserts a citation. A journal submission system could flag an unqualified restatement when a registered replication is already indexed, leaving the author and editor to decide whether it is relevant. Any implementation would need

transparent coverage, human review, and monitoring for false positives. The certainty and acknowledgment measures used here could serve as outcomes in evaluations of these interventions.

Supplementary information. Additional file 1 contains Supplementary Text, Supplementary Tables 1–7, and Supplementary Figs. 1–15.

List of abbreviations

LLM: large language model; RPP: Reproducibility Project: Psychology; EERP: Experimental Economics Replication Project; SSRP: Social Sciences Replication Project.

Declarations

Ethics approval and consent to participate. Not applicable. This study is a computational analysis of publicly available scholarly metadata and published citing sentences and did not involve human participants, human tissue, or animals.

Consent for publication. Not applicable.

Availability of data and materials. The datasets supporting the conclusions of this article are available in the Claim Drift GitHub repository, https://github.com/lijiajia3/claim-drift. They include the claim list and replication-status labels, 15,467 retrieved citing sentences, scores from all four language models, the annotation codebook, and de-identified human validation ratings. Citing sentences were retrieved from the Semantic Scholar Academic Graph and are redistributed with attribution under its ODC-BY terms. Replication outcomes are from the cited projects' public data. Analysis code and scripts that regenerate every figure and reported result are available in the same repository.

Competing interests. The authors declare that they have no competing interests.

Funding. This work received no specific funding.

Authors' contributions. D.G. and J.L. conceived the study and designed the analysis. D.G. implemented the retrieval and scoring pipeline, performed the statistical analysis, and wrote the first draft. J.L. supervised the project, validated the analysis, and revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements. The authors thank the maintainers of the registered-replication archives and the Semantic Scholar Academic Graph for making the underlying scholarly data available, and the annotators who completed the blinded validation samples.

Authors' information. Not applicable.

References

[1] Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716.

[2] Camerer CF, Dreber A, Forsell E, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016;351(6280):1433–1436.

[3] Camerer CF, Dreber A, Holzmeister F, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour. 2018;2(9):637–644.

[4] Klein RA, Vianello M, Hasselman F, et al. Many Labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science. 2018;1(4):443–490.

[5] Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005;2(8):e124.

[6] Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011;22(11):1359–1366.

[7] Gelman A, Loken E. The statistical crisis in science. American Scientist. 2014;102(6):460–465.

[8] Rosenthal R. The file drawer problem and tolerance for null results. Psychological Bulletin. 1979;86(3):638–641.

[9] Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. Science. 2014;345(6203):1502–1505.

[10] John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science. 2012;23(5):524–532.

[11] Kerr NL. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review. 1998;2(3):196–217.

[12] Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. Nature. 2005;435:737–738.

[13] Serra-Garcia M, Gneezy U. Nonreplicable publications are cited more than replicable ones. Science Advances. 2021;7(21):eabd1705.

[14] Clark CJ, Costello T, Mitchell G, Tetlock PE. Failing to replicate predicts citation declines in psychology. Proceedings of the National Academy of Sciences. 2023;120(29):e2304862120.

[15] Dreber A, Pfeiffer T, Almenberg J, et al. Using prediction markets to estimate the reproducibility of scientific research. Proceedings of the National Academy

of Sciences. 2015;112(50):15343–15347.

[16] Altmejd A, Dreber A, Forsell E, et al. Predicting the replicability of social science lab experiments. PLoS ONE. 2019;14(12):e0225826.

[17] Yang Y, Youyou W, Uzzi B. Estimating the deep replicability of scientific findings using human and artificial intelligence. Proceedings of the National Academy of Sciences. 2020;117(20):10762–10768.

[18] Youyou W, Yang Y, Uzzi B. A discipline-wide investigation of the replicability of psychology papers over the past two decades. Proceedings of the National Academy of Sciences. 2023;120(6):e2208863120.

[19] Herzenstein M, Rosario S, Oblander S, Netzer O. The language of (non)replicable social science. Psychological Science. 2024;35(8):887–899.

[20] van der Vet PE, Nijveen H. Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. Research Integrity and Peer Review. 2016;1:3. https://doi.org/10.1186/s41073-016-0008-5.

[21] Schneider J, Woods ND, Proescholdt R, et al. Reducing the inadvertent spread of retracted science: recommendations from the RISRS report. Research Integrity and Peer Review. 2022;7:6. https://doi.org/10.1186/s41073-022-00125-x.

[22] Baethge C, Jergas H. Systematic review and meta-analysis of quotation inaccuracy in medicine. Research Integrity and Peer Review. 2025;10:13. https://doi.org/10.1186/s41073-025-00173-z.

[23] Hosseini M, Eve MP, Gordijn B, Neylon C. MyCites: a proposal to mark and report inaccurate citations in scholarly publications. Research Integrity and Peer

Review. 2020;5:13. https://doi.org/10.1186/s41073-020-00099-8.

[24] Treves A, Chinn S, Krofel M. Irreproducible research and a typology of replication efforts. Research Integrity and Peer Review. 2026;11:19. https://doi.org/10.1186/s41073-026-00203-4.

[25] Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. BMJ. 2009;339:b2680.

[26] Hardwicke TE, Szűcs D, Thibault RT, et al. Citation patterns following a strongly contradictory replication result: Four case studies from psychology. Advances in Methods and Practices in Psychological Science. 2021;4(3):1–14.

[27] Schafmeister F. The effect of replications on citation patterns: Evidence from a large-scale reproducibility project. Psychological Science. 2021;32(10):1537–1548.

[28] Kinney R, et al.: The Semantic Scholar Open Data Platform. ArXiv:2301.10140.

[29] Hyland K. Hedging in scientific research articles. Amsterdam: John Benjamins. 1998;.

[30] Salager-Meyer F. Hedges and textual communicative function in medical English written discourse. English for Specific Purposes. 1994;13(2):149–170.

[31] Medlock B, Briscoe T. Weakly supervised learning for hedge classification in scientific literature. Proceedings of ACL. 2007;p. 992–999.

[32] Farkas R, Vincze V, Móra G, Csirik J, Szarvas G. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. Proceedings of CoNLL: Shared Task. 2010;p. 1–12.

[33] Small HG. Cited documents as concept symbols. Social Studies of Science. 1978;8(3):327–340.

[34] Simkin MV, Roychowdhury VP. Read before you cite! Complex Systems. 2003;14:269–274.

[35] Teplitskiy M, Duede E, Menietti M, Lakhani KR. How status of research papers affects the way they are read and cited. Research Policy. 2022;51(4):104484.

[36] Horbach SPJM, Halffman W. The extent and causes of academic text recycling or 'self-plagiarism'. Research Policy. 2019;48(2):492–502.

[37] Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. Proceedings of the National Academy of Sciences. 2017;114(14):3714–3719.

[38] Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery. 2011;10:712.

[39] Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012;483:531–533.

[40] Errington TM, Mathur M, Soderberg CK, et al. Investigating the replicability of preclinical cancer biology. eLife. 2021;10:e71601.

[41] Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biology. 2015;13(6):e1002165.

[42] Lu SF, Jin GZ, Uzzi B, Jones B. The retraction penalty: Evidence from the Web of Science. Scientific Reports. 2013;3:3146.

[43] Peng H, Romero DM, Horvát EÁ. Dynamics of cross-platform attention to retracted papers. Proceedings of the National Academy of Sciences. 2022;119(25):e2119086119.

[44] Scheel AM, Schijen MRMJ, Lakens D. An excess of positive results: Comparing the standard psychology literature with registered reports. Advances in Methods and Practices in Psychological Science. 2021;4(2).

[45] Hoekstra R, Vazire S. Aspiring to greater intellectual humility in science. Nature Human Behaviour. 2021;5:1602–1607.

Table 1 Cohort, corpus, and analysis samples by replication project. Study-level claims are compiled from the three projects' public data; a claim is *resolved* when it matches a unique Semantic Scholar record (title match with year tolerance ±2, disambiguated to the most-cited candidate, discarded below 20 citations). *Refuted* denotes a failed registered replication and *robust* a successful one; the split is fixed by the projects, not by us. *Assertion sentences* restate the claim itself rather than citing the work as a method, paradigm, dataset, or background, and are the input to every certainty analysis. The event year is the publication of each project's results, used as the correction event.

	Psychology (RPP)	Economics (EERP)	Social sci. (SSRP)	All
(a) Replication projects				
Project results published (event year)	2015	2016	2018	n/a
Study-level claims compiled	112	18	5	135
Claims resolved in Semantic Scholar	91	15	4	110
Refuted (failed replication)	60	6	2	68
Robust (successful replication)	31	9	2	42
(b) Citing sentences				
Scored sentences	12,357	2,550	560	15,467
Assertion sentences	7,086	1,162	350	8,598
Median sentences per claim	121	193	138	127
(c) Analysis samples (claims)				
Trend analysis	71	10	4	85
Event analysis	59	10	4	73

Adequacy rules (Methods): a claim enters the trend analysis with at least three year bins spanning at least four years, and the event analysis with at least ten assertion sentences on each side of its event. Retrieval totaled 15,624 dated sentences, of which 15,467 were scored.

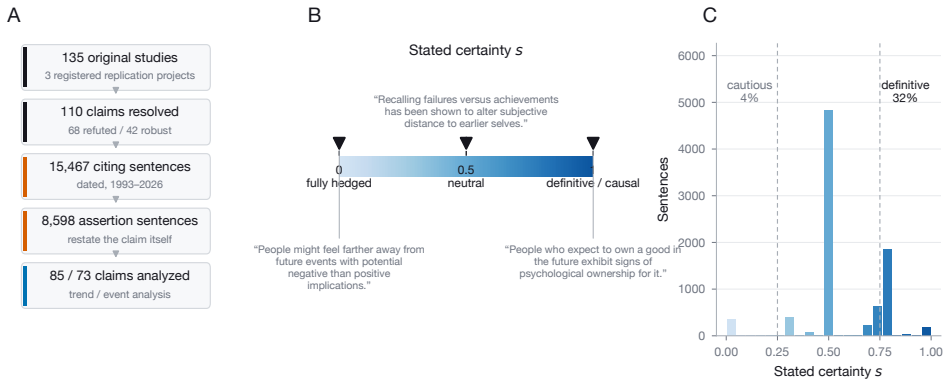


Fig. 1 Design and measurement. (**A**) Cohort construction: original studies of three registered replication projects, resolved in the Semantic Scholar Academic Graph; replication status is assigned by the projects, not by us. (**B**) The stated-certainty scale with anchor definitions and three verbatim corpus sentences at $s = 0$, 0.5, and 1. The scorer reads the rhetoric of the sentence, never the identity or truth of the claim. (**C**) The rhetoric of the citing literature: distribution of s across all 8,598 assertion sentences; 32% restate their claim definitively ($s \geq 0.75$), 4% cautiously ($s \leq 0.25$).

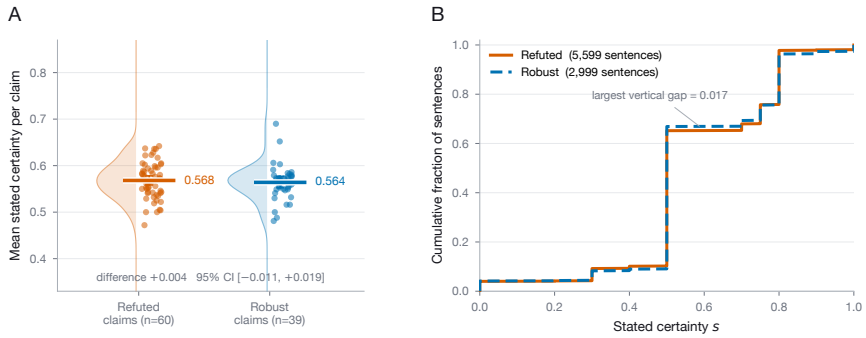


Fig. 2 Refuted and replicated claims are restated with the same certainty. (**A**) Mean stated certainty of restating sentences for each claim, by replication status; horizontal bars are group means (0.568 refuted, 0.564 robust; difference 95% CI −0.011 to +0.019). (**B**) Cumulative distributions of sentence-level certainty for the two groups; the curves are nearly coincident.

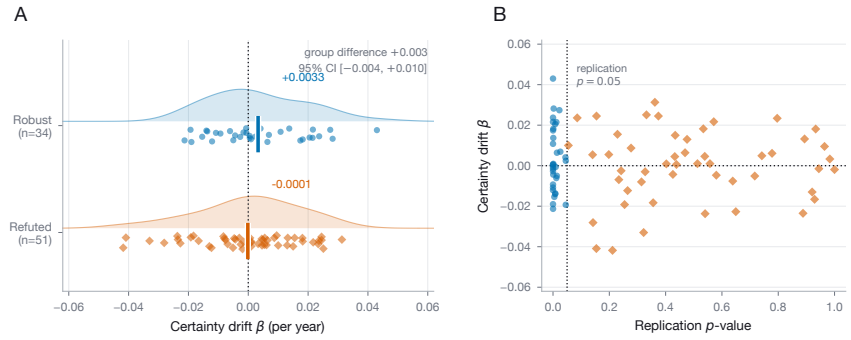


Fig. 3 No differential drift, and no dose-response to the replication evidence. (**A**) Per-claim yearly drift rate β of stated certainty by replication status; vertical bars are group means. (**B**) β against the replication study's p-value; the dotted line marks the projects' significance criterion.

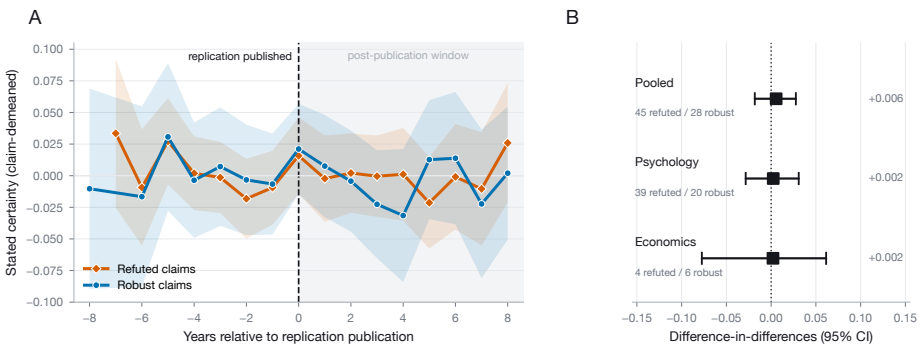


Fig. 4 The publication of a failed replication does not change how the claim is restated. (**A**) Mean claim-demeaned stated certainty by year relative to the replication project's publication; bands are 95% confidence intervals. (**B**) Difference-in-differences estimates with cluster-bootstrap 95% confidence intervals: the pooled estimate over all 73 claims, and, shown as descriptive within-project checks only, the psychology cohort (59 claims) and the much smaller economics cohort (10 claims), whose wide interval carries no independent inferential weight. The social-sciences project (4 claims) is omitted as underpowered.

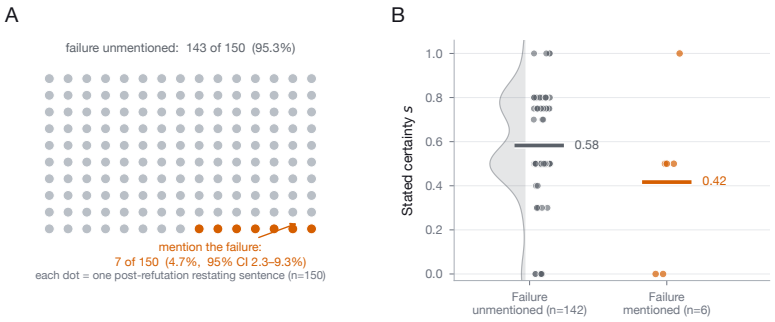


Fig. 5 The silence of the citing literature. (**A**) Each dot is one post-refutation sentence that restates a refuted claim (stratified sample, $n = 150$). Vermillion dots mention the replication failure or any controversy (7 of 150, 4.7%; Wilson 95% CI 2.3–9.3%); grey dots do not mention the failure (143 of 150, 95.3%). (**B**) Stated certainty among the 148 sentences with matched certainty scores, split by whether they mention the failure: the six acknowledging sentences hedge (mean 0.42), whereas the other 142 have a mean of 0.58. The acknowledging group is small; the contrast is a within-corpus indication that the scorer registers real caution, not a precise estimate.

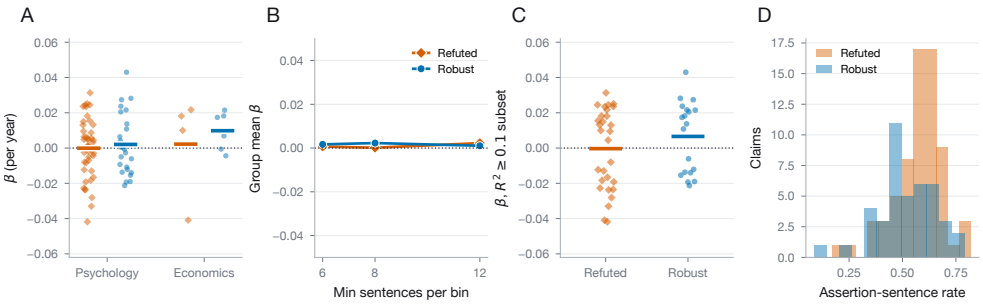


Fig. 6 The null is not an artifact of analytic choices. (**A**) Drift rates β by replication status within each project with adequate claim counts. (**B**) Group mean β under alternative minimum bin sizes. (**C**) The group comparison restricted to claims with well-determined trajectories ($R^2 \geq 0.1$). (**D**) Distribution of assertion-sentence rates by group: the filter treats both groups alike.

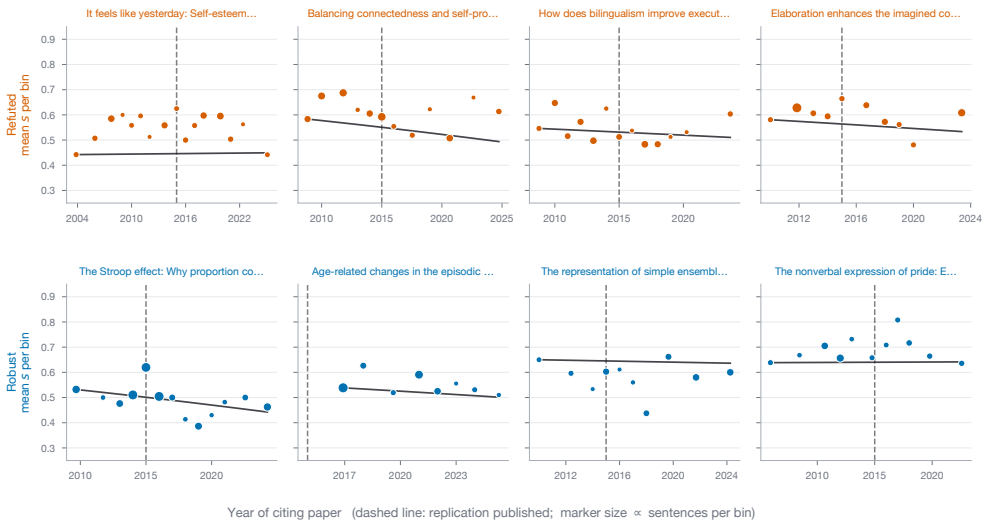


Fig. 7 Certainty trajectories of the eight most-cited claims. Mean stated certainty per adaptive year bin (marker size ∝ sentences) for the four most-cited refuted (top) and robust (bottom) claims, with the fitted drift curve and the replication-publication year (dashed). The refutation year is not visible in any trajectory.

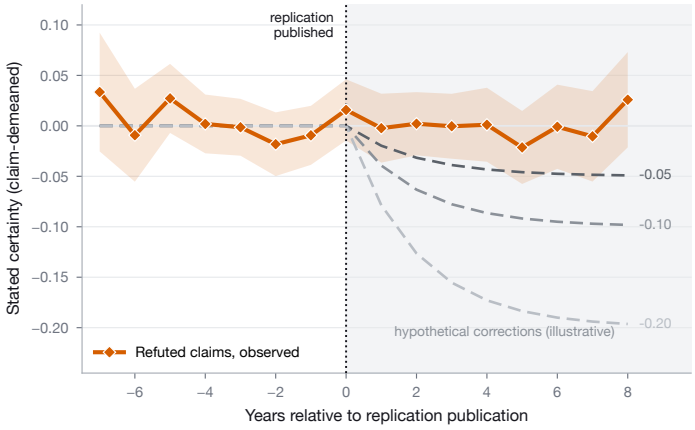


Fig. 8 The correction that did not happen. Observed claim-demeaned certainty of refuted claims around the replication event (orange, 95% confidence band), against illustrative trajectories of hypothetical corrections that lower post-event certainty by 0.05, 0.10, and 0.20 (dashed grey; drawn, not fitted). Even the smallest lies outside the observed band.

Supplementary Files

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

- Additionalfile1LaTeXsource.zip