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.

What does Retraction Watch's "computer-generated content" reason capture? A metadata analysis of 9,161 retractions

Jonas Mandalunes

jonas.mandalunes24414@oed.com.ph

AMA Computer University

Research Article

Keywords: Retraction, Research integrity, Paper mills, Generative artificial intelligence, Bibliographic databases, Scholarly publishing

Posted Date: August 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10546212/v1

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

Additional Declarations: No competing interests reported.

Abstract

Retraction Watch records a retraction reason for computer-aided or computer-generated content, and that reason is increasingly read as evidence about generative artificial intelligence in research. We asked what the reason captures when it is compared against the rest of the retraction record. From a snapshot of 71,448 records we retained the 66,013 classified as retractions and isolated the 9,161 (13.9%) carrying this reason. The category was absent before 2010, grew to 41.8% of retractions recorded for 2023 (95% CI 40.9–42.6), and one publisher accounts for 70.0%. No meaningful difference in median time to retraction was detected (difference 3 days, cluster-robust 95% CI − 17 to 43). Crude timing appeared far less dispersed (mean absolute deviation 268 versus 769 days, ratio 0.35), but this proved compositional: across 5,119 pairs matched exactly on retraction year, publication year and publisher the ratio moved to 0.97, leaving a residual of 5 days, and removing the dominant publisher from both cohorts reversed it. A paper-mill flag co-occurred in 71.1% of the cohort against 9.1% elsewhere; stratification attenuated a crude risk ratio of 7.78 to 1.43 (cluster-robust 95% CI 1.36–1.52) with extreme heterogeneity (Cochran Q = 248.2 on 7 degrees of freedom), so no common association holds. A crude post-retraction citation deficit of − 9.5 percentage points attenuated to − 1.3 (95% CI − 3.08 to 0.51) once pairs were matched on citations accumulated through the retraction year and on publisher. Both indicators come from one Reason field, so these are co-coding patterns and cannot establish what the articles contain.

1 Introduction

Retractions have grown faster than the literature producing them, and more than ten thousand articles were retracted in 2023, a record driven largely by mass withdrawals of paper-mill output (Van Noorden, 2023). The reasons recorded alongside retractions are the closest available register of how the scholarly record fails, and they are increasingly used as data (Khademizadeh et al., 2024; Kocyigit et al., 2023; Santos-d'Amorim et al., 2025; Steen et al., 2013).

One recorded reason concerns text not written by a person. Retraction Watch applies a reason labeled "Computer-Aided Content or Computer-Generated Content" when a notice or investigation attributes part of an article to automated generation. That label spans phenomena that differ in mechanism and in what they imply. The oldest is machine-generated nonsense: SCIgen output and its descendants, detected by fingerprinting (Cabanac & Labbé, 2021; Labbé & Labbé, 2013). A second is automated paraphrasing, whose signature is the tortured phrase (Cabanac et al., 2021). The newest is large-language-model text, publicly available from late 2022 and detectable at population level (Gray, 2024; Liang et al., 2024).

The distinction matters for policy. Machine-generated nonsense and automated paraphrasing are instruments of organized fabrication, produced at volume and sold. Large-language-model text is available to any author, and disclosed use is permitted by many publishers. If one category covers all three, counts drawn from it will be read as measuring generative-AI misuse when they may be measuring something else.

A related conflation runs through recent work. Three distinct objects are studied under similar names: retracted articles whose *subject matter* is artificial intelligence (Samala et al., 2025; Sridharan & Sivaramakrishnan, 2026); retractions whose recorded *reason* is computer-generated content (Saqr, 2026); and generative-AI authorship itself, which is inferred rather than recorded (Gray, 2024; Liang et al., 2024). Studies of the first report concentration in 2023 and median retraction times of 510 to 550 days (Sridharan & Sivaramakrishnan, 2026), along with slower detection of AI-related misconduct than of traditional misconduct (Guo et al., 2026). Those findings do not transfer to the second object, which is the subject here.

Saqr (2026) has examined that second object in a preprint, isolating the reason for 3,974 retracted DOIs, linking them to OpenAlex, reporting co-occurrence with paper-mill signatures in China and India, and describing a compressed median time to retraction of about 600 days against 1,300 or more for the United States and Japan. That analysis compares countries with one another. It does not compare the category against the rest of the retraction record, and it is that comparison which determines whether the category is distinctive at all.

Five questions follow. How large is the category, and how has its share of recorded retractions changed (RQ1)? How concentrated is it, and how much arises from shared notices rather than independent cases (RQ2)? Does time to retraction differ from other reasons in location and in dispersion, once clustering is respected and once year and publisher composition is held fixed (RQ3)? Which integrity flags co-occur, and does the association survive stratification by retraction year and publisher and increasingly strict definitions of the post-model era (RQ4)? Are these articles cited after retraction relative to matched retracted controls, and is their status visible in a major open index (RQ5)?

Our contribution is deliberately bounded. We quantify the empirical distribution and co-coding structure of a database category, and we show where a naive comparison misleads. Three results follow. First, the category's apparently distinctive timing is mostly an artifact of when and where these retractions occurred: it survives matching only as a small residual. Second, the paper-mill association widely attributed to the category attenuates sharply under stratification and is strongly heterogeneous, with reversals inside two large publishers. Third, an apparent post-retraction citation deficit is substantially attenuated once citation visibility measured through the retraction year is balanced, and is then no longer statistically distinguishable from zero. We do not claim to establish what the underlying articles contain; that requires notice-level and full-text validation, which we specify as the necessary next study.

As sole author, Jonas Mandalunes made substantial contributions to the conception and design of the work and to the acquisition, analysis and interpretation of the data; drafted and critically revised the work; approved the version to be published; and agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

2 Background

2.1 Machine-generated and machine-altered text in the scholarly record

Automatically generated papers entered the literature well before generative language models became public. Labbé and Labbé (2013) identified 85 SCIgen articles using inter-textual distance, and Cabanac and Labbé (2021) detected 243 nonsensical generated papers, mostly in one publisher's proceedings. Both rely on the text carrying a detectable fingerprint of machine origin.

Automated paraphrasing is a different mechanism with a different motive. Cabanac et al. (2021) showed that running existing text through synonym-substituting rewriting software leaves characteristic distortions, the tortured phrases, in which established terms are replaced by near-synonyms that no specialist would use. Here the machine is not composing new content but disguising copied content, typically to evade text-similarity screening. The output is therefore evidence of concealed plagiarism as much as of automation, which is a materially different finding from a manuscript generated from nothing. Grouping both under one reason code, as the database does, merges fabrication with obfuscation.

Large language models weakened the fingerprint assumption again, and detection now rests on distributional evidence. Gray (2024) estimated the prevalence of model-associated vocabulary across the 2023 literature, and Liang et al. (2024) estimated the fraction of conference peer reviews substantially modified by such models. Both infer population-level prevalence and identify no individual article with confidence, and nothing in either approach reaches a retraction notice.

2.2 Paper mills and mass retraction

Paper mills sell authorship of fabricated manuscripts, and their output is produced with templates and automated text tools. Candal-Pedreira et al. (2022) analyzed 186 paper-mill retractions and showed concentration in specific journals and countries; a follow-up examined reference and citation behavior (Candal-Pedreira et al., 2024). Because mills operate at volume, discovery produces volume responses: a publisher identifying a compromised special-issue pipeline may retract hundreds of articles under a single notice. Records sharing a notice are therefore not independent, and any timing analysis must accommodate that.

The period covered by this snapshot contains one such episode of unusual scale. Wiley, which had acquired Hindawi in 2021, reported that indicators of large-scale manipulation of the peer-review process were identified in Hindawi's special-issue program by its own research-integrity staff and by independent researchers working in parallel; the publisher paused special-issue publication from around October 2022, closed several titles, and in December 2023 announced that it would discontinue the Hindawi imprint and fold the remaining journals into its own portfolio, by which point several thousand articles had been retracted (Van Noorden, 2023). We describe this episode because it is a matter of public record reported by the publisher and in the trade and news literature, and because it bears directly on interpretation: any full-cohort analysis of retractions recorded in 2023 and 2024 is describing it in substantial part. We take no position on the adequacy of any publisher's response, which is outside the scope of a metadata study.

2.3 Retraction timing and post-retraction citation

Steen et al. (2013) showed that retraction growth reflected both increased detection and shortening time to retraction. Separately, retracted articles continue to be cited: Bar-Ilan and Halevi (2017) found that post-retraction citations frequently show no awareness of the retraction, and Hsiao and Schneider (2021) confirmed this across biomedicine. García-Romero et al. (2026) modeled the determinants of post-retraction citation and found that journal and institutional prestige predict continued citation while recorded misconduct severity does not, which implies that reason labeling matters for research about retraction rather than for citing behavior. Retraction Watch data have supported comparable analyses by field (Kwee & Kwee, 2023) and country (Kocyigit et al., 2023; Samala et al., 2025; Santos-d'Amorim et al., 2025). Table 1 positions this study.

Table 1
Position of this study relative to the closest prior work

Study	Object of study	Records	Comparator	Timing: location and dispersion	Clustering by notice	Year and publisher held fixed
Labbé & Labbé (2013); Cabanac & Labbé (2021); Cabanac et al. (2021)	Machine text in full texts	85; 243; n/a	None	No	No	No
Steen et al. (2013)	Retraction growth	All retractions	Within retractions	Location only	No	No
Candal-Pedreira et al. (2022, 2024)	Paper-mill retractions	186	None	Partly	No	No
Bar-Ilan & Halevi (2017); Hsiao & Schneider (2021); García-Romero et al. (2026)	Post-retraction citation	Varies; 531	Self-controlled	No	No	Partly
Sridharan & Sivaramakrishnan (2026); Samala et al. (2025); Guo et al. (2026)	Retracted articles *about* AI	335; national; disciplinary	None or disciplinary	Location only	No	No
Saqr (2026)	Computer-aided content reason	3,974	Between countries	Location only	No	No
This study	Computer-aided content reason	9,161	All other retraction reasons	Both	Yes	Yes, by exact matching

3 Data and methods

3.1 Data source and provenance

Crossref acquired the Retraction Watch database from the Center for Scientific Integrity in September 2023 and distributes it publicly. Crossref has since retired the Labs endpoint through which the data were formerly served and states that data obtained there will be out of date; the current distribution is a CSV file in a public repository, rebuilt each working day (Crossref, 2024).

We analyzed that repository file rather than the retired endpoint. The file retraction_watch.csv was downloaded on 31 July 2026 at commit dda8c8749c6ac35923f69162936c53fc9b9f5e3c, corresponding to the build titled "Retraction Watch data for 2026-07-29." The downloaded file was 65,517,426 bytes with SHA-256 checksum 1B61BABC73AD7F9861175ECB5DECC2B5C3BE33E0464C40489355A5BB96B9EF2F and contained 71,448 records across 20 fields. Fields used were Record ID, Title, Subject, Journal, Publisher, Country, ArticleType, RetractionDate, RetractionDOI, OriginalPaperDate, OriginalPaperDOI, RetractionNature, Reason and Notes. We describe this as the reference distribution in the sense that it is the copy the maintainers publish and version; we make no claim that its reason coding is complete or independently validated.

The secondary source was OpenAlex (Priem et al., 2022), queried through its REST API with an authenticated key, for citation counts, per-year citation series, primary topic and field, and retraction flags.

3.2 The reason code and its taxonomy history

Two features of the category's provenance bear directly on interpretation and are documented here because they are prior to any analysis.

First, the reason is defined broadly by the database itself. The maintainer's published vocabulary defines the category as content "created via a text generator, randomizing algorithm, generative AI, etc." and gives Mathgen, SCIgen and ChatGPT as examples of such tools (Retraction Watch, 2025a). Changes to the vocabulary are separately recorded by the maintainer (Retraction Watch, 2025b). The mixing of older random-generation tools with modern language models is therefore stated in the published definition, not inferred by us from the data, and the three tools named in that definition span precisely the distinction drawn in Section 2.1.

Second, the category has been renamed. The current snapshot contains no record carrying the earlier label "Randomly Generated Content"; every record in the relevant category carries the present wording, 9,162 occurrences across all record types and 9,161 among retractions. Because a category that once existed under a narrower name is entirely absent while its successor is populated back to 2010, historical records must have been mapped into the renamed category rather than left under the old wording. This is an observation about the state of the distributed file, and it is consistent with the maintainer's own practice of revising the vocabulary and applying it to the database as a whole. Part of the temporal pattern reported in Section 4.1 may therefore reflect the evolution of coding policy rather than a change in publisher behavior.

We could not quantify that component. The distributed file records the date an article was published and the date it was retracted, but not the date a record entered the database or the date a reason was assigned or revised. No field in the snapshot allows the date of coding to be separated from the date of retraction. This is a limitation of the available data rather than a choice, and it bounds what any analysis of this file, including ours, can say about when the label came to be applied.

3.3 Record selection

The database contains update types other than retractions, namely expressions of concern, corrections and reinstatements, identified by the RetractionNature field. Because the research questions concern retractions, analysis was restricted to records with RetractionNature equal to "Retraction" (Table 2).

Table 2
Record selection

Step	Records
Repository snapshot, all record types	71,448
Excluded: expressions of concern	−3,585
Excluded: corrections	−1,499
Excluded: reinstatements	−160
Excluded: blank record type	−191
Analyzed: RetractionNature = Retraction	66,013
Machine-attributed cohort	9,161
Comparison cohort, all other retraction reasons	56,852
Machine cohort with usable publication and retraction dates	9,161
Comparison cohort with usable publication and retraction dates	56,852

3.4 Cohort and indicator definitions

A record entered the machine-attributed cohort if its Reason field matched the case-insensitive pattern Computer-Aided Content|Computer-Generated Content. The Reason field is a semicolon-delimited list, so records normally carry several reasons and matching is inclusive. All remaining retractions formed the comparison cohort; no record was excluded by publisher, country, subject or date.

Time to retraction was the difference in days between OriginalPaperDate and RetractionDate, excluding records with missing or unparseable dates or with retraction preceding publication. Binary indicators were derived from the Reason field by case-insensitive substring match on the following patterns, given exactly as applied so the figures below can be reproduced: paper mill, Paper Mill; any peer-review concern, Peer Review; plagiarism, Plagiarism; duplication, Duplication. The peer-review pattern is deliberately inclusive and matches three strings in this snapshot, Concerns/Issues about Peer Review (11,253 records), Compromised Peer Review (10,866) and Taken via Peer Review (23); the last denotes text taken during review rather than a defect in review, and is retained because it is 23 records in 66,013 and excluding it changes no reported figure at two decimal places. Each record was assigned to a retraction notice using RetractionDOI, with records lacking a usable notice DOI treated as singleton clusters.

The maintainer publishes a definition for each reason. The category studied here is defined as content "created via a text generator, randomizing algorithm, generative AI, etc.", with Mathgen, SCIgen and ChatGPT given as examples (Retraction Watch, 2025a). The breadth of the category, and specifically its coverage of both random generators and large language models under one code, is therefore explicit in the published vocabulary. The paper-mill code is defined as applying when "a journal/publisher/institution states that a paper came from a paper mill or the paper is in a list of paper mill papers," and may also be applied where the compiler "has identified multiple paper mill criteria applying to this paper" (Retraction Watch, 2025a). Both codes are thus editorial determinations recorded by the compiler, which is the basis for treating their joint occurrence as co-coding rather than as independent corroboration.

Throughout, we describe the joint occurrence of two reason codes on the same record as *co-coding*. Both the machine-attributed indicator and the paper-mill indicator are drawn from the same Reason field, frequently from the same notice and the same investigation. No analysis in this paper observes paper-mill origin independently of the database's own coding.

Scope of claims. Three statements bound everything that follows. First, the unit of analysis is a database record, and every quantity reported is a property of how the Retraction Watch coding system has been applied, not a finding about the content of any article. Second, the retraction decisions themselves were made and published by the responsible journals and publishers; where this paper reports that records from a named publisher carry particular reason codes, it is describing those publishers' own published notices as recorded by a third party, and it makes no assertion about any publisher's conduct, editorial standards or good faith, nor about the conduct of any author. Third, where counts are reported by country, the country field records author affiliation as given in the article, so those counts describe the affiliations attached to retracted records and must not be read as characteristics of a national research community. Named entities appear because the concentration is itself the finding and suppressing the names would make the analysis unverifiable, not to attribute fault.

3.5 Statistical analysis

Continuous variables are summarized as median with a distribution-free 95% confidence interval from order statistics, quartiles, mean, standard deviation and mean absolute deviation from the median. Location was compared with the Mann–Whitney U test using a normal approximation with exact tie correction, and effect size reported as Cliff's delta with a 95% confidence interval; values below 0.11 are treated as negligible. Dispersion was compared with the Brown–Forsythe test, a Welch t-test on absolute deviations from the group median. Location and dispersion are reported separately for every subset, because a change in one does not imply a change in the other. Proportions carry Wilson 95% confidence intervals and are compared with two-proportion z-tests, risk ratios and differences in percentage points.

Because records sharing a notice are not independent, headline estimates carry cluster-robust 95% confidence intervals from a nonparametric bootstrap resampling whole notices with replacement, using 10,000 replicates and seed 20260731. Monte Carlo error on the interval endpoints is reported so that the replicate count can be judged sufficient. Unclustered p-values are retained for transparency but are not the basis of any conclusion.

Stratification by retraction year and publisher was implemented in two ways. For the binary co-coding outcome we report a Mantel–Haenszel adjusted risk ratio and odds ratio, with strata defined by retraction year crossed with publisher group (the six largest publishers in the cohort plus a residual category), a Robins–Breslow–Greenland variance for the odds-ratio interval, and a Cochran Q test of heterogeneity. For the continuous timing outcome, stratification alone is insufficient because calendar opportunity for a long lag differs by publication year as well, so we use exact matching as described in Section 3.6.

No survival analysis is reported: the at-risk denominator for retraction is the set of all published articles, which is unobserved, so time to retraction is analyzed as a continuous outcome among retracted articles only. All tests are two-sided. Effect sizes and cluster-robust intervals govern interpretation; p-values below 1e-15 are reported as such, the resolution limit of the normal approximation used.

3.6 Matched analysis of timing dispersion

A crude comparison of dispersion between these cohorts is confounded by composition. The machine-attributed cohort is concentrated in recent retraction years and in a small number of publishers, whereas the comparison cohort spans several decades. An article retracted in 2024 and published in 2022 cannot exhibit a ten-year lag, so a cohort concentrated in recent years is mechanically constrained to a narrower range of possible lags. Excluding the dominant publisher does not resolve this, because the remaining records still differ in retraction year and publication year composition.

We therefore repeated the dispersion comparison with year and publisher held fixed by construction. Strata are the exact cross of retraction year, publication year and the exact publisher string, normalised only for case and internal whitespace, giving 1,688 distinct publishers. An earlier version of this analysis collapsed publishers into the six largest plus a pooled residual category; that specification is retained only as a reported sensitivity, because pooling permits a record from one small publisher to be matched against a record from a different small publisher, which defeats the purpose of holding publisher fixed.

Three analyses are reported. First, stratum-specific dispersion, restricted to strata containing at least five records in each arm, reported as the ratio of the machine-attributed mean absolute deviation to the comparison mean absolute deviation, and the corresponding interquartile-range ratio. Second, a pooled within-stratum difference in mean absolute deviation, weighted by the smaller arm in each stratum, tested by permuting the cohort label only within strata, 10,000 permutations, so that the null preserves the composition entirely. Third, a 1:1 sample matched exactly within strata and drawn without replacement, compared directly on mean absolute deviation and interquartile range, with a within-pair permutation test. The matched sample necessarily covers only that part of the cohort for which a contemporaneous same-publisher control exists, and the unmatched remainder is reported.

Because 70% of the cohort comes from one publisher, the analysis is also repeated with Hindawi removed from **both** cohorts. This is stated explicitly because a publisher-exclusion test that removes records from the exposed group only, while retaining that publisher's records in the comparison group, does not hold publisher composition fixed and is not a valid sensitivity analysis; an earlier implementation of this study did exactly that, and the corrected two-sided exclusion is reported in Section 4.4.

3.7 Notice-level aggregation

Rather than selecting one arbitrary record per notice, each notice was summarized by the median, minimum and maximum time to retraction of its constituent articles and by the proportion of those articles carrying a paper-mill flag. Notice-level median lags were then compared between cohorts using the same location and dispersion tests.

3.8 Regime definitions, follow-up restriction and terminology search

Post-model regimes were defined by original publication date at four cut-points: 30 November 2022 (the public release of ChatGPT), 1 July 2023, 1 January 2024 and 1 July 2024, and at 1 June 2022 as a falsification cut-point preceding that release. Because recently published articles have had less opportunity to be investigated, each cut-point was re-examined restricting the sample to articles published at least 12, 18 and 24 months before the snapshot date, so that every retained article had that minimum possible follow-up.

The terminology search applied twenty case-insensitive, word-boundary regular expressions to the Title and Notes fields. They were written to tolerate the surface variation that defeats a naive keyword list: internal spacing and hyphenation (so that "ChatGPT", "Chat GPT" and "Chat-GPT" all match), singular and plural forms, version suffixes on model names, and both spelled-out and abbreviated forms of the same term. Coverage spans named systems, generic model vocabulary, descriptions of machine production, the older nonsense generators, and the paraphrasing tools associated with tortured phrases. Each pattern is reported verbatim alongside its own match count with the terminology results in Section 4.9, so any reader can reproduce the search exactly; the machine-readable list is also supplied as a column in Online Resource 3. The Notes field is not the full text of a retraction notice, so this searches available metadata rather than notices. All matches are listed for adjudication in Online Resource 1.

3.9 Citation analysis with matched comparison cohorts

Two matched designs are reported, because the first has a known weakness that the second repairs.

The initial design matched 1:1 without replacement, exact on publication year, retraction year and publisher, with a fallback to publication year and retraction year alone where no unused publisher match remained; the match tier is recorded for every pair. Citations accumulated through the retraction year could not enter this design, because they are not observable until the record has been retrieved and are therefore unavailable at the moment a control is selected. That is also why neither propensity-score nor coarsened exact matching on prior citations was possible at the design stage: the covariate that most strongly predicts the outcome is revealed only by the retrieval itself. Residual imbalance on that covariate was consequently handled after retrieval, by caliper restriction and by stratification within bands.

The second design removes the weakness by rematching within the retrieved data. The control candidate pool was constructed as follows, before any outcome was examined. The source population was all retractions in the snapshot not carrying the machine-attributed reason and holding a parseable original DOI and date pair. From that population, every record sharing a publication year and retraction year with at least one machine-attributed record was eligible, since a control outside those years can never satisfy the match key; records outside those year combinations were therefore not queried, which is an efficiency restriction and not a selection on outcome. Eligible DOIs were queried in OpenAlex in publication-year order until the daily request allowance was reached, yielding 24,976 candidates with citation data. Publisher and journal were taken from the Retraction Watch record rather than from OpenAlex, so publisher agreement is defined on the same field in both cohorts. Primary topic and field were retrieved separately for every DOI in both cohorts. The pool is frozen and supplied in Online Resource 3 so the matching is reproducible without further querying. Because the allowance truncated the queue rather than sampling it, the pool over-represents publication years shared with the largest machine-attributed cells; this is the reason the rematched analyses retain fewer pairs than the initial design, and it is reported as a limitation rather than corrected.

From that pool, each machine-attributed record was matched exactly on citations accumulated through the retraction year, publication year and retraction year, once additionally requiring exact publisher agreement and once leaving publisher free. Publisher for the machine-attributed records is read from the Retraction Watch snapshot directly, for all 9,142 records with a usable DOI, so the eligible population for the second design does not depend on whether a record was successfully paired under the first. An earlier implementation took publisher from the initial matched-pair table, which contains only the 8,594 records that first design managed to pair; that made the repaired analysis partly conditional on success in the analysis it was meant to replace, and excluded 548 records from publisher-exact rematching regardless of whether an eligible control existed. Correcting this adds 31 usable pairs and leaves every machine record with a publisher value, so none is now excluded for want of one. Only the first retrieved record per DOI is retained, consistent with the deduplication described above; 62 repeated machine-cohort rows were dropped on this basis. Because the pool is fixed, this involves no further querying, and because baseline citations enter the match key, balance on them is exact by construction rather than approximate. Sensitivity to the arbitrary element of matching, namely which of several eligible controls is drawn, is assessed by repeating the assignment twenty times under different random seeds.

For both designs the analysis is paired: binary outcomes are compared with McNemar's test and a paired confidence interval for the difference in proportions, with a conditional odds ratio estimated from discordant pairs, and citation rates with the Wilcoxon signed-rank test on within-pair differences together with a paired bootstrap of the median difference. Covariate balance is reported as standardized mean differences.

Where the interface returned more than one record for a single DOI, which occurs when a work has more than one OpenAlex entry, records were ordered by the interface's own default relevance ranking and the first retained. This affected 143 distinct DOIs, 1.6% of those queried, and no pair was lost by it. Because the choice is arbitrary, we note it as a minor source of unquantified variation rather than a principled rule.

Because OpenAlex citation data are resolved to calendar years rather than exact dates, all follow-up measures are defined on calendar years. Citations recorded in years after the retraction year are classified as post-retraction and the remainder as accumulated through the retraction year, so citations occurring after the retraction date but within the same calendar year count toward the earlier total. For that reason we write "citations accumulated through the retraction year" throughout rather than "prior citations." Exposure is the snapshot year minus the retraction year. Every follow-up measure requires the corresponding exposure in both members of a pair, so denominators differ between measures and are reported for each.

3.10 Reproducibility, computational validation and use of generative AI

All quantities reported in this study were generated by reproducible scripts that take the pinned Retraction Watch snapshot and, where applicable, the frozen OpenAlex extract as inputs and produce the reported tables, figures and derived result files. The snapshot itself is not redistributed, because it is rebuilt daily by its maintainers; it is instead identified immutably by commit hash, byte length and checksum so that the analysed build can be recovered exactly. The source code, frozen OpenAlex extract, software requirements, fixed random-number seeds and instructions for reproducing the analyses are provided in the Online Resources.

Several estimators and resampling procedures were implemented in custom code rather than invoked through dedicated statistical functions. To check implementation accuracy, selected principal statistics were therefore reimplemented in a separate Python validation workflow using NumPy 2.5.1 and SciPy 1.18.0 and compared with the primary implementation. The cross-check covered the crude timing and dispersion comparisons, co-occurrence proportions, Mantel-Haenszel point estimates and a 2,000-replicate notice-cluster bootstrap.

The two implementations produced identical Mann-Whitney *U* statistics and Cliff's delta estimates. Absolute differences were below 0.0003 for the Welch statistics and Mantel-Haenszel risk and odds ratios. The Cochran *Q* statistics differed by 0.013 because the implementations accumulated stratum contributions in a different order; this difference did not affect the reported conclusion of extreme heterogeneity. Bootstrap confidence-interval endpoints differed by one to six days because the implementations used independently generated resampling sequences. Because identical bootstrap samples were not supplied to both implementations, this comparison was treated as an assessment of consistency rather than a test of numerical identity, and no pass-or-fail tolerance was imposed on the interval endpoints. The comparison table and validation script are provided in Online Resource 2.

Three analyses introduced during revision were available in one Python implementation only: the exact-publisher matched dispersion analysis, the citation rematching and the notice-cluster bootstrap for the Mantel-Haenszel estimates. For these analyses, reproducibility is supported by archived input data, versioned source code, fixed random-number seeds, within-stratum or within-pair permutation procedures where applicable, repeated matching under multiple seeds and complete derived result tables for third-party re-execution. Fixed seeds permit exact repetition of stochastic procedures but are not, by themselves, presented as evidence that an implementation is correct.

During the preparation of this manuscript, the author used a large language model to assist with identifying potentially relevant literature and duplicate bibliographic records, drafting and editing manuscript text, and producing preliminary analysis-code drafts. Candidate references identified with AI assistance were included only after the author checked their bibliographic details and relevance to the claims for which they were cited; every reference carrying a digital object identifier or arXiv identifier was additionally verified against the issuing registry, and the verification output is supplied in Online Resource 3. All AI-assisted code was inspected, revised where necessary and executed by the author. Its inputs and outputs were checked against the archived source data and, for the principal statistics described above, against the separate validation implementation.

The large language model was not used to create or modify the source data, make final decisions about record inclusion, autonomously execute the reported analyses, select results for reporting, determine the interpretation of the findings or introduce unverified references. No AI system is listed as an author. The author retains full responsibility for the accuracy, originality, integrity, interpretation and final content of the manuscript.

4 Results

4.1 Scale and temporal distribution

Of 66,013 retractions, 9,161 (13.9%) carried a computer-aided or computer-generated content reason. The category is entirely absent from retractions recorded before 2010: none of the 3,241 such records carries the reason. Across 2010 to 2019 it remains negligible at 28 records among 24,053 retractions, so 28 records in the 27,294 retractions recorded before 2020. It then rose sharply and unevenly (Table 3; Fig. 1). Records for retraction year 2023 accounted for 41.8% of retractions recorded that year (95% CI 40.9–42.6); the share then settled at 23.1% in 2024 and 22.1% in 2025. As Section 3.2 notes, this series is indexed by retraction year, and the snapshot does not permit the date of coding to be separated from it, so an unknown part of the rise may reflect the renaming and back-mapping of the category rather than contemporaneous publisher behavior.

Table 3
Machine-attributed retractions by retraction year, with share of all retractions recorded for that year and Wilson 95% confidence intervals. Rows cover all 66,013 analyzed retractions except one record whose retraction date is not parseable

Retraction year	Machine-attributed	All retractions	Share % (95% CI)
Before 2010	0	3,241	0.00 (0.00–0.12)
2010–2019	28	24,053	0.12 (0.08–0.17)
2020	39	3,079	1.27 (0.93–1.73)
2021	481	3,905	12.32 (11.32–13.39)
2022	383	5,570	6.88 (6.24–7.57)
2023	5,514	13,208	41.75 (40.91–42.59)
2024	1,417	6,145	23.06 (22.02–24.13)
2025	1,256	5,697	22.05 (20.99–23.14)
2026, partial	43	1,114	3.86 (2.88–5.16)
Total	9,161	66,012	13.88

Because the two cohorts occupy different calendar spans, the timing analyses in Sections 4.3 to 4.5 are also reported on common calendar support: the matched analysis in Section 4.4 conditions on retraction year and publication year exactly, so it uses only strata in which both cohorts are present, which is the stricter form of the same restriction.

4.2 Concentration and notice structure

The cohort is concentrated on every axis examined (Table 4). Hindawi accounted for 6,409 records (70.0%) and IOS Press for 1,039 (11.3%). The Country field records every country appearing in the author affiliations as a semicolon-delimited list, so it must be read carefully: the field is exactly "China" in 6,590 records (71.9%) and exactly "India" in 816 (8.9%), while China appears anywhere in the field in 7,223 records (78.9%) and India in 1,263 (13.8%). The larger figures are the correct ones for a statement about affiliations; the smaller are the correct ones for a statement about single-country records. Table 4 reports the exact-field counts because they are mutually exclusive categories.

This concentration has a specific and documented cause rather than being a diffuse property of the literature. As described in Section 2.2, the Hindawi special-issue investigation and the mass retractions that followed it fall squarely inside this snapshot's 2023 and 2024 windows. The full-cohort figures below therefore describe that episode in large part, which is why every headline timing result is accompanied by a without-Hindawi variant and, in Section 4.4, by an analysis that holds publisher fixed.

The 9,161 records arise from 8,127 distinct retraction notices, and the structure is unusual: only two notices cover more than one article. Those two, of 657 and 379 articles, together account for 1,036 records, which is 1,034 in excess of one record per notice. Within them the articles are heterogeneous, with a median spread between earliest and latest time to retraction of 4,210 days. This structure is why headline estimates carry cluster-robust intervals and why notice-level results are reported alongside record-level results.

Table 4
Concentration of the machine-attributed cohort (n = 9,161)

Dimension	Largest category	n	%	Second largest	n	%
Publisher	Hindawi	6,409	69.96	IOS Press	1,039	11.34
Journal	Journal of Intelligent & Fuzzy Systems	1,039	11.34	Journal of Healthcare Engineering	600	6.55
Country	China	6,590	71.94	India	816	8.91

4.3 Time to retraction: location and crude dispersion

Median time to retraction was 491 days (95% CI 486–497) in the cohort and 488 days (95% CI 481–490) in the comparison group (Table 5). The unclustered Mann–Whitney test was significant (U = 274,613,105; z = 8.39; p < 1e-15) but the effect size was negligible (Cliff's delta 0.055, 95% CI 0.042–0.067), and under cluster-robust inference the median difference was 3 days with a 95% confidence interval of − 17 to 43 days, which includes zero (Table 6). There is no evidence that machine-attributed articles are retracted faster than other retracted articles.

Crude dispersion differs sharply (Fig. 2). The interquartile range was 325 days against 1,073 days, and the mean absolute deviation from the median 268 against 769 days, a ratio of 0.35 (Brown–Forsythe t = − 75.25, df = 37,255, p < 1e-15), with a cluster-robust difference of − 501 days (95% CI − 563 to − 423). Section 4.4 shows that most of this gap is attributable to cohort composition rather than to the label, so the crude figures should not be read as a substantive finding on their own.

Table 5
Time from publication to retraction, in days

Cohort	n	Median (95% CI)	Q1	Q3	IQR	Mean	SD	Mean absolute deviation	Max
Machine-attributed	9,161	491 (486–497)	369	694	325	621.4	484.0	267.5	4,505
All other retraction reasons	56,852	488 (481–490)	102	1,175	1,073	931.6	1,341.3	768.5	29,622

Table 6
Cluster-robust 95% confidence intervals from a bootstrap resampling whole retraction notices, 10,000 replicates, with Monte Carlo standard error of each interval endpoint. Records whose retraction DOI is blank or the literal value "unavailable" carry no notice identity and are treated as singleton clusters, giving 61,135 clusters over the 66,013 analyzed retractions

Estimate	Point estimate	Cluster-robust 95% CI	Monte Carlo SE of endpoints
Median time-to-retraction difference, days	3	−17 to 43	0.00, 0.33
Mean-absolute-deviation difference, days	−501.0	−563.2 to − 423.0	0.23, 0.44
Paper-mill co-coding difference, percentage points	61.98	57.80 to 66.59	0.03, 0.04

4.4 The dispersion difference is largely compositional

Holding retraction year, publication year and the exact publisher fixed removes most of the crude dispersion gap (Table 7; Fig. 3).

The exact cross of those three variables yields 9,811 strata, of which 39 contain at least five records in each arm. Within them the mean absolute deviation is small in *both* arms, which is the mechanism at issue: in the largest stratum, articles published in 2022 and retracted in 2023 by Hindawi, it is 75.6 days for 4,253 machine-attributed records against 91.8 days for 3,215 comparison records. Most of the dispersion visible in Table 5 is variation *between* years and publishers, not within them.

Stratum-specific ratios are consistent in direction but modest in size. The median stratum ratio of mean absolute deviations is 0.71, with 31 of 39 strata below 1.0, and 30 of 39 interquartile-range ratios below 1.0. The pooled within-stratum difference is − 18.9 days, against − 501 days unadjusted. Permuting the cohort label only within strata, so that composition is preserved exactly, gives a null distribution spanning − 2.5 to 2.2 days, and the observed value falls far outside it ($p < 1e\text{-}4$, 10,000 permutations).

The matched comparison is the most direct. Drawing a contemporaneous same-publisher control without replacement for each machine-attributed record yields 5,119 pairs; 4,042 records could not be matched because no unused control existed in their stratum, so this analysis covers 56% of the cohort and is weighted toward the larger publisher-year cells. Within those pairs the mean absolute deviation is 175.0 days against 180.2, a ratio of 0.972, and the interquartile ranges are 260 against 259 days, a ratio of 1.004. The paired difference of − 5.1 days remains statistically detectable under a within-pair permutation test ($p = 0.0009$) because the sample is large, but it is roughly a hundredth of the unadjusted difference. Repeating the matched draw under twenty different random seeds gives a mean ratio of 0.971 with a standard deviation of 0.002 and a range of 0.967 to 0.976, and the eligible pair count is fixed at 5,119 by the match keys, so the ratio does not depend on which eligible control happens to be drawn.

Two further specifications bound the result. Collapsing publishers into the six largest plus a pooled residual category, as an earlier version of this analysis did, gives 5,216 pairs and a ratio of 0.968; the difference from the exact specification is small, but the exact specification is the defensible one and is reported as primary. More consequentially, removing Hindawi from both cohorts leaves 2,752 machine-attributed records, 31 informative strata and 811 matched pairs, and **reverses the direction**: the matched mean absolute deviation is 401.4 days for the machine-attributed members against 377.9 for their controls, a ratio of 1.062, with a paired difference of + 23.5 days ($p = 0.0002$). Outside the dominant publisher, these retractions are if anything slightly *more* dispersed in timing than comparable retractions from the same publisher and years.

The conclusion is that the machine-attributed cohort is not more tightly timed than comparable retractions. Its crude appearance of narrow timing is a consequence of being concentrated in recent years and in a few publishers, where lags are short and similar for everything those publishers retracted. A small residual survives matching within the full cohort, but it does not survive removal of the dominant publisher, and no reading of these results supports treating timing homogeneity as a distinguishing feature of the category.

Table 7
Dispersion of time to retraction, crude and with retraction year, publication year and publisher held fixed. The primary specification uses the exact publisher string; the publisher-group row is retained only to show the effect of pooling

Analysis	Machine records	MAD, machine (days)	MAD, comparison (days)	Ratio	Test
Crude, whole cohort	9,161	267.5	768.5	0.348	Brown–Forsythe p < 1e-15
Stratum-specific, median of 39 informative strata, exact publisher	8,011	—	—	0.714	31 of 39 strata below 1.0
Pooled within-stratum difference, exact publisher	8,011	−18.9 days	—	—	within-stratum permutation p < 1e-4
Matched 1:1, exact publisher (primary)	**5,119 pairs**	**175.0**	**180.2**	**0.972**	within-pair permutation p = 0.0009
Matched 1:1, exact publisher, 20 random control draws	5,119 pairs in every draw	—	—	mean 0.971, SD 0.002, range 0.967–0.976	—
Matched 1:1, exact publisher, interquartile range	5,119 pairs	260.0	259.0	1.004	—
Matched 1:1, publisher group (sensitivity)	5,216 pairs	210.8	217.8	0.968	within-pair permutation p < 1e-4
Matched 1:1, exact publisher, Hindawi removed from both cohorts	811 pairs	401.4	377.9	1.062	within-pair permutation p = 0.0002

4.5 Sensitivity of the timing result to notice structure

Location and dispersion behave differently under sensitivity analysis, and conflating them would invert the conclusion (Table 8). Because only two notices cover more than one article, the informative sensitivity analyses remove those two notices individually and together rather than an arbitrary set of the largest notices. Removing either reduces what little location difference exists and lowers the crude dispersion further; removing both leaves 8,125 singleton records with a mean absolute deviation of 214 days. Aggregating to notice level, by summarizing each notice with the median lag of its constituent articles, gives a notice-level median of 475 days against 471 in the comparison cohort, a negligible location difference (Cliff's delta 0.035, z = 5.10, p = 3.4e-7).

These analyses establish that the crude dispersion gap is not created by the two large notices. They do not address composition by year and publisher, which Section 4.4 shows is the substantive explanation.

Table 8
Sensitivity analyses, with location and dispersion reported separately. Notices A and B are the only two notices covering more than one article

Analysis	n	Median	Mean absolute deviation	Cliff's delta (location)	Location p	Dispersion ratio	Dispersion p	Paper-mill co-coding %
Full machine cohort	9,161	491	267.5	0.055	< 1e-15	0.348	< 1e-15	71.12
Excluding notice A (657 articles)	8,504	478	225.3	0.028	3.5e-5	0.293	< 1e-15	68.93
Excluding notice B (379 articles)	8,782	489	259.1	0.049	1.1e-13	0.337	< 1e-15	69.87
Excluding both multi-article notices	8,125	475	214.1	0.021	0.002	0.279	< 1e-15	67.48
Notice level, median lag per notice	8,127	475	214.1	0.035	3.4e-7	0.279	< 1e-15	not applicable
Excluding Hindawi from both cohorts	2,752	575.5	551.0	0.115	< 1e-15	0.669	< 1e-15	47.75

The Hindawi row removes that publisher's records from the machine-attributed cohort **and** from the comparison cohort, leaving 2,752 and 51,774 records respectively, so the comparison mean absolute deviation rises from 768.5 to 823.5 days and the dispersion ratio is 0.669. An earlier implementation of this study removed Hindawi from the exposed cohort only while retaining that publisher's 5,077 records in the comparison cohort, which left the comparison statistic unchanged and understated the ratio at 0.717. That specification does not hold publisher composition fixed and is not reported. The remaining rows of this table concern notice structure and are unaffected.

4.6 Co-occurring integrity codes, crude and stratified

The machine-attributed reason rarely appears alone (Table 9). A paper-mill code co-occurred in 71.1% of the cohort against 9.1% of the comparison group, a difference of 62.0 percentage points with a cluster-robust interval of 57.8 to 66.6 and a crude risk ratio of 7.78. Some peer-review concern was recorded for 93.5% against 22.1%. Codes associated with individual authorial misconduct ran the other way: plagiarism 0.6% against 10.7% (risk ratio 0.06) and duplication 0.2% against 18.3% (risk ratio 0.01).

Stratifying on retraction year and publisher attenuates the association sharply, from a crude risk ratio of 7.78 to a Mantel–Haenszel 1.43 across 23 contributing strata (Table 10). The Greenland–Robins interval for that estimate is 1.39 to 1.48, but it assumes records are independent, which they are not: records sharing a retraction notice are correlated. Resampling whole notices instead gives a cluster-robust interval of **1.36 to 1.52** for the risk ratio and 2.07 to 2.59 for the odds ratio, roughly 1.7 times wider, and these are the intervals we report and quote in the abstract. The independence-based interval is retained in Table 10 only to show the size of the understatement. Heterogeneity is extreme: Cochran Q was 248.2 on 7 degrees of freedom. The degrees of freedom are fewer than the number of contributing strata because the stratum-specific log odds ratio and its variance are undefined wherever any cell is zero, so the Q test uses only the 8 strata in which all four cells are non-zero.

We therefore do not interpret the pooled ratio as an estimate of a common association, and we do not present it as the finding. With heterogeneity of this magnitude the pooled value is a weighted average of dissimilar quantities, and even a dependence-aware interval around it does not make it a common effect. The substantive result is the heterogeneity itself, set out in Tables 11 and 12.

Table 9
Co-occurrence of other integrity codes, crude comparison

Code	Machine-attributed % (95% CI)	All other reasons % (95% CI)	Difference, pp (95% CI)	Risk ratio	z	p
Paper mill	71.12 (70.18–72.04)	9.14 (8.90–9.38)	61.98 (61.02–62.94)	7.78	144.12	< 1e-15
Any peer-review concern	93.47 (92.95–93.96)	22.06 (21.73–22.41)	71.41 (70.80–72.02)	4.24	136.00	< 1e-15
Plagiarism	0.64 (0.50–0.83)	10.67 (10.42–10.92)	−10.02 (− 10.32 to − 9.72)	0.06	−30.69	< 1e-15
Duplication	0.22 (0.14–0.34)	18.29 (17.97–18.61)	−18.07 (− 18.40 to − 17.74)	0.01	−44.02	< 1e-15

Table 10
Paper-mill co-coding, crude and stratified by retraction year and publisher

Estimate	Value
Crude risk ratio	7.78
Mantel–Haenszel risk ratio, cluster-robust 95% CI (notice bootstrap, 10,000 replicates)	1.43 (1.36 to 1.52)
Mantel–Haenszel odds ratio, cluster-robust 95% CI	2.31 (2.07 to 2.59)
Same estimates under Greenland–Robins, which assumes independence	RR 1.43 (1.39–1.48); OR 2.31 (2.15–2.48)
Contributing strata	23
Strata with all four cells non-zero, contributing to the Q test	8
Cochran Q for heterogeneity (df)	248.2 (7)
Interpretation	Heterogeneity far exceeds df; pooled value is not a common effect

4.7 The paper-mill co-coding pattern varies by year and publisher

Stratification shows where the aggregate association comes from. By retraction year (Table 11) it is absent before 2023 and then reverses direction. No machine-attributed record carried a paper-mill code in any pre-2023 period: none of the 28 records retracted in 2010–2019, none of the 39 in 2020, none of the 481 in 2021 and none of the 383 in 2022, while 23.8% and 18.0% of other retractions in 2021 and 2022 did. From 2023 the pattern inverts, reaching 82.5% against 1.76% in 2025. By publisher (Table 12) the pattern is strong in Hindawi, IOS Press and Wiley, absent in Springer Nature's Nature Publishing Group imprint, and reversed in Springer and IOP Publishing, where machine-attributed records were *less* likely to carry a paper-mill code than other retractions from the same publisher.

A co-coding rate that moves from exactly zero to above 80% within two years is more readily explained by changing coding and investigation practice than by a change in the articles themselves, particularly given the taxonomy history in Section 3.2.

Table 11
Paper-mill co-coding by retraction year

Retraction year	Machine-attributed %	All other reasons %	Difference, pp (95% CI)	Risk ratio	p
2010–2019	0.00 (n = 28)	not tested	not tested	0.00	not tested
2020	0.00 (n = 39)	not tested	not tested	0.00	not tested
2021	0.00	23.80	−23.80 (− 25.23 to − 22.38)	0.00	< 1e-15
2022	0.00	17.99	−17.99 (− 19.03 to − 16.94)	0.00	< 1e-15
2023	76.80	33.10	43.70 (42.17–45.23)	2.32	< 1e-15
2024	87.79	11.87	75.93 (73.99–77.86)	7.40	< 1e-15
2025	82.48	1.76	80.73 (78.59–82.87)	46.96	< 1e-15

Table 12
Paper-mill co-coding within publisher, six largest publishers in the cohort

Publisher	Machine-attributed % (n)	All other reasons % (n)	Difference, pp	Risk ratio	p
Hindawi	81.15 (6,409)	43.18 (5,077)	37.98	1.88	< 1e-15
IOS Press	99.62 (1,039)	3.14 (733)	96.48	31.75	< 1e-15
Springer	0.49 (608)	2.88 (3,888)	−2.39	0.17	5.3e-4
IOP Publishing	1.65 (485)	62.89 (795)	−61.24	0.03	< 1e-15
Wiley	85.90 (305)	8.46 (3,771)	77.44	10.15	< 1e-15
Springer Nature (Nature Publishing Group)	0.00 (96)	0.99 (2,829)	−0.99	0.00	0.327

Publishers appear here because they are the axis along which the association varies, and because the reversals in Springer and IOP Publishing are only visible when they are named. Each percentage is the share of that publisher's *own* retracted records carrying a paper-mill code as recorded by Retraction Watch, so differences between publishers reflect differences in what their notices state and in how those notices have been coded. They are not a measure of the prevalence of paper-mill activity in any publisher's active portfolio, and a higher share may equally reflect a more thorough investigation and fuller disclosure. No comparative judgment about publishers is intended or supported.

4.8 Regime definitions are confounded by differential follow-up

Under the release-date definition, 8,194 cohort articles were published before 30 November 2022 and 967 on or after it, and the paper-mill code was statistically unchanged across that boundary (69.6% against 71.3%). Under progressively stricter cut-points the paper-mill share among recently published articles appears to fall monotonically, reaching 0.6% for articles published from 1 July 2024 against 72.5% before that date.

That apparent decline is substantially an artifact of differential follow-up (Table 13). Restricting each comparison to articles with at least 24 months of possible follow-up reverses the direction: the paper-mill share among later-published articles rises to 84.3% at the November 2022 cut-point, 82.8% at July 2023 and 77.8% at January 2024. At the strictest cut-point only seven articles survive a 24-month restriction, so that comparison is uninformative. The pattern does persist at 12 and 18 months of follow-up, at 0.61% and 0.63% among 165 and 159 articles, but recently published and already retracted articles are a highly selected set of rapidly detected cases, as their median time to retraction of 136 days indicates. We therefore treat the strictest cut-point as an exploratory recent-cohort pattern rather than evidence of a completed regime change.

The falsification cut-point behaves as expected for a boundary marking nothing: placed at 1 June 2022, five months before ChatGPT's public release, it produces an *increase* in the paper-mill share of 11.6 percentage points, consistent with cut-points separating publisher cohorts rather than technological regimes.

Table 13
Paper-mill co-coding by regime cut-point and minimum possible follow-up. A minimum follow-up of k months retains only articles published on or before the snapshot date minus k months. The complete grid, including all four follow-up levels at every cut-point, is in Online Resource 3

Cut-point (original publication date)	Min follow-up	n on or after	Paper mill % before	Paper mill % on or after (95% CI)	Median lag on or after
30 Nov 2022	none	967	71.30	69.60 (66.62–72.41)	402
30 Nov 2022	24 months	797	71.30	84.32 (81.63–86.68)	456
1 Jul 2023	none	635	71.89	60.79 (56.94–64.51)	390
1 Jul 2023	24 months	465	71.89	82.80 (79.10–85.95)	442
1 Jan 2024	none	395	72.31	44.56 (39.73–49.49)	283
1 Jan 2024	24 months	225	72.31	77.78 (71.90–82.72)	369
1 Jul 2024	none	177	72.51	0.56 (0.10–3.13)	137
1 Jul 2024	24 months	7	72.51	0.00 (0.00–35.43)	399

4.9 Explicit generative-AI terminology is rare in the searched metadata

Across all 66,013 retractions, the twenty patterns of Section 3.8 matched 718 records. The total is dominated by the generic phrase "artificial intelligence," which matched 700 records, of which 315 are machine-attributed, and which in this database almost always reflects the subject matter of the retracted article rather than the means of its production. Patterns naming a specific generative-AI system or technique matched 21 records (Table 14). We call these "specific" only to distinguish them from the generic phrase; a match does not by itself indicate that the article was machine-produced, and indeed most are not.

Broadening the search relative to a narrower initial pattern set changed the count by two records, from 19 to 21, and did not change the picture. Eleven of the twenty patterns matched nothing at all: bare LLM or LLMs, machine-generated or machine-written, automatically generated, automated generation, synthetic text, neural text generator, hallucinated citations or references, generated references, SCIgen or Mathgen, tortured phrases, and paraphrasing or text-spinning tools.

All 21 records were classified individually against rules fixed before classification, into subject matter (the matched term names the object of study), production (the metadata indicates machine generation of the article's own content), false positive (the string is not a reference to a generative system) or indeterminate. The classification, the supporting note and the identity of the adjudicator are recorded for every record in Online Resource 1b. The outcome was 18 subject matter, 2 false positives and 1 production, and **no record was classified as production while also carrying the computer-generated-content reason**, which is the specific claim made here. Adjudication was performed by the sole author against Title and Notes metadata; there was no second rater, so no inter-rater agreement statistic is reported and the classification should be read as one trained reader's judgment rather than as a verified consensus coding. This is a limitation of the present design, and it is the first thing the notice audit proposed in Section 5.6 would remedy.

Four of the 21 are machine-attributed, and all four were classified as subject matter: studies of generative-AI tasks for language learning, of sentiment analysis from rule-based lexicons to large language models, of ChatGPT in educational transformation, and of large-language-model fine-tuning for news summarization. These four illustrate why subject-matter terminology in Title and Notes cannot validate production-related coding: an article *about* language models here carries the reason code for computer-generated content, but because we did not obtain the notices we cannot say whether the code was applied on evidence recorded elsewhere. Their presence is not by itself evidence of mistaken coding. The two false positives are records in which "Gemini" denotes gemini surfactants, a class of amphiphilic compounds unrelated to the language model of that name. The single record classified as production concerns AI-generated figures and links to a report of image manipulation; it does **not** carry the computer-generated-content reason.

Two conclusions follow, both bounded. First, the reason code is not, in general, accompanied by explicit generative-AI attribution in these fields. Second, the one record whose metadata does indicate AI-generated material does not carry the code. We do not present that as a coding error. A compiler can only code what a notice discloses, and a notice that describes manipulated figures without characterizing them as automatically generated gives no basis for applying this reason; the observation indicates a limit on what the field can capture rather than a mistake in applying it. Because the Notes field is not the full text of a retraction notice, absence of a tool name in metadata is not evidence of absence of generated content in the article, and we draw no such inference. The full 718-record frame is provided in Online Resource 1.

Table 14
Search patterns and records matching generative-AI terminology in Title or Notes fields (n = 66,013 retractions). Patterns are Python regular expressions app eleven patterns that matched nothing are listed together in the final row; the complete per-pattern grid is in Online Resource 3

Search target	Regular expression	Records matching	Ma att
artificial intelligence, generic	\bartificial\s+intelligence\b	700	31
ChatGPT, any spacing or hyphenation	\bchat\s*[-]?\s*gpt\b	8	1
language model, singular or plural	\blanguage[\s\-]models?\b	5	2
named assistants	\b(?:bard|gemini|copilot|claude|llama|mistral|deepseek|qwen|grok)\b	5	0
large language model, singular or plural	\blarge[\s\-]language[\s\-]models?\b	3	2
generative AI or GenAI	\b(?:generative\s + a\.?i\.?|gen[\s\-]?ai)\b	1	1
GPT with version number	\bgpt[\s\-]?[2–9](?:\.\d)?\b	1	0
GPT, bare token	\bgpt\b	1	0
AI-generated	\bai[\s\-]generated\b	1	0
Eleven patterns matching no record: bare LLM \bllms?\b; machine-generated or written \bmachine[\s\-](?:generated|written)\b; automatically generated \bautomatic(?:ally)?[\s\-]generated\b; automated generation \bautomated\s+(?:\w+\s+)?generation\b; synthetic text \bsynthetic\s+text\b; neural text generator \bneural\s+text\s+generat\w+\b; hallucinated citations \bhallucinat\w+\s+(?:citation|reference)s?\b; generated references \bgenerated\s+references?\b; SCIgen or Mathgen \b(?:sci\s*gen|math\s*gen)\b; tortured phrases \btortured\s+phrase\w*\b; paraphrasing and spinning tools \b(?:paraphras\w+|spinbot|article\s+spinner|text\s+spinn\w+)\b	as listed	0	0
Distinct records matching any pattern	union of the twenty	718	—
Distinct records matching a specific generative-AI-related pattern	union excluding the generic phrase	21	4

4.10 Citation after retraction, against matched controls

Of 9,142 machine-attributed records queried, 8,594 were paired under the initial design and 8,448 pairs had both members returned by OpenAlex (Table 15). Balance was exact on retraction year, exposure and publisher match and near-exact on publication year, but controls had accumulated more citations through the retraction year (mean 8.93 against 6.00, standardized mean difference − 0.152), which exceeds the conventional 0.10 threshold and indicates residual imbalance in citation visibility measured through the retraction year.

Under that design, machine-attributed articles appeared to be cited less after retraction than their own controls. Among the 8,407 pairs in which both members had at least one year of exposure, 55.44% against 64.98% were cited in a post-retraction calendar year, a paired difference of − 9.54 percentage points (95% CI − 10.98 to − 8.10, McNemar $p < 1e-15$). The gap narrowed with longer windows, from − 9.44 points in the first post-retraction calendar year to − 3.06 points across the first three. The median within-pair difference in citations per article-year was exactly zero, with a paired bootstrap interval of zero to zero, because most pairs accrue identical low rates; the mean within-pair difference was − 0.42 and the Wilcoxon signed-rank test on 6,403 non-zero

differences gave z = − 16.10, p < 1e-15. Retraction status was slightly better propagated for the machine-attributed articles (98.39% against 97.86%, paired difference 0.53 points, 95% CI 0.14 to 0.93, p = 0.010).

Rematching with citations accumulated through the retraction year built into the match key attenuates that difference substantially (Table 16). Requiring exact agreement on baseline citations, publication year, retraction year and publisher yields 4,747 pairs with a baseline standardized mean difference of zero by construction, 65.05% against 66.34% cited, a paired difference of − 1.29 percentage points (95% CI − 3.08 to 0.51), McNemar p = 0.167, and a conditional odds ratio of 0.937 (95% CI 0.856 to 1.026). Leaving publisher free while still matching exactly on baseline citations yields 7,698 pairs and a difference of − 3.81 percentage points (95% CI − 5.17 to − 2.45, conditional odds ratio 0.814, 95% CI 0.757 to 0.877), which shows that publisher carries part of the residual. Repeating the publisher-exact assignment under twenty different random seeds gives a mean difference of − 1.24 percentage points with a standard deviation of 0.09 and a range of − 1.45 to − 1.05; the number of eligible pairs is fixed at 4,747 by the match keys and does not vary with the seed, so only which eligible control is drawn changes.

We do not describe the difference as having disappeared. Failing to reject a null is not evidence of equivalence, and the publisher-exact interval still admits differences of about three percentage points in either direction. The defensible statement is that the crude deficit is substantially attenuated once comparisons are restricted to articles with the same prior citation count from the same publisher in the same years, that the remaining estimate is small and no longer statistically distinguishable from zero, and that its interval is not narrow enough to exclude a modest effect. Stratifying by citations accumulated through the retraction year under the original design shows the same pattern from another direction, with the difference largest where those citations are absent and reversing among the most cited pairs; that table is in Online Resource 3, and we do not emphasize the reversal because pairs in the highest band are not matched on that count within the band.

Table 15
Citation analysis selection, attrition and covariate balance under the initial design

Step or covariate	Value
Machine-attributed records with usable DOI and dates	9,142 records
Pairs formed 1:1 without replacement	8,594
Records not pairable	548 records
Duplicate returned records collapsed, affecting 143 distinct DOIs	no pair loss
Pairs lost: machine member not returned by OpenAlex	−92
Pairs lost: control member not returned by OpenAlex	−49
Pairs lost: neither member returned	−5
Complete pairs analyzed	8,448
Citations through retraction year, machine mean versus control mean (SMD)	6.00 versus 8.93 (− 0.152)
Publication year, machine versus control (SMD)	2021.54 versus 2021.51 (0.019)
Retraction year and exposure years (SMD)	exact (0.000)
Publisher matched exactly (proportion)	0.60

Table 16
Paired probability of citation in any post-retraction calendar year, under the initial design and after rematching on citations accumulated through the retraction year. Both rematched rows retain only the first retrieved record per DOI, consistent with the deduplication described in Section 3.9

Design	Pairs	Baseline SMD	Machine %	Control %	Paired difference, pp (95% CI)	Conditional OR (95% CI)	p
Initial: year and publisher only	8,407	−0.168	55.44	64.98	−9.54 (− 10.98 to − 8.10)	—	< 1e-15
Rematched: baseline citations, years and publisher exact	4,747	0.000	65.05	66.34	−1.29 (− 3.08 to 0.51)	0.937 (0.856–1.026)	0.167
Rematched: baseline citations and years exact, publisher free	7,698	0.000	56.34	60.15	−3.81 (− 5.17 to − 2.45)	0.814 (0.757–0.877)	4.8e-8
Rematched, publisher exact: 20 random control assignments	4,747 in every assignment	0.000	—	—	mean − 1.24, SD 0.09, range − 1.45 to − 1.05	—	—

5 Discussion

5.1 Principal findings

The category is large, recent and highly concentrated: 28 records before 2020, two in five recorded retractions in 2023, and 70% of records from a single publisher whose special-issue program was the subject of a documented mass-retraction episode in exactly that window.

Its timing is not distinctive once composition is respected. Time to retraction is no faster than for other retracted articles under cluster-robust inference, and the crude appearance of markedly tighter timing does not survive holding retraction year, publication year and publisher fixed: the dispersion ratio moves from 0.35 to 0.97 in a sample matched exactly on the publisher string, and the interquartile-range ratio to 1.00. A residual of about five days remains and is unlikely to be chance given the sample size, but it is two orders of magnitude smaller than the crude figure, and it does not survive removing the dominant publisher from both cohorts, where the direction reverses to + 23.5 days. Tight timing is not a signature of the label. This is a correction to the natural reading of the unadjusted numbers, and it is the main methodological result of the paper.

The paper-mill association widely attributed to the category attenuates from a crude risk ratio of 7.78 to 1.43 under stratification (cluster-robust 95% CI 1.36 to 1.52), did not exist before 2023, and reverses inside two large publishers. Given heterogeneity of Q = 248.2 on 7 degrees of freedom, the pooled figure describes no common effect, and the heterogeneity is the finding. Because both codes come from the same Reason field and frequently the same investigation, this is co-coding within one database and not independent evidence about article origin.

Under the broadened terminology search, 21 of 66,013 records match specific generative-AI-related terminology, four of them machine-attributed, and in all four the term describes the article's subject rather than its production. Eleven of twenty patterns match nothing. This says that the code is not being applied on the basis of explicit tool attribution in these fields; it does not say that generated content is absent from the articles.

An apparent post-retraction citation deficit attenuates from − 9.5 to − 1.3 percentage points once pairs are matched exactly on citations accumulated through the retraction year and on publisher, and is then no longer statistically distinguishable from zero, though the interval still admits a difference of about three points in either direction. Differences in citation visibility measured through the retraction year account for most of the crude association. Because that measure can include same-calendar-year post-retraction citations, we do not present this as a clean separation of confounding from effect.

5.2 What the reason code appears to capture, and what this study cannot say

Within this snapshot, the computer-generated-content reason is concentrated in a recent, publisher-specific mass-retraction episode and co-occurs overwhelmingly with peer-review and paper-mill codes, while showing no distinctive timing once composition is held fixed and no accompanying explicit tool attribution in the searched fields.

What that does not establish is what the underlying articles contain. Concentration in one publisher and one period is not evidence that a code was misapplied; a publisher conducting a coordinated investigation could in principle detect computer-generated material at scale, and the concentration would look identical. Our evidence concerns the distribution and co-coding structure of a database field, not the provenance of text. Distinguishing the two requires reading notices and, ultimately, articles.

This is why we frame the contribution as bounded. The defensible claim is that this reason code should not be used as a proxy for generative-language-model authorship without notice-level validation, and that any analysis using it must stratify by retraction year and publisher, since crude aggregate ratios conceal reversals and overstate association by roughly a factor of five. The stronger claim, that the label predominantly reflects something other than generative AI, would require the notice audit set out in Section 5.6.

5.3 Relation to prior work

Saqr (2026) reported a compressed median time to retraction for this category relative to other countries. Our results are compatible with that country-level contrast while showing that the category is not faster than the retraction record as a whole, and that its apparently tighter dispersion is largely compositional. A country-level comparison cannot separate those effects, which is the methodological point.

Our concentration pattern matches what Candal-Pedreira et al. (2022) reported for paper-mill retractions identified by other means, supporting substantial overlap in the 2023 wave, though our stratified analysis shows that overlap is period- and publisher-specific. The detection literature (Cabanac & Labbé, 2021; Cabanac et al., 2021; Labbé & Labbé, 2013) identifies machine text from linguistic evidence in individual articles; our data show that when such evidence reaches a notice it usually arrives alongside evidence of compromised peer review.

Studies of retracted articles whose subject is artificial intelligence (Guo et al., 2026; Samala et al., 2025; Sridharan & Sivaramakrishnan, 2026) report medians of 510 to 550 days and slower detection of AI-related misconduct. Those concern a different object, and our terminology search quantifies the conflation risk directly: 700 records mention artificial intelligence, almost all as subject matter, against 21 matching specific generative-AI-related terminology. Estimates of generative-AI prevalence in published text (Gray, 2024; Liang et al., 2024) concern a third object again.

Our citation results support the persistence literature rather than qualifying it. Bar-Ilan and Halevi (2017) and Hsiao and Schneider (2021) documented continued citation of retracted work, and more than half of these articles were cited in a post-retraction calendar year. That the crude deficit against matched controls is substantially attenuated after balancing citation visibility measured through the retraction year is the direction García-Romero et al. (2026) would predict, having found post-retraction citation driven mainly by the pre-retraction citation rate. We describe this as attenuation rather than as an identified mechanism, for the measurement reason set out in Section 4.10.

5.4 Implications

For users of the database, the category should not be treated as an indicator of generative-AI misuse. Any analysis using it should stratify by retraction year and publisher, and any timing comparison should hold publication year and publisher fixed, because the unadjusted contrast is dominated by composition.

For database maintainers, we offer the following as suggestions for analytic value rather than as criticism of a resource that is compiled largely by hand and made freely available, and on which this study depends entirely. Distinguishing random generation, automated paraphrasing and large-language-model

authorship would separate phenomena that currently share one code and that differ in mechanism, as Section 2.1 sets out, though we recognize this is only possible where a notice supplies the distinction. Recording the date a reason was assigned or revised, and flagging records whose codes were remapped when the taxonomy changed, would let analysts separate coding history from publisher behavior; the absence of such fields is what prevents us from quantifying that here. Marking records that share a retraction notice would let any analysis separate independent cases from batch events.

For publishers, the concentration result is relevant to where screening effort is likely to repay itself. The records in this category arise overwhelmingly from special-issue and proceedings workflows at a small number of venues rather than being spread across the literature, which suggests that submission pathways operating at volume and with delegated editorial control merit closer monitoring than the general submission stream. This is an observation about where these retractions originated, not an assessment of any publisher's current controls.

5.5 Limitations

The cohort is defined by attribution in notices and investigations, not by verified machine authorship, and under-recording is likely because many notices give no detailed reason.

The snapshot does not record when a record was added or when a reason was assigned or revised, so the date of coding cannot be separated from the date of retraction. Because the earlier category name is entirely absent while its successor is populated back to 2010, records were evidently remapped, but the size of that contribution to the temporal series is unquantifiable from this file.

Coverage is not exhaustive. Retraction Watch is the most complete available source but does not capture every retraction, and reporting practices differ across publishers, affecting both numerator and denominator of the annual shares.

Concentration limits generalizability: with 70% of records from one publisher, full-cohort results substantially describe that publisher's retraction behavior. A without-Hindawi variant accompanies the timing analyses in Tables 7 and 8; the co-coding and citation analyses instead hold publisher fixed by stratification and by exact matching, which addresses the same concern by a different route. We do not report a separate without-Hindawi variant of every analysis.

The matched dispersion analysis covers 5,119 of 9,161 machine-attributed records, because the remainder have no contemporaneous control at the same publisher. It is therefore weighted toward the larger publisher-year cells and is not a statement about the unmatched remainder. The two-sided Hindawi exclusion is stricter still, retaining 811 pairs, so its reversal of direction rests on a small subset and is reported as a bound rather than as a primary estimate.

The regime analysis uses publication date as a proxy for the availability of large language models and cannot establish that any article used one. Articles published shortly after a cut-point may have been written before it.

The terminology search covers Title and Notes metadata, not full notice texts. Absence of an explicit tool name in metadata is not evidence that the article contains no generated content.

Notice membership was identified by retraction DOI.Of the 9,161 cohort records, 9,152 carried a valid retraction DOI, none was blank and 9 carried the literal value "unavailable"; those 9 were treated as singleton clusters. The stratified co-coding analysis is a Mantel–Haenszel estimator rather than a multilevel model with random intercepts for journal, publisher and notice, which would be a stronger specification and is left to further work; the stratified estimate is now accompanied by a notice-cluster bootstrap interval, which is about 1.7 times wider than the independence-based one, so the reported uncertainty no longer understates dependence; a multilevel model would additionally borrow strength across strata, which the bootstrap does not.

For the citation analysis, the rematched design achieves exact balance on citations accumulated through the retraction year but does so within a fixed retrieved pool, which reduces the number of usable pairs and may select toward denser publisher-year cells. Field and article type were retrieved and are supplied in Online Resource 3, but a field-matched variant is not reported as a headline because coverage of the topic field was incomplete across the candidate pool. Where multiple OpenAlex records existed for one DOI the first was retained, which is arbitrary.

Finally, this is a cross-sectional analysis of records selected on the occurrence of retraction. It cannot estimate the probability that a machine-generated article is retracted, only how retracted articles carrying this attribution differ from other retracted articles.

5.6 Future work

The decisive next study is a notice-level audit. A stratified random sample of 300 to 500 full retraction notices, stratified by publisher, retraction year and paper-mill co-coding and coded independently by two raters into categories that separate random generation, automated paraphrasing, tortured phrases, explicit language-model use, generated figures, unspecified automation and no supporting detail, with agreement reported as Cohen's kappa or Krippendorff's alpha and disagreements adjudicated, would establish what the code is applied to. That is the evidence required to convert the bounded claim made here into a claim about article content, and Online Resource 1 provides the sampling frame for the terminology component.

Two further studies follow. A multilevel logistic model with random intercepts for journal, publisher and notice would improve on the stratified co-coding estimate and would supply dependence-aware uncertainty. Applying linguistic detection methods to articles published after 2023 would bound the under-recording suspected here.

6 Conclusion

We characterized the 9,161 records in a Retraction Watch retraction snapshot attributed to computer-aided or computer-generated content. The category grew from 28 records before 2020 to two in five recorded retractions in 2023 and is concentrated in one publisher and a handful of journals, in the same

window as a documented special-issue investigation at that publisher.

Two crude contrasts that invite over-reading do not survive adjustment. Timing appears far more tightly clustered than for other retractions, but holding retraction year, publication year and the exact publisher fixed moves the dispersion ratio from 0.35 to 0.97 and leaves a residual of about five days, and removing the dominant publisher from both cohorts reverses the direction entirely, so tight timing is not a signature of the label. An apparent post-retraction citation deficit of − 9.5 percentage points attenuates to − 1.3 and loses statistical distinguishability once pairs share a baseline citation count and a publisher. The paper-mill co-coding association attenuates from a risk ratio of 7.78 to 1.43 (cluster-robust 95% CI 1.36 to 1.52), is absent before 2023, and reverses within two large publishers, with heterogeneity so extreme that the pooled value describes no common effect.

Under a twenty-pattern terminology search, 21 records carry specific generative-AI-related terminology and none of the four machine-attributed among them indicates that the article's own text was generated. Because both the machine-attributed and paper-mill indicators come from a single Reason field, and because the searched fields are metadata rather than notices, these results describe the structure of a database category rather than the provenance of text. The practical conclusion is narrow and firm: this reason code should not be used as a proxy for generative-language-model authorship, and any analysis built on it must hold retraction year and publisher fixed before drawing a contrast.

Declarations

Supplementary information Online Resource 1: all 718 records matching generative-AI terminology, with snippets and adjudication columns. Online Resource 1b: individual adjudication of the 21 records matching specific generative-AI-related terminology. Online Resource 2: independent Python validation comparison table and script. Online Resource 3: all analysis scripts, derived result tables, the frozen OpenAlex extract and the complete grids abbreviated in Tables 13 and 14, with a README for reproduction.

Competing interests

The author has no competing interests to declare that are relevant to the content of this article.

Ethics approval

Not applicable. The study analyzes publicly available bibliographic records and involves no human participants or animals.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Use of generative AI

Documented in Section 3.10.

Funding

No funding was received for conducting this study.

Author Contribution

Jonas Mandalunes is the sole author. He conceived and designed the study, acquired and curated the data, wrote all analysis code, performed the analysis and interpretation, produced the figures and tables, drafted and critically revised the manuscript, and approved the submitted version. He agrees to be accountable for all aspects of the work.

Acknowledgments

The author thanks Retraction Watch and the Center for Scientific Integrity for compiling the retraction database, and Crossref for acquiring it and distributing it as an open resource. The author also thanks the OpenAlex team at OurResearch for maintaining an open bibliographic index. Neither organization was involved in the design, analysis or interpretation of this study.

Data Availability

All data are publicly available. The Retraction Watch database is distributed by Crossref at https://gitlab.com/crossref/retraction-watch-data (https://gitlab.com/crossref/retraction-watch-data/-/commit/dda8c8749c6ac35923f69162936c53fc9b9f5e3c); the analysed file is pinned at commit dda8c8749c6ac35923f69162936c53fc9b9f5e3c. Scripts, derived tables and the frozen OpenAlex extract are in Online Resource 3.

Code availability

All analysis scripts, including the matched dispersion reanalysis, the citation rematching script and the independent Python validation script, are provided in Online Resource 3 together with a README describing how to regenerate every table and figure from the public snapshot.

References

1. Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: A case study. *Scientometrics*, 113(1), 547–565. https://doi.org/10.1007/s11192-017-2242-0
2. Cabanac, G., & Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, 72(12), 1461–1476. https://doi.org/10.1002/asi.24495
3. Cabanac, G., Labbé, C., & Magazinov, A. (2021). *Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals*. arXiv. https://doi.org/10.48550/arXiv.2107.06751
4. Candal-Pedreira, C., Guerra-Tort, C., Ruano-Ravina, A., Freijedo-Fariñas, F., Rey-Brandariz, J., Ross, J. S., & Pérez-Ríos, M. (2024). Retracted papers originating from paper mills: A cross-sectional analysis of references and citations. *Journal of Clinical Epidemiology*, 172, 111397. https://doi.org/10.1016/j.jclinepi.2024.111397
5. Candal-Pedreira, C., Ross, J. S., Ruano-Ravina, A., Egilman, D. S., Fernández, E., & Pérez-Ríos, M. (2022). Retracted papers originating from paper mills: Cross sectional study. *BMJ*, 379, e071517. https://doi.org/10.1136/bmj-2022-071517
6. Crossref. (2024). *Retraction Watch data* [Data set]. GitLab. https://gitlab.com/crossref/retraction-watch-data
7. García-Romero, A., González-Mancebo, S., Espallardo-Ortega, C., & Concepción, M. A. (2026). The enduring influence of retracted research: Determinants and citation patterns across three disciplines. *Scientometrics*. https://doi.org/10.1007/s11192-026-05699-7
8. Gray, A. (2024). *ChatGPT "contamination": Estimating the prevalence of LLMs in the scholarly literature*. arXiv. https://doi.org/10.48550/arXiv.2403.16887
9. Guo, Z., Mao, H., & Zhang, F. (2026). The attention mismatch: Mapping the structural academic governance deficit in the age of generative AI. *Publications*, 14(2), 27. https://doi.org/10.3390/publications14020027
10. Hsiao, T.-K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, 2(4), 1144–1169. https://doi.org/10.1162/qss_a_00155
11. Khademizadeh, S., Danesh, F., Esmaeili, S., Lund, B., & Santos-d'Amorim, K. (2024). Evolution of retracted publications in the medical sciences: Citations analysis, bibliometrics, and altmetrics trends. *Accountability in Research*, 31(8), 1182–1197. https://doi.org/10.1080/08989621.2023.2223996
12. Kocyigit, B. F., Zhaksylyk, A., Akyol, A., & Yessirkepov, M. (2023). Characteristics of retracted publications from Kazakhstan: An analysis using the Retraction Watch database. *Journal of Korean Medical Science*, 38(46), e390. https://doi.org/10.3346/jkms.2023.38.e390
13. Kwee, R. M., & Kwee, T. C. (2023). Retracted publications in medical imaging literature: An analysis using the Retraction Watch database. *Academic Radiology*, 30(6), 1148–1152. https://doi.org/10.1016/j.acra.2022.06.025
14. Labbé, C., & Labbé, D. (2013). Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? *Scientometrics*, 94(1), 379–396. https://doi.org/10.1007/s11192-012-0781-y
15. Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). *Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews*. arXiv. https://doi.org/10.48550/arXiv.2403.07183
16. Oransky, I. (2023, September 12). *The Retraction Watch Database becomes completely open — and RW becomes far more sustainable*. Retraction Watch. https://retractionwatch.com/2023/09/12/the-retraction-watch-database-becomes-completely-open-and-rw-becomes-far-more-sustainable/
17. Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv. https://doi.org/10.48550/arXiv.2205.01833
18. Retraction Watch. (2025a). *Retraction Watch database user guide appendix B: Reasons*. https://retractionwatch.com/retraction-watch-database-user-guide/retraction-watch-database-user-guide-appendix-b-reasons/
19. Retraction Watch. (2025b). *Retraction Watch database user guide appendix D: Changes*. https://retractionwatch.com/retraction-watch-database-user-guide/retraction-watch-database-user-guide-appendix-d-changes/
20. Samala, A. D., Ferdian, F., Fiandra, Y. A., Hanafiah, M. H., Rawas, S., & Asad, M. M. (2025). Retraction trends in Indonesian research: Misconduct, AI risks, and policy gaps. *Quality & Quantity*, 60, 4507–4531. https://doi.org/10.1007/s11135-025-02453-8
21. Santos-d'Amorim, K., Sanz-Casado, E., & dos Santos, R. N. M. (2025). Errors, questionable practices, or misconduct? A bibliometric and altmetric review covering two decades of retractions in Latin America. *Scientometrics*, 130(7), 3679–3706. https://doi.org/10.1007/s11192-025-05343-w
22. Saqr, K. (2026). *Dissecting AI-related paper retraction across countries and institutions* [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202601.0314.v1
23. Sridharan, K., & Sivaramakrishnan, G. (2026). Artificial intelligence in the retraction spotlight: Trends, causes and consequences of withdrawn AI literature through a systematic bibliometric review. *Frontiers in Research Metrics and Analytics*, 10. https://doi.org/10.3389/frma.2025.1737168

24. Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLoS ONE*, 8(7), e68397. https://doi.org/10.1371/journal.pone.0068397
25. Van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023 — a new record. *Nature*, 624, 479–481. https://doi.org/10.1038/d41586-023-03974-8

Figures

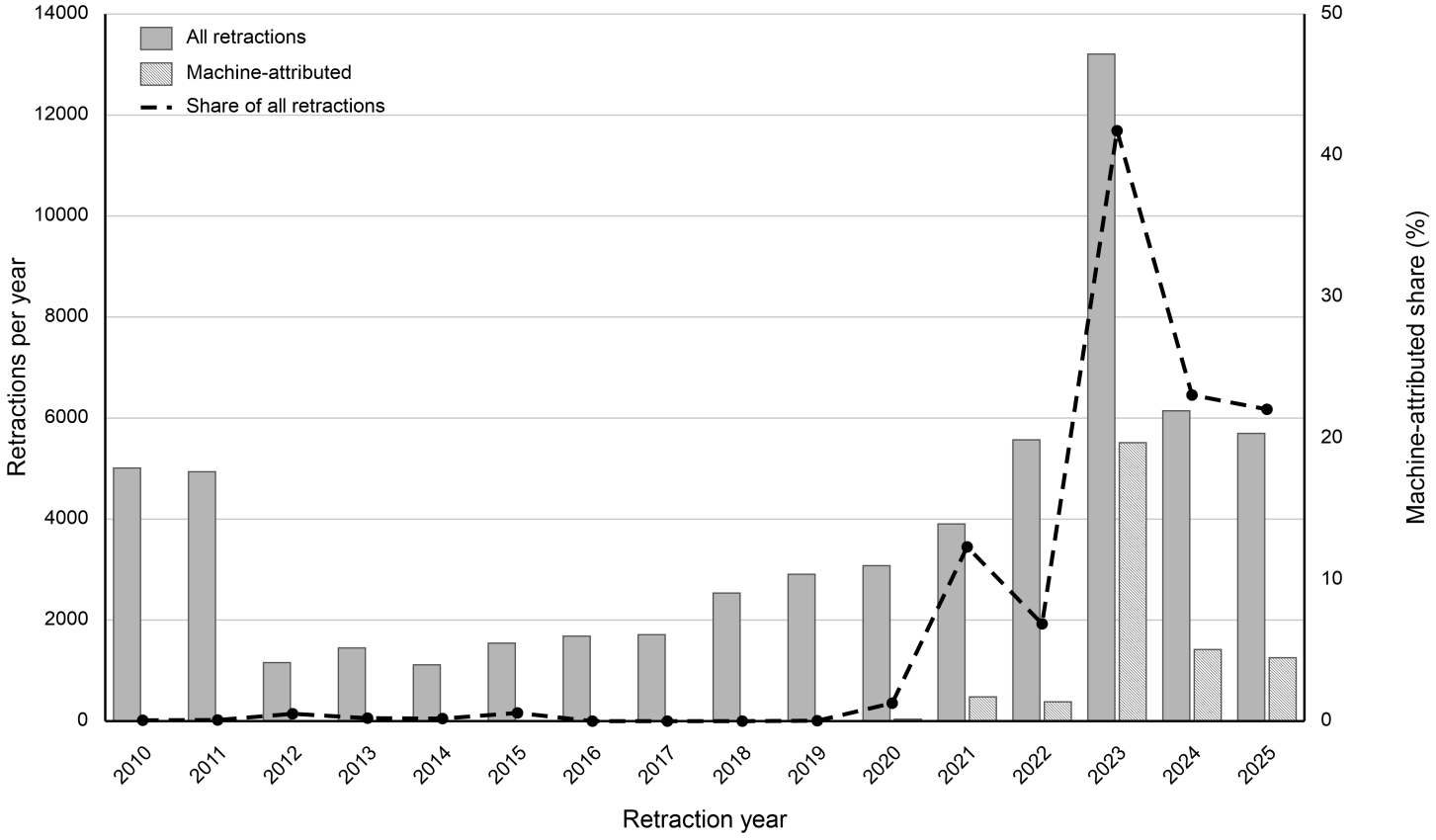


Figure 1

Annual count of all retractions and of machine-attributed retractions, 2010 to 2025, with the machine-attributed share of all retractions on the secondary axis. Series are distinguished by fill pattern as well as shade. Data from Table 3

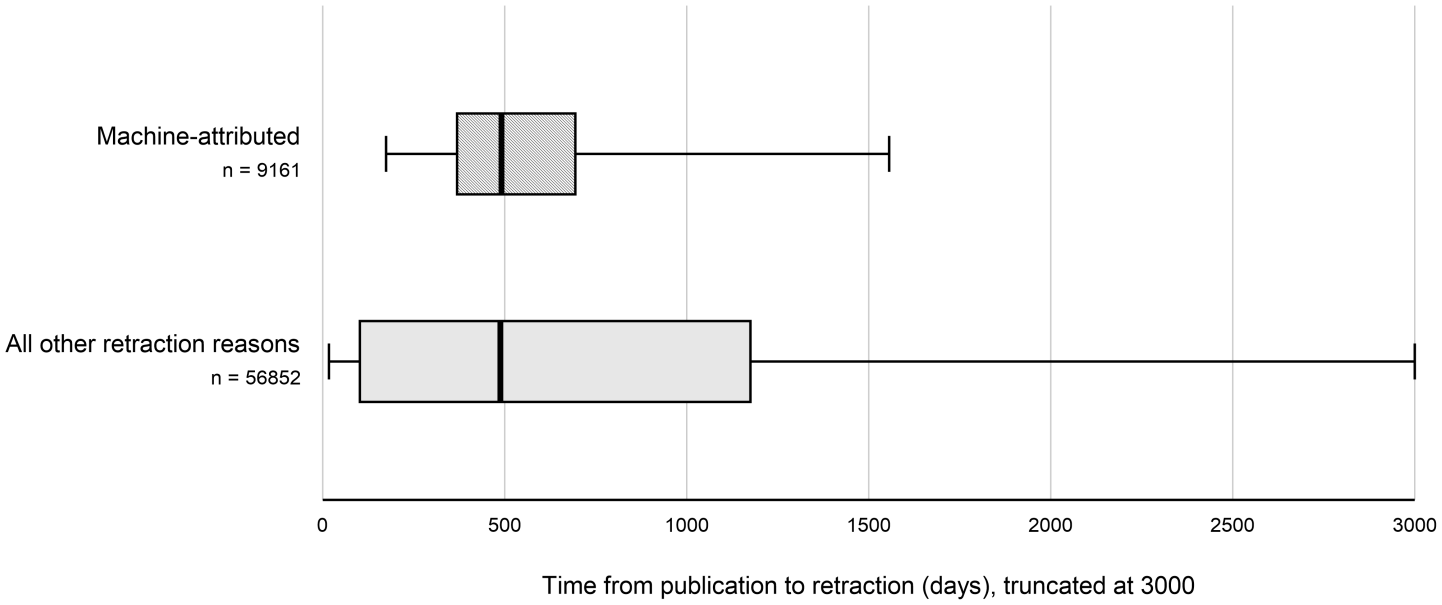


Figure 2

Distribution of time from publication to retraction for the machine-attributed cohort and the comparison cohort. Boxes span the interquartile range, the interior line marks the median, and whiskers span the 5th to 95th percentiles. The horizontal axis is truncated at 3,000 days for display, so the upper whisker of the comparison cohort is clipped. Data from Table 5

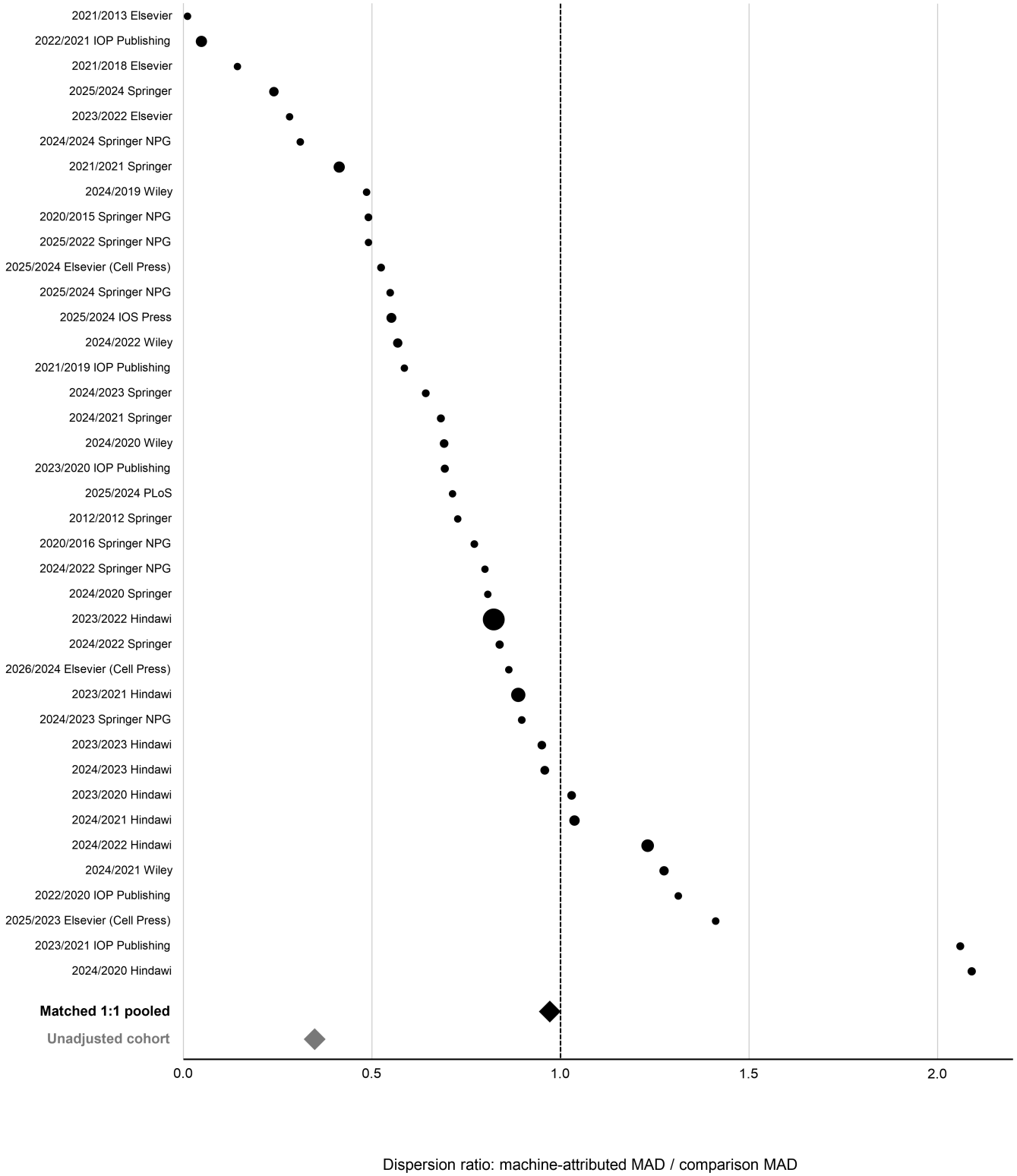


Figure 3

Stratum-specific dispersion ratios for time to retraction, machine-attributed mean absolute deviation from the median divided by the comparison mean absolute deviation from the median, for the 39 exact-publisher strata containing at least five records in each arm. Strata are labeled by retraction year, publication year and publisher, and ordered by ratio. The publisher-group specification, which yields 46 strata, is plotted in Online Resource 3 and is not shown here. Circles denote strata, with area proportional to the number of machine-attributed records; the dashed vertical line at 1.0 denotes equal dispersion, so circles to its left indicate a cohort less dispersed than its comparison. The black diamond is the pooled estimate from the 1:1 matched sample and the grey diamond the unadjusted whole-cohort estimate. Data from Table 7 and Online Resource 3

Supplementary Files

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

- ESM3scriptsandtables.zip
- ESM1baitermadjudication.csv
- ESM2pythonvalidation.md
- ESM2validatestats.py
- ESM1aitermmatches.csv