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.

Divergent responses to AI-era submission pressure: a comparative analysis of medicine and computer science, 2015–2026

Kangwei Wang

wkwyqrmyy@wmu.edu.cn

Yueqing People's Hospital

Ce Shi

Wenzhou Vocational College of Science and Technology

Research Article

Keywords: scholarly publishing, generative AI, peer review, preprints, research integrity, bibliometrics

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11412416/v1

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

Additional Declarations: No competing interests reported.

Divergent responses to AI-era submission pressure: a comparative analysis of medicine and computer science, 2015–2026

Kangwei Wang[1], Ce Shi[2]

[1] The Affiliated Yueqing Hospital of Wenzhou Medical University, Yueqing, Wenzhou, Zhejiang, China

[2] School of Digital Economics, Wenzhou Vocational College of Science and Technology, Wenzhou, China

Corresponding author: Kangwei Wang. Email: wkwyqrmyy@wmu.edu.cn. ORCID: 0009-0002-2553-8862. Co-author ORCID: Ce Shi, 0009-0008-5780-9691.

Abstract

ChatGPT's release in late 2022 and the arrival of agentic AI systems from 2025 coincided with reports of rising manuscript volumes and editorial strain. Whether this pressure is uniform across research fields, or absorbed differently by each, has not been tested comparatively. We contrast two ecosystems with sharply different barriers to entry: computer science (CS), where submission carries almost no procedural cost, and medicine, where ethics review and data governance constrain what can be produced. Using open data for 2015–2026 — monthly submission and output volumes from arXiv, medRxiv, bioRxiv and PubMed; submission and acceptance figures for thirteen CS conferences; and review times for 38 medical journals — we measured AI-associated writing with a 95-word excess-vocabulary detector applied identically to 30,000 CS abstracts and 13,000–24,000 PubMed abstracts per year. CS preprint submissions rose 99.3% between 2022 and 2025, against 38–41% for the life-science preprint servers and 5.5% for PubMed-indexed output, and several medical specialties contracted; the divergence holds within arXiv itself. CS conferences absorbed the surge by expanding capacity, with NeurIPS submissions up 195% at an unchanged acceptance rate. AI-associated writing rose sharply from 2023 in both fields, the proportional increase nearly identical from a higher CS baseline, with the break appearing in all eight medical specialties tested. Review time for accepted papers showed no aggregate change. Medicine and CS therefore experienced comparable growth in AI-associated writing but opposite responses: CS expanded output and review capacity, while biomedicine held output flat and routed the pressure into editorial triage.

Keywords: scholarly publishing; generative AI; peer review; preprints; research integrity; bibliometrics

1. Introduction

1.1 A sudden change in the cost of writing

In November 2022, ChatGPT placed fluent academic prose generation within reach of any researcher with a browser. The change was immediately visible in the scholarly record: in the three years that followed, word frequencies in biomedical abstracts shifted in ways that a decade of prior trend could not explain (Kobak et al., 2025). By 2025, agentic systems could plan, retrieve, analyse and draft with limited supervision (Dell'Acqua et al., 2023), and by October 2026 arXiv introduced per-author submission limits of two papers per month after single-month submissions exceeded 40,000 for the first time (arXiv, 2026).

These developments coincided with widespread reports of strain in peer review. *Nature* described an overloaded system (The peer-review crisis, 2025); publisher surveys documented reviewer shortages and

declining willingness to review (NISO, 2026; Silverchair, 2026); and editorials in management and information systems journals warned of a structural crisis rather than a temporary backlog (Organization Science, 2026). Meanwhile experimental evidence indicated that generative tools raise individual research productivity (Noy & Zhang, 2023; Toner-Rodgers, 2024), which makes a rise in submitted volume the expected downstream consequence rather than a surprise.

What has not been established is whether this pressure is *uniform* across research fields, or whether different disciplines experience and absorb it differently. That distinction matters both for explanatory accuracy — an undifferentiated "AI submission crisis" narrative may be wrong in ways that mislead policy — and for practical advice to researchers deciding where and how to submit.

1.2 Two ecosystems with different barriers to entry

Generative tools lower the cost of *writing*. They do not lower the cost of *doing* research. This asymmetry suggests that fields should respond very differently depending on which cost binds.

Computer science (CS) sits at one extreme. A paper can be drafted and posted to arXiv immediately; conference submission involves no ethical approval, no trial registration, no data-access agreement, and review capacity that can be expanded by enlarging a programme committee. The marginal cost of an additional submission is close to zero.

Medicine sits at the other. A clinical study requires ethics approval, often prospective registration and data-governance steps, before it can be conducted at all. Journals have close to fixed page budgets and depend on volunteer reviewers. Adoption of generative tools in writing may be slowed further by early, explicit and comparatively restrictive AI-disclosure policies at the leading medical journals, and by the heightened institutional scrutiny that accompanies clinical research.

If the binding constraint in a field is *doing*, cheaper writing should produce little growth in submission volume — and whatever pressure does arise should surface as congestion in the review system. If the binding constraint is *writing*, cheaper writing should produce near-immediate volume growth, followed by adaptation on the capacity side. These are testable predictions, and they imply that the same upstream technology may generate two different "crises" that happen to share a name.

1.3 Prior work and the gap

Three bodies of quantitative work are directly relevant. Each studies a single ecosystem, and this journal has published much of the recent work in the area.

Biomedical corpora. Kobak et al. (2025) introduced the excess-vocabulary method: words whose frequency in a corpus rises anomalously after ChatGPT serve as markers of LLM-assisted writing, and the share of documents containing such words estimates prevalence. Applied to biomedical abstracts, the method revealed a sharp post-2023 break and estimated that a substantial minority of recent abstracts carry LLM signals. The approach has since been replicated in medical education (Matsui, 2025) and extended to medical abstract style specifically, where Qiu and Pan (2026) characterised lexical-complexity differences between model-generated and human-written medical abstracts.

Preprint corpora. Bao et al. (2025) documented linguistic shifts in preprint writing before and after the launch of ChatGPT, establishing that the change is visible in preprints as well as in indexed literature.

Citation consequences. Paklina et al. (2026) analysed the association between LLM-assisted writing and citation advantage across scientific publications, again on a single corpus.

Liang et al. (2024) separately examined AI-modified content at scale in computer science venues.

The gap. Every one of these studies is single-ecosystem. The strata that have been examined are biomedical abstracts, preprints of unspecified discipline, and one CS corpus — never the two side by side, and never with a common instrument. This matters because the practical question researchers face is comparative: not "is AI-assisted writing present?" but "how is the field I publish in responding, and what does that mean for how I should submit?"

The limitation is acknowledged in the work itself. Paklina et al. (2026) note that their empirical analysis rests on an arXiv-linked corpus, which they describe as disproportionately concentrated in the quantitative sciences, and they flag the resulting constraint on generalisability. That is the gap this paper addresses: by measuring medicine and CS on the same detector over the same period, we test whether the patterns documented in one corpus hold in another, and whether the fields differ in ways that a single-corpus design cannot reveal.

A second gap is substantive rather than geographic. Existing work connects AI assistance to *writing* outcomes — vocabulary, complexity, citations — but not to *editorial* outcomes: submission volume, acceptance rates, review duration. The literature therefore documents that writing changed, and separately that publishing is under strain, without establishing how the two relate or whether the relationship is the same across fields.

Contribution relative to prior work. The distinction is summarised as follows. Where prior work asks whether AI-assisted writing is detectable in a given corpus, we ask how two ecosystems with different barriers to entry absorb the same upstream change. Where prior work measures writing outcomes, we measure writing *and* editorial outcomes on a common time axis. And where prior work is explicitly constrained to one discipline's corpus, the cross-ecosystem design is the contribution rather than a limitation.

1.4 This study

We assemble open data covering 2015–2026 for both ecosystems and measure three things with a common instrument and a common time axis:

1. **Submission and output volume** — monthly series for arXiv (all fields and, reconstructed from a full metadata snapshot, by category), medRxiv, bioRxiv, and PubMed (overall and by specialty); plus submission and acceptance figures for thirteen CS conferences.
2. **AI-associated writing** — prevalence of excess-vocabulary markers in 30,000 CS abstracts and 13,000–24,000 PubMed abstracts per year, using a detector identified *independently within each corpus* and then restricted to a common word set so that the two fields are measured on the same instrument.
3. **Editorial outcomes** — conference acceptance rates, and submission-to-acceptance times for a fixed panel of medical journals.

We address three questions. Each is stated with the directional expectation that follows from the barriers-to-entry argument in §1.2, so that the results can be read against a prediction rather than merely described.

- **Q1 (volume).** Do submission and output volumes diverge between the two ecosystems after 2022, and is the change localised to a dated breakpoint? *We expect CS volume growth to substantially exceed medicine's, because cheaper writing expands the volume of work submitted wherever writing — rather than doing — is the binding constraint.*
- **Q2 (writing).** Does the prevalence of AI-associated writing markers rise in both fields, and if so, to the same extent? *We expect it to rise in both, since the tools are equally available to both communities. Whether the* extent *is similar is genuinely open: restrictive disclosure norms and institutional scrutiny in medicine could damp adoption, or editors' and reviewers' expectations could leave the underlying writing behaviour largely unchanged. We therefore treat the magnitude as an empirical question rather than a prediction, and consider a comparable proportional rise the more informative outcome.*
- **Q3 (absorption).** How does each ecosystem absorb the resulting pressure — through capacity expansion, or through congestion? *We expect the two to differ by capacity regime: CS venues can add review capacity and should absorb growth largely through expanded output at a stable acceptance rate, whereas medicine, with close-to-fixed capacity, should route the pressure into rejection and queueing rather than into additional published output.*

A fourth question — whether marker prevalence at journal level predicts changes in editorial outcomes — is reported descriptively rather than modelled. The association would be jointly determined by journal reputation, editorial policy, field growth and author-language composition, none of which this design observes, so a fitted coefficient would not be interpretable. We report the observed pattern and note that identifying the question properly requires within-journal submission-level data.

1.5 Contribution

First, we provide the first side-by-side measurement of AI-era submission pressure in medicine and CS using an identical detector, which allows the question "is this one phenomenon or two?" to be asked directly. Second, we show that the proportional growth in AI-associated writing is nearly identical across the two fields while their volume responses differ by more than an order of magnitude — a dissociation that a single-field study could not detect. Third, we translate the resulting pattern into practical guidance for researchers deciding how to write and where to submit.

1.6 Scope and claims

The design is ecological. We measure populations of documents and platform-level volumes, not individual manuscripts. We therefore report temporal co-movement and cross-ecosystem difference, and we do not claim that generative AI *caused* any specific volume change; funding cycles, open-access mandates and post-pandemic recovery are among the plausible contributors that our design cannot exclude. Where we use the phrase "AI-associated", it denotes stylistic markers statistically associated with the post-2022 period, not confirmed use of any particular tool.

2. Methods

2.1 Study design

Comparative bibliometric study of two research ecosystems with sharply different barriers to entry, over January 2015 – October 2026. We use population-level series (submission and output volumes) and abstract-level text measurement (prevalence of AI-associated writing markers). The design is ecological: it

supports statements about temporal co-movement and cross-ecosystem difference, not about individual papers or causal attribution to any single technology.

The two ecosystems are:

- **Computer science (CS)** — proxied by arXiv `cs.*` and by thirteen CS conferences. Submission cost is essentially zero: no ethics review, no data access agreement, and conference review capacity that can be expanded.
- **Medicine** — proxied by PubMed-indexed journal articles, plus the medRxiv and bioRxiv preprint servers. Submission requires ethics approval, frequently trial registration and data-governance steps, and journals have close to fixed editorial capacity.

2.2 Data sources and acquisition

All data are open and were collected between 2026-10-16 and 2026-10-17. Each source, endpoint, coverage and derived field is listed below and in the data dictionary (`logs/p0_validation_report.md`). Sources are cited in the reference list: arXiv statistics and the metadata snapshot (arXiv, n.d.; Hugging Face, n.d.), the preprint server APIs (medRxiv & bioRxiv, n.d.), NCBI E-utilities (National Center for Biotechnology Information, n.d.), conference statistics (Xu, n.d.), and the published marker vocabulary and analysis code (Berens lab, n.d.; Kobak et al., 2025).

- **arXiv submission statistics** — `arxiv.org/stats/get_monthly_submissions`; monthly counts from 1991-07 to 2026-10 (424 months), official.
- **arXiv metadata snapshot** — `librarian-bots/arxiv-metadata-snapshot` on Hugging Face (11 Parquet files, 2.5 GB); 3,205,432 papers, category and first-version submission date from `versions[0].created`, abstracts for the CS subset.
- **medRxiv / bioRxiv** — `/details/{server}/{start}/{end}/0` on `api.medrxiv.org` and `api.biorxiv.org`; monthly posting totals from 2019-06 and 2013-12 respectively to 2026-10 (244 months, no missing values).
- **PubMed output volume** — NCBI E-utilities `esearch` with `retmax=0`; annual counts by MeSH-defined specialty and by publication type, 2000–2026.
- **PubMed abstracts** — `esearch` plus `efetch` (POST); 211,817 records from 2018-01 to 2026-09.
- **CS conference statistics** — cs-conf-stats (`github.com/Xovee/cs-conf-stats`); 13 conferences, 351 conference-years, 1988–2026.
- **Publication-to-acceptance times** — PubMed `<History> PubStatus="received"` and `"accepted"`, 2015–2026, restricted to the journal panel in §2.5.

Two acquisition notes are material to interpretation.

First, arXiv's public statistics endpoint returns only the all-field total and ignores category parameters. Category-level series were therefore reconstructed from the full metadata snapshot. A submission date is taken as the timestamp of version 1 (`versions[0].created`), i.e. the original posting, not the most recent revision. All 3,205,432 records yielded a parseable date (0 failures).

Second, the PubMed sample is stratified by publication month rather than drawn at random from the year. For each month we ran a count query restricted to `Journal Article[pt] AND hasabstract[text]`, then retrieved the first 2,000 records in the default result order and fetched them with `efetch`. This excludes books, editorials and records without abstracts, and produces roughly

uniform coverage across the calendar year. Systematic rather than random sampling means that no formal sampling weights are applied. We report sample sizes with every estimate and attach binomial (Wilson) confidence intervals to all prevalence figures; because the sampling is systematic rather than random, these intervals should be read as approximate.[^ci]

[^ci]: Wilson score intervals assume independent Bernoulli draws. Under systematic sampling from an ordered result list, the effective variance may differ from the binomial ideal, so intervals indicate the precision implied by sample size rather than a guaranteed coverage level.

What the PubMed sample represents. The general sample is drawn uniformly from all indexed records, not from clinical journals specifically, so it reflects the full breadth of what PubMed indexes. That breadth is deliberate — the comparison of interest is with all of biomedical publishing — but it means the sample includes many small and non-clinical titles, and that the leading clinical journals contribute only a handful of records each. The resulting estimate should therefore be read as representing PubMed-indexed biomedical literature as a whole, not high-impact clinical research in particular. The specialty-stratified samples described in §2.3 provide an independent check on whether this matters: in 2025 the general sample gives 26.0% of abstracts at the ≥3 threshold, while the eight MeSH-defined specialties, drawn directly and with no dependence on the general sample, range from 17.4% to 27.7% (median 24.3%). The general estimate falls within that range, so the breadth of the sampling frame does not appear to distort the level or the trend reported here.

2.3 Measuring AI-associated writing

We adopted the excess-vocabulary approach of Kobak et al. (2025), in which words whose frequency rises abnormally after the release of ChatGPT serve as markers of LLM-assisted writing.

Independent marker identification per corpus. Applying a purely data-driven selection to CS abstracts recovers topic drift rather than prose style — the highest-ranked "excess" terms are `mamba`, `mllm`, `agentic`, `rag`, `llama`, `deepseek` and `claude`, which reflect what CS papers are *about* after 2022, not how they are *written*. We therefore restricted the candidate vocabulary to writing-style words before selection.

Procedure. Applied separately to each corpus:

4. Tokenise each abstract to lowercase alphabetic tokens (`[a-z]{3,}`), forming a set per document.
5. Restrict candidates to the 434 style-type words of the published 900-word reference table (content and domain terms excluded).
6. Compute each candidate's document frequency by year.
7. Baseline rate = mean over 2018–2022 (pre-ChatGPT window).
8. Recent rate = mean over 2024–2025.
9. Retain words with excess ratio ≥ 2.0 *and* recent rate ≥ 0.2%.

This yielded **171 CS markers** and **123 medicine markers**, listed in full in Table S1. Of these, 121 and 92 respectively appear in the published Kobak list, which we treat as cross-corpus validity evidence rather than a target.

Prevalence measures. For each corpus-year we report the share of abstracts containing at least 1, 2, 3 or 5 markers, and the mean number of distinct markers per abstract. Because a single marker can occur by chance in a long abstract, the threshold sweep, rather than any single cut, carries the signal; the cross-corpus comparison rests on the *trend*, which is robust to the threshold chosen.

Common detector. Marker lists of different sizes (171 vs 123) are not directly comparable, so all CS-versus-medicine comparisons use the **95 words identified as excess in both corpora** (71 of which are in the published list; Table S2). Using one shared detector removes list-size as an explanation for any observed difference in level.

Corpora: 30,000 arXiv `cs.*` abstracts per year for 2015–2026 (397,225 abstracts; capped per year to balance the panel), and the full PubMed sample by publication year.

Subfield samples. For the disaggregated analysis (Tables S3–S4, Fig. S1) the medicine corpora above are not reused, because MeSH descriptors are present on only ~25% of records in the general sample and leave specialty groups too small to test. Instead each specialty was sampled directly with the same MeSH query PubMed uses for its own indexing (e.g. `"Neoplasms"[MeSH Terms]`), so hierarchical expansion is handled by PubMed and every retrieved abstract belongs to the specialty by construction (107,987 abstracts; 8 specialties × 2018–2026 × ~1,500 per specialty-year). These queries are identical to those used for the population counts in §3.1, so counts and samples refer to the same definitions. The CS subfield analysis groups the arXiv sample above by primary category; cross-listed papers are counted once, under the first-listed category.

2.4 Volume and breakpoint analysis

For each monthly series we estimated a segmented regression with harmonic seasonality and heteroskedasticity- and autocorrelation-consistent standard errors (Newey & West, 1987; 12 lags), following the interrupted time-series approach described by Lopez Bernal et al. (2017, 2018):

$$y_t = \beta_0 + \beta_1 \cdot t + \Sigma_k \left[\beta_{D,k} \cdot D_k + \beta_{S,k} \cdot (t - T_k) \cdot D_k \right] + \gamma_1 \cdot \sin(2\pi t/12) + \gamma_2 \cdot \cos(2\pi t/12) + \varepsilon_t$$

where t indexes months and D_k is an indicator for $t \geq T_k$. Breakpoints were fixed a priori at **2022-11** (public release of ChatGPT) and **2025-01** (general availability of agentic systems), not selected from the data. We report the pre-break slope, the slope change at each break, and the resulting post-break slope. Level-shift terms are estimated but not emphasised, because in this parameterisation they are collinear with the slope terms.

Preprint series (medRxiv, bioRxiv) begin in 2021 rather than 2015: medRxiv launched in mid-2019, and both servers show a COVID-era peak followed by decline, so a pre-2020 baseline would conflate post-pandemic recovery with any AI-era effect. The arXiv series, which show no such distortion, begin in 2015.

Alongside the models we report annual volumes and simple growth ratios (2019→2022 and 2022→2025), plus an **acceleration factor** = (2022→2025 growth) ÷ (2019→2022 growth), which is 1.0 when a trend continues unchanged and >1 when it steepens.

2.5 Editorial outcomes

CS conferences. Submission and acceptance counts are taken as published by each meeting (via cs-conf-stats) and are used as reported, without modelling. For comparability across meetings of very different size, submissions are indexed to 2019 = 100.

Medical journals. Submission-to-acceptance time was to be derived from the `received` and `accepted` timestamps that some journals deposit in PubMed `<History>` metadata, computed as `accepted − received` in days and summarised by publication year for a fixed journal panel.

A pre-analysis validation step established a material constraint on this measure. Sampling ~60 recent articles from each of twelve leading medical journals, only about **32% deposit both dates**, and deposition is determined by publisher rather than by field or prestige:

Received and accepted dates are deposited by the Lancet titles (67–78%), BMC Medicine (95%), Journal of Clinical Epidemiology (95%), PLoS Medicine (68%) and Nature Medicine (43%), while NEJM, JAMA, BMJ, Annals of Internal Medicine, Radiology, Journal of Clinical Oncology and Circulation deposit neither date.

Crucially, within depositing journals the deposit rate is **stable over time** (94–100% in every year from 2018 to 2025), so the restriction should not generate a spurious time trend. The consequence is nonetheless that this measure rests on a publisher-biased subset and is reported as *supporting* evidence rather than as a primary outcome. The full journal panel is reported with the results.

Because acceptance rates for individual medical journals are not published as consistent time series, we do not report journal-level acceptance-rate trends. Journal-level association between marker prevalence and review-time change is reported as **exploratory only** (§3.5): journal reputation, editorial policy, field growth and author-language composition are all plausible common causes, and the design cannot separate them.

Why review time is reported "conditional on acceptance". The `received → accepted` interval exists only for accepted manuscripts, so manuscripts rejected before or during review never enter the sample. A journal absorbing volume pressure by raising its desk-rejection rate will therefore show an *unchanged* interval for the papers it does accept — the rejected ones leave the denominator rather than shortening the interval. This is a structural property of the measure, not a data-quality problem, and it is stated explicitly wherever the result is used (§3.3, §4.4).

A direct measure of triage speed — time to first decision, available in publisher-reported journal metrics — would close this gap. We attempted to reconstruct such a series from archived snapshots of publisher journal-metric pages via the Internet Archive; the endpoint proved unreliable for this purpose and the attempt was abandoned rather than reported. Instead we corroborate the triage mechanism with published editorial statistics covering the same period (desk-rejection growth, rejections per acceptance, reviewer invitation acceptance rates), treated as independent supporting evidence rather than as a controlled series (§3.3).

2.6 Design safeguards

Five design choices carry the interpretive weight of this study. Each was adopted to close off a specific alternative explanation, and together they — rather than the limitations section — define what the results can support.

Independent marker identification within each corpus. The detector vocabulary is derived separately for CS and for medicine rather than borrowed from one field and imported into the other. This is not a refinement but a requirement: a naively data-driven selection applied to CS returns *topic* markers — `mamba`, `mllm`, `agentic`, `rag`, `llama`, `claude` — because post-2022 CS papers are about these things. Any measure built on such a list would track shifts in research subject matter while presenting itself as a measure of writing. We verified the correction empirically: the style-restricted CS list is led by `underscoring`, `underscore`, `overlooking`, `underexplored` and `prompting`, with 121 of 171 markers also present in the published biomedical list.

A common detector for cross-corpus comparison. Per-corpus lists differ in size (171 vs 123), and a longer list mechanically produces higher measured prevalence. All CS-versus-medicine comparisons therefore use the 95 words identified as excess in *both* corpora, which removes list length as a candidate explanation for any level difference between fields.

Within-platform control. arXiv provides a natural control unavailable in cross-platform designs: the same submission process, the same infrastructure and the same policies govern `cs.*` and, say, `math.*`. If the post-2022 surge were driven by platform-level factors — arXiv policy, a general shift toward preprinting, or increased overall research activity — it would appear across categories. The observation that `cs.*` grew 271% from 2018 to 2025 while mathematics grew 32% and physics 38% therefore localises the change to the CS community rather than to the platform.

A fixed journal panel for review times. Review-time trends are computed over journals that each contribute ≥500 fully dated records, so that the panel's composition cannot change over time and manufacture a trend. Depositing journals are additionally verified to have stable deposit rates across years (94–100%), which rules out a change in *what gets reported* being mistaken for a change in *how long review takes*.

Breakpoints fixed a priori. The ChatGPT (2022-11) and agent-era (2025-01) breakpoints were specified before estimation and not selected by searching for the best-fitting split. Data-driven breakpoint selection would inflate apparent significance and is a known route to spurious structural-change findings.

Two further verifications support the measurement itself. The pipeline reproduces the published per-word trend for reference markers before extension (the frequency of *delves* in biomedical abstracts rises from 0.0075% in 2022 to 0.36% in 2024), and the medicine prevalence we estimate is of the same order as the published biomedical estimate. Both per-corpus marker lists, the common detector and the selection code are released as supplementary material, so the marker-identification step can be audited or re-run on other corpora.

2.7 Software and reproducibility

Analysis is in Python 3.12 (pandas 3.0, statsmodels, pyarrow, matplotlib). Scripts are numbered by pipeline stage in `scripts/`; each writes its output to `data/` (raw), `analysis/` (derived) or `figures/`.

Quality control includes: cross-checking each collected series against an independent published aggregate where one exists; confirming 0 date-parse failures on the full arXiv snapshot; verifying the completeness of the preprint series (0 missing months); and inspecting every figure after rendering.

2.8 Ethics

No human participants, no personal data, no intervention. All sources are public scholarly metadata and text. No ethics approval was required.

2.9 Use of large language models

In accordance with the journal's policy on large language models, we disclose their use here and confirm human accountability for the final text.

Large language models were used in two capacities.

In the research process. Analysis code was written with AI assistance. Every script was executed and its output inspected by the authors; all numerical results reported in this paper were produced by running that code on the public data described in §2.2, and the derived files are listed in `README_DATA.md`. No result was generated, estimated or transcribed by a language model.

In preparing this manuscript. The text was drafted with AI assistance and then reviewed, revised and approved by the authors, who directed the study design, the analytical choices, the interpretation and the conclusions.

No large language model is an author, and none meets the journal's authorship criteria: authorship carries accountability that cannot be applied to a model. The exemption in the journal's policy for "AI assisted copy editing" — defined there as improvements to readability, style, grammar, spelling, punctuation and tone — covers part of the editing performed. The assistance described above extends beyond that exemption, because it included drafting and generative editorial work, and is therefore declared here rather than omitted. The authors take full responsibility for the content of the final version and confirm that it reflects their original work.

3. Results

All prevalence figures carry Wilson 95% intervals (approximate, see §2.2). Percentages are of abstracts in the stated corpus and year.

3.1 The volume divergence: CS doubled, biomedicine was flat

Between 2022 and 2025, arXiv `cs.*` submissions rose **99.3%** (64,663 → 128,857). Over the same period medRxiv rose 40.6% and bioRxiv 38.3%. Biomedical journal output indexed in PubMed rose only **5.5%** (1,783,569 → 1,881,982) — and this small increase sits inside the range of ordinary year-to-year variation for PubMed, whose annual output has grown by roughly 2–6% per year for two decades.

Several medical specialties contracted outright over the same period, with cardiology and infectious disease falling most sharply (Table 1, panel c).

(Table 1; Fig. 1.)

Table 1 Annual submission and output volumes by series, with growth over 2019→2022 and 2022→2025, and the ratio of the two (acceleration factor).

(a) arXiv

Series	2019	2022	2025	Growth 2019→2022 (%)	Growth 2022→2025 (%)	Acceleration factor
arXiv cs.*	43316	64663	128857	49.3	99.3	2.01
arXiv econ.*	669	1320	2307	97.3	74.8	0.77
arXiv q-bio.*	1856	1946	3220	4.8	65.5	13.5
arXiv q-fin.*	793	819	1207	3.3	47.4	14.45
arXiv eess.*	6095	8870	12582	45.5	41.8	0.92
arXiv stat.*	4338	4915	6935	13.3	41.1	3.09
arXiv physics	65233	68838	86501	5.5	25.7	4.64
arXiv math	33617	34616	42553	3.0	22.9	7.72

(b) Preprint servers

Series	2019	2022	2025	Growth 2019→2022 (%)	Growth 2022→2025 (%)	Acceleration factor
medRxiv	1055	12996	18270	1131.8	40.6	0.04
bioRxiv	39862	48588	67211	21.9	38.3	1.75

(c) PubMed

Series	2019	2022	2025	Growth 2019→2022 (%)	Growth 2022→2025 (%)	Acceleration factor
PubMed: AI methods	14585	29692	51754	103.6	74.3	0.72
PubMed: radiology & imaging	108296	80190	94783	-26.0	18.2	-0.7
PubMed: public health	457075	450342	477869	-1.5	6.1	-4.15
PubMed: all biomedicine	1418988	1783569	1881982	25.7	5.5	0.21
PubMed: psychiatry	64821	64847	67978	0.0	4.8	120.37
PubMed: neurology	112364	117147	114559	4.3	-2.2	-0.52
PubMed: surgery	142627	125780	119284	-11.8	-5.2	0.44
PubMed: oncology	155094	166379	156205	7.3	-6.1	-0.84
PubMed: cardiology	100293	103455	88329	3.2	-14.6	-4.64
PubMed: infectious disease	19710	19076	15584	-3.2	-18.3	5.69

(d) CS conferences

Series	2019	2022	2025	Growth 2019→2022 (%)	Growth 2022→2025 (%)	Acceleration factor
ICLR	1591	3391	11565	113.1	241.0	2.13
ACL	1740	3378	8360	94.1	147.5	1.57
ICML	3424	5630	12107	64.4	115.0	1.79
NEURIPS	6743	10411	21575	54.4	107.2	1.97
EMNLP	2876	4190	8174	45.7	95.1	2.08
KDD	1179	1659	2955	40.7	78.1	1.92
CVPR	5165	8161	13008	58.0	59.4	1.02
NAACL	1733	2103	3099	21.4	47.4	2.22
AAAI	7095	9020	12957	27.1	43.6	1.61
SIGIR	426	793	1071	86.2	35.1	0.41
WWW	1247	1822	2062	46.1	13.2	0.29

Note. Each row is one series, taken verbatim from the source table; rows within a panel follow the source order. Panel (a) contains the arXiv category aggregates, panel (b) the two preprint servers, panel (c) PubMed all-biomedicine plus the nine MeSH-defined specialties, and panel (d) the eleven CS conferences used in the analysis. 2019, 2022 and 2025 are calendar-year totals: 2019 is the last complete pre-ChatGPT year, 2022 the last year before the post-ChatGPT period, and 2025 the most recent complete year. Growth is the percentage change between those totals. The acceleration factor is the 2022→2025 growth divided

by the 2019→2022 growth: 1.0 means the trend continued unchanged, above 1.0 means it steepened. Values are reported as stored in the source data and are not recomputed or rounded here.[a]

*[a] 2026 is excluded throughout: PubMed indexing for the current year is incomplete and the arXiv snapshot ends mid-year, so partial-year totals would understate both.

The acceleration is specific to CS rather than a general property of preprint platforms. Within arXiv — a single platform with a single submission process — `cs.*` grew 271% between 2018 and 2025, against 32% for mathematics and 38% for physics. In other words, the CS series is an outlier even against its platform-mates, which rules out platform-level explanations such as arXiv policy or a general shift toward preprinting.

The interrupted time-series models locate the change at the ChatGPT release. For arXiv `cs.*` the fitted slope rises from **53.8 to 163.7 submissions per month** at 2022-11, a change of **+109.9/month (p < 0.00001)**. arXiv totals show a further significant steepening at the agent-era breakpoint (+263.4/month, p < 0.00001). bioRxiv shows **no significant post-ChatGPT slope change** in the same model. *(Table 2.)*

Table 2 Interrupted time-series estimates: pre-break slope, slope change and resulting post-break slope at 2022-11 (ChatGPT) and 2025-01 (agent era), with Newey–West standard errors.

Series	R^2	Pre-break slope	Slope change at 2022-11	Post-break slope	p (2022-11)	Slope change at 2025-01	Post-break slope	p (2025-01)	Last 12 months vs 2019
medRxiv	0.709	-28.22	43.38	15.16	0.0	1.94	17.1	0.67303	11.14
arxiv_econ	0.904	1.47	-0.01	1.46	0.97113	1.16	2.61	0.22667	3.88
arxiv_cs	0.916	53.82	109.9	163.72	0.0	40.05	203.77	0.45723	3.59
arxiv_total	0.924	90.94	148.33	239.26	0.0	263.39	502.65	0.0	2.19
arxiv_eess	0.881	10.12	1.34	11.45	0.3452	-13.09	-1.64	0.00118	2.02
arxiv_stat	0.707	2.51	1.8	4.32	0.01655	5.65	9.96	0.05854	1.92
bioRxiv	0.722	-21.49	71.15	49.66	0.0	-24.06	25.59	0.17973	1.77
arxiv_q-fin	0.664	0.33	0.92	1.25	0.0	0.21	1.46	0.77206	1.75
arxiv_math	0.675	8.21	10.29	18.5	0.00041	84.87	103.37	1e-05	1.62
arxiv_q-bio	0.547	0.69	2.46	3.15	0.0	-5.88	-2.73	0.00066	1.51
arxiv_physics	0.658	14.02	22.64	36.66	0.0	17.17	53.83	0.62014	1.44

Note. Segmented regression on monthly series, with harmonic seasonality and Newey–West standard errors (12 lags). Breakpoints were fixed a priori at 2022-11 (public release of ChatGPT) and 2025-01 (general availability of agentic systems), not selected from the data. Slopes are in submissions per month; a slope change is the difference between the post- and pre-break slopes, and the post-break slope column gives the resulting level. R^2 is the coefficient of determination of the fitted model. The final column is the ratio of the mean of the last twelve months to the 2019 mean. Preprint series begin in 2021 to avoid the COVID-era peak (§2.4). Values are taken verbatim from the source table.[a]

*[a] p-values are reported as stored; values shown as 0.0 are below the output precision of the estimation routine.

3.2 The capacity contrast: expansion versus queueing

CS conferences absorbed the surge by expanding capacity. NeurIPS submissions rose from 10,411 (2022) to 30,709 (2026), **+195%**; accepted papers rose from 2,672 to 7,900, **+196%**; the acceptance rate was **unchanged at 25.7%**. ICLR grew from 3,391 to 19,525 (**+476%**) and ICML from 5,630 to 23,918 (**+325%**). Across the six largest meetings we examined, acceptance rates stayed broadly flat between

2022 and 2026 (Fig. 2). Growth was absorbed by adding programme-committee capacity, not by raising the bar.

Biomedical journal publishing behaved in the opposite way. Because journal capacity is close to fixed, the same underlying increase in submission pressure cannot be absorbed by publishing more. Consistent with this, PubMed output was essentially flat while preprint posting continued to rise: the pressure appears in the submission pipeline rather than in final output.

3.3 Review time conditional on acceptance: no aggregate change

This measure is the submission-to-acceptance interval for papers that were eventually accepted. Its scope is deliberately narrow, and the narrowness is what makes the result interpretable — see the framing below.

On a fixed panel of **38 medical journals** (each contributing ≥500 fully dated records; Table S5), median submission-to-acceptance time was essentially unchanged: 107 days over 2019–2022 (n = 24,911) against 106 days over 2023–2025 (n = 18,270), a change of **−1 day (−0.9%)**. Yearly medians ranged narrowly from 100 to 114 days across 2015–2026 with no trend (Fig. 3 left).

What this measure cannot see. Because `received → accepted` exists only for manuscripts that were accepted, *rejected* manuscripts are absent from the denominator by construction. A journal responding to volume pressure by rejecting more at triage — the expected response when capacity is fixed — will therefore not show any change in this statistic: the affected manuscripts simply disappear from the sample rather than appearing as shorter intervals. The measure is structurally blind to the mechanism through which we would most expect pressure to be absorbed.

The null is where the main hypothesis predicts it. Our central claim is that medicine's capacity is close to fixed, so pressure exits the system as rejection rather than as output (§3.1–3.2). If that is right, the surviving accepted papers need not show longer review times, and the flat panel median is consistent with — rather than contrary to — the main finding. It is a weak test either way, and we do not treat it as evidence against strain.

Corroborating evidence from published editorial reporting. Aggregate statistics reported by publishers and editorial analysts over the same period point to exactly the triage mechanism this measure cannot observe:

- Desk rejections grew **72%** between 2022 and 2025, nearly double the 43% growth in total submission decisions over the same period (Scholarly Kitchen, 2026a).
- The ratio of desk rejections to acceptances rose from **1.69 in 2022 to 2.49 in 2025** — the editorial filter is rejecting faster than submissions are growing (Scholarly Kitchen, 2026a).
- Reviewer invitation acceptance fell every year from 2018 to 2024, from **43% to 22.3%**, with an average of 4.5 invitations now needed per completed review (Scholarly Kitchen, 2026b; Silverchair, 2026).
- Growth was concentrated in small journals: titles receiving fewer than 15 submissions per quarter saw an **81%** increase in early 2026, versus **20%** for those receiving more than 1,500 — a pattern consistent with low-effort submissions targeting the venues least equipped to screen them (Scholarly Kitchen, 2026a).

These are independent, largely qualitative sources rather than a controlled series, so they corroborate the mechanism rather than establish it. Taken together with the volume and capacity evidence, they support reading the flat review-time panel as *pressure exiting through triage* rather than as an absence of pressure.

Journal-level spread, with both extremes flagged. The aggregate stability conceals wide and two-directional journal-level heterogeneity. Among journals with ≥30 fully dated records in both windows, per-journal changes centre on zero (median **+3 days**) but span roughly −50 to +170 days:

- **Longer**: eLife 152 → 323 d (+171), Nature Genetics 284 → 368 d (+84), BMC Medicine 112 → 161 d (+49), European Heart Journal 126 → 158 d (+32), PLoS Biology 178 → 203 d (+26), Thorax 154 → 179 d (+25), Lancet 23 → 47 d (+24), Nature Communications 210 → 232 d (+22)
- **Shorter**: Journal of Hepatology 64 → 20 d (−43), Gastroenterology 40 → 7 d (−33), BMC Cancer 159 → 141 d (−18), BMC Psychiatry 167 → 149 d (−18), Critical Care 66 → 50 d (−16), BMC Infectious Diseases 162 → 146 d (−15)

Two cautions apply to the extremes, and we flag them rather than present the distribution as clean. **eLife** changed its review model in 2023, which alters what "accepted" means for that journal and makes its interval non-comparable across the break. **Gastroenterology** (7 days) and **Journal of Hepatology** (20 days) show intervals short enough to indicate idiosyncratic editorial routing rather than a general speed-up. Neither end should be read as evidence about review speed in general, and no journal-level inference is drawn.

(Fig. 3 left: panel median with IQR band; right: distribution of per-journal changes.)

Position in the argument. This section is supporting evidence. The primary results are the volume divergence (§3.1) and the writing-marker trend (§3.4); the review-time panel is reported to show what a conventional "is review slowing down?" measure does and does not capture.

3.4 AI-associated writing markers rose sharply in both fields

Using the 95-word common detector (§2.3), AI-associated writing markers rose sharply from 2023 in both corpora, and the timing is essentially identical (Table 3):

Table 3 Marker prevalence by corpus and year under the common detector, at density thresholds ≥1, ≥2, ≥3 and ≥5, with mean markers per abstract.

Corpus	Year	n	Proportion with ≥1 marker	Proportion with ≥2 markers	Proportion with ≥3 markers	Proportion with ≥5 markers	Mean markers per abstract
CS	2016	19950	0.386817	0.119298	0.032982	0.001805	0.5482
CS	2017	25982	0.402394	0.127511	0.034293	0.002155	0.5753
CS	2018	30000	0.433467	0.143967	0.042367	0.002933	0.6341
CS	2019	30000	0.4568	0.157767	0.0465	0.002667	0.6764
CS	2020	30000	0.488133	0.1791	0.055733	0.004333	0.7451
CS	2021	30000	0.519133	0.206167	0.0706	0.007167	0.8282
CS	2022	30000	0.555133	0.236267	0.089633	0.012767	0.9379
CS	2023	30000	0.719367	0.4686	0.294967	0.112033	1.9445
CS	2024	30000	0.825467	0.635767	0.469667	0.2186	2.8048
CS	2025	30000	0.851633	0.655233	0.4636	0.179233	2.6503
CS	2026	30000	0.649533	0.356833	0.177667	0.0366	1.3333
Medicine	2018	23963	0.353837	0.112006	0.034345	0.00242	0.5125
Medicine	2019	15986	0.351433	0.104091	0.029401	0.002752	0.4981
Medicine	2020	13987	0.383713	0.125974	0.04111	0.003718	0.5675
Medicine	2021	15990	0.398374	0.134647	0.043777	0.00444	0.5981
Medicine	2022	13996	0.40433	0.142755	0.045656	0.004644	0.6134

Medicine	2023	15964	0.444124	0.184603	0.079053	0.018792	0.7908
Medicine	2024	13972	0.56513	0.323862	0.192957	0.074721	1.4085
Medicine	2025	13995	0.651661	0.417578	0.260164	0.10075	1.7406
Medicine	2026	13990	0.680415	0.407005	0.22752	0.065547	1.5783

Note. All 20 corpus-year cells are shown. Proportions are the share of abstracts containing at least the stated number of distinct markers from the 95-word common detector, and the final column is the mean number of distinct markers per abstract. Values are stored as proportions in the source data and are reproduced here without conversion, recomputation or rounding; multiply by 100 for percentages. Detection uses the common detector defined in §2.3, applied identically to 30,000 arXiv `cs.*` abstracts and 13,000–24,000 PubMed abstracts per year.[a]

*[a] 2026 is included for completeness but excluded from every test and every inference in the text, because coverage of the most recent year is incomplete in both sources (PubMed indexing lag and the arXiv snapshot boundary).

Three features stand out.

A sharp, synchronised break in 2023. In both corpora the 2022→2023 step is an order of magnitude larger than any year-on-year change in the preceding decade. CS mean markers per abstract rise from 0.94 to 1.94; medicine from 0.61 to 0.79 and then to 1.41. The synchrony across two corpora with very different selection pressures argues for a common external cause, though our design cannot exclude concurrent changes in writing norms or editorial practice. The break is not an artefact of aggregation: repeating the analysis within subfields, it appears in **all eight medical specialties** (each sampled directly by MeSH query, $n \approx 1{,}500$ per specialty-year, all $p < 1\times10^{-16}$) and in 23 of 26 arXiv categories with adequate samples, the three exceptions all falling below the $n \geq 500$ per-year floor and so read as underpowered rather than as counterexamples (Tables S3–S4, Fig. S1).

Nearly identical proportional growth. From 2022 to 2025 the CS mean rose ×2.83 and the medicine mean ×2.84. On this measure the two fields' adoption of AI-associated phrasing is indistinguishable in rate.

A persistent absolute gap. Despite identical proportional growth, CS sits at roughly **1.8-fold** the medicine level throughout (2.81 vs 1.41 markers per abstract in 2024; 47.0% vs 19.3% at the ≥3 threshold).

Per-corpus marker identification (§2.3) reproduces this independently: 171 CS and 123 medicine markers, of which 121 and 92 respectively overlap the published Kobak list. The ten strongest medicine markers (*underscores*, *intricate*, *emphasizing*, *fostering*, *necessitating*, *encompassing*, *advancements*, *underscore*, *avenue*, *stands*) all appear in the published list, and the CS list is led by *underscoring*, *underscore*, *underscores*, *overlooking*, *underexplored*, *prompting* — stylistic rather than topical terms, confirming that the style-restricted selection is doing what it is meant to.

Comparability with the published biomedical estimate. Because our method is an adaptation of Kobak et al. (2025) rather than an independent instrument, the medicine-side estimate is a direct check on whether the adaptation behaves as the original. The two are of the same order: our medicine-side prevalence rises from 0.61 markers per abstract in 2022 to 1.41 in 2024, and the share of abstracts at the ≥3 threshold rises from 4.6% to 19.3% over the same window, against the published estimate that a substantial minority of 2024 biomedical abstracts carry LLM signals. We do not claim numerical identity

— the corpora differ (our PubMed sample is stratified and capped, theirs is the full baseline snapshot), the marker vocabularies differ (our medicine list is re-derived and then restricted to the 95-word common detector), and the summary statistic differs (we report a threshold share rather than a modelled prevalence). The agreement is at the level of order of magnitude and of timing, which is what a method-transfer check can establish: the sharp 2023 break appears in our medicine corpus as it does in the published analysis, using an instrument tuned only on the style-restricted vocabulary and never fitted to reproduce their number.

(Fig. 4.)

3.5 Marker prevalence and editorial outcomes

A deliberately unscoped question. We report what the data show and do not fit an association, because journal-level inference here is not identified.

Marker prevalence rises in essentially every depositing journal, while review-time change varies in both directions across the same journals, so no consistent journal-level relationship is apparent in these data.

We report that pattern rather than fitting a model to it. A journal-level association between marker prevalence and review-time change would be jointly determined by journal reputation, editorial policy, field growth and author-language composition — none of which this design observes — so a fitted coefficient would be uninterpretable regardless of its sign or significance. The result that the two quantities move independently across journals is itself informative, and it is the reason the paper's argument rests on the cross-ecosystem comparison (§3.1, §3.4) rather than on within-medicine heterogeneity. Identifying the journal-level question properly requires within-journal submission-level data, which is a different study.

3.6 Summary of results

10. CS preprint volume **doubled** (2022→2025) while biomedical journal output was **flat**, with several specialties contracting. The divergence holds within arXiv itself.
11. CS conferences absorbed the surge by **expanding capacity** and holding acceptance rates flat; journal publishing has no equivalent mechanism.
12. Review time **conditional on acceptance** shows **no aggregate change**, but the measure is structurally blind to desk rejection; published editorial statistics over the same period (desk rejections +72%, rejections per acceptance 1.69 → 2.49, reviewer acceptance 43% → 22.3%) indicate that this is where the pressure exited.
13. AI-associated writing markers rose in **both** fields from 2023, at **nearly identical proportional rates** (≈2.8×), from a higher CS base.

The combination of (4) with (1) is the central finding: **similar upstream change, opposite downstream response.**

4. Discussion

4.1 Same upstream change, opposite downstream response

The central result is a dissociation. The proportional rise in AI-associated writing from 2022 to 2025 was almost exactly the same in medicine and in computer science — ×2.83 and ×2.84 on mean markers per abstract (§3.4). The downstream volume response could hardly be more different: the CS literature

doubled (+99.3%) while biomedical journal output was flat (+5.5%), with several specialties contracting outright (§3.1).

This is not explained by differing exposure to the technology. The tools were equally available to both communities, and the writing evidence says both communities used them to a similar relative degree. What differs is not the *input* but the *permeability of the system*: how readily a change in writing behaviour converts into a change in submitted and published volume.

Three features of the two ecosystems account for the gap.

Cost of doing versus cost of writing. Generative tools reduce the cost of producing text. They do not reduce the cost of producing evidence. A clinical study still needs ethics approval, often prospective registration, data-governance clearance and a recruited sample before a manuscript exists. A CS paper often needs none of these. When writing is the binding constraint, cheaper writing translates directly into more submissions; when doing is the binding constraint, it does not.

Absorptive capacity. Conference review scales: programme committees can be enlarged, and the evidence shows they were — NeurIPS accepted 196% more papers in 2026 than in 2022 with its acceptance rate unchanged at 25.7% (§3.2). Journal capacity does not scale the same way: page budgets, editorial teams and volunteer reviewer pools are close to fixed, and reviewer willingness to accept invitations fell from 43% to 22.3% between 2018 and 2024. Under fixed capacity, additional submissions become additional rejections.

Gatekeeping norms. Medical journals were early and comparatively strict in setting explicit AI-disclosure requirements, and clinical research operates under institutional and regulatory scrutiny without a CS equivalent. These norms may suppress the *volume* effects of cheaper writing even while leaving the *stylistic* effects detectable, which is one reading of why the marker gap (CS at roughly 1.8× medicine) persists without narrowing.

4.2 Why the level gap does not explain the volume gap

It is worth being explicit about what the 1.8-fold marker gap can and cannot account for. CS abstracts carry more AI-associated phrasing, and one might suppose this drives the volume divergence. It does not, for two reasons.

First, the gap is *stable*: it is present in 2020, before any of this mattered, and it persists through 2025 without widening. A constant difference cannot produce a divergence that appears at a specific date.

Second, the marker levels moved *together* after 2023 (×2.83 vs ×2.84), while the volumes moved apart. If writing markers drove volume, the two fields' volumes should have moved similarly, as their markers did.

The divergence therefore has to be attributed to the system's response to a common stimulus, not to different-sized stimuli. That is the substantive claim, and it is the one our data can support.

4.3 Two crises wearing one name

The literature and the trade press commonly describe a single "AI-driven submission crisis." Our results suggest that this framing conflates two problems that require different responses.

In CS, the pressure is a **throughput** problem with an available answer. Volume grew, and venues grew with it; the mechanism that has been reached for — submission limits, as at arXiv in October 2026 — addresses rate rather than quality, and its costs fall on productive authors generally.

In medicine, the pressure is a **triage** problem with no equally available answer. Output cannot expand to match, so the system absorbs the load by rejecting earlier and more often: desk rejections grew 72% between 2022 and 2025 against 43% growth in submission decisions, and the ratio of desk rejections to acceptances rose from 1.69 to 2.49 (§3.3). The load is real, but it is invisible in the metric most often used to assess it — the review time of papers that were accepted — because those papers are precisely the ones that survived the filter.

Treating these as one phenomenon risks applying the wrong remedy: capacity expansion answers the CS problem and does nothing for the medical one, while tightened desk screening addresses the medical problem but would be inappropriate as a general response to CS volume.

One consequence is epistemic rather than procedural. When triage becomes the binding filter, the selection that shapes the literature is made earlier, on less information, and by fewer people — a shift whose effects on what becomes known are difficult to observe after the fact (Messeri & Crockett, 2024). Our data cannot measure that effect; they only identify where in the system it is likely to arise.

4.4 Implications for researchers

The evidence above supports practical guidance that differs by ecosystem.

For authors targeting biomedical journals. The binding constraint has moved toward triage. Budget for a higher probability of desk rejection, and consider whether a manuscript's framing makes its contribution legible in the first page an editor reads. Because the review time of accepted papers has not lengthened (§3.3), the expected cost of submission is increasingly concentrated in the early filter rather than spread across a slow review.

For authors submitting to CS venues. Capacity has expanded and acceptance rates have held roughly constant, so growth in submissions has not translated into a comparably lower acceptance probability. The relevant risk is different: as volume rises, the variance of review outcomes rises with it, and the value of a paper's distinctiveness relative to a larger pool becomes the operative consideration.

On disclosure. Policies have converged on two rules — AI cannot be an author, and substantive use must be disclosed — while disclosure practice lags considerably. The two rules are set out by the International Committee of Medical Journal Editors (2024) and the Committee on Publication Ethics (2023), and their rationale is that authorship carries accountability which cannot be attached to a model (Flanagin et al., 2023; Zielinski et al., 2024). Write the disclosure statement before drafting rather than reconstructing it afterwards, and note that journals increasingly require disclosure of *assistance* separately from *authorship*.

On detectability. The markers measured here are stylistic rather than semantic — words and constructions models overuse — and they are detectable at corpus scale with simple tooling. Heavy reliance on LLM drafting is therefore observable in principle by editors and reviewers. Light-touch editing (structure, clarity, language) retains most of the productivity benefit at a much smaller stylistic footprint; wholesale drafting does not, and increasingly carries reputational risk independent of whether the content is sound.

4.5 Limitations

Measurement. The marker measure identifies stylistic similarity to LLM output rather than confirmed use of any tool. This defines what is being reported and makes the estimate a **conservative lower bound** on

AI-associated writing: the detector keys on stylistic distinctiveness, so as models become less distinctive its sensitivity falls and prevalence is understated rather than overstated.

Design. The study is ecological. We measure populations of documents and platform-level volumes, not individual manuscripts, and we therefore report temporal co-movement and cross-ecosystem difference. We do not attribute any volume change to generative AI causally: funding cycles, open-access mandates, post-pandemic recovery and AI-independent capacity decisions are all plausible contributors that this design cannot separate. The same constraint is why §3.5 reports the journal-level pattern without fitting an association to it.

Routine constraints. Three standard sampling issues apply and do not materially affect interpretation. The marker vocabulary is identified independently within each corpus but originates in part from a biomedical reference list, so residual domain effects may remain despite the common-detector restriction. The 38-journal review panel is publisher-biased, because deposit habits for `received`/`accepted` dates vary by publisher rather than by field. And the PubMed and arXiv abstract sets are systematic rather than random samples, so the reported intervals are approximate. None of these alters the direction or order of magnitude of the results.

4.6 Conclusion

The AI-era submission surge is real, but it is not one phenomenon. In medicine and in computer science the growth in AI-associated writing was proportionally the same; the responses were not. CS doubled its literature and expanded review capacity to match, holding acceptance rates steady. Biomedicine held output flat and absorbed the pressure through rejection, where it is least visible to the measures most often used to look for it.

Knowing which of these a field resembles is the first step to knowing what to do about it — and the two cases call for different remedies.

Statements and Declarations

Funding

The authors received no specific funding for this work.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable. This study used only publicly available scholarly metadata and published text. It involved no human participants, no animal subjects, no personal data and no intervention, so ethical approval and informed consent were not required.

Consent to participate / Consent for publication

Not applicable — no human participants and no individual-level data.

Data availability

The datasets and analysis code generated and analysed during this study are available in the Zenodo repository:

https://doi.org/10.5281/zenodo.23436255

The archive contains the analysis code (`code/`), the derived data underlying every table and figure (`data/`, `results/analysis/`), the generated figures (`results/figures/`), and a file manifest with SHA-256 checksums (`DATA_MANIFEST.csv`).

Three notes on scope:

- The 2.5 GB arXiv metadata snapshot is not deposited because it is a public, versioned dataset; the supplied script `code/p1b_arxiv_snapshot.py` re-downloads it from Hugging Face, and all data derived from it are included.
- **The abstract text itself is not deposited.** The study measures word usage in roughly 400,000 arXiv abstracts and 320,000 PubMed abstracts, but redistributing that text as a bundled corpus is not something we can do: NLM states that its data include abstracts that may be protected by copyright and disclaims responsibility for redistribution, and most arXiv articles carry a licence that does not let arXiv grant redistribution rights to others. The supplied scripts re-fetch the abstracts from the public endpoints instead. All measurements derived from them — document counts, marker prevalence, marker word lists and review dates — contain no abstract text and are included in full.
- All primary sources are public and are listed with their endpoints in `README_DATA.md`: arXiv submission statistics and metadata, the medRxiv and bioRxiv APIs, NCBI E-utilities (PubMed), and cs-conf-stats.

PubMed-derived data in the archive is provided courtesy of the U.S. National Library of Medicine, which has not endorsed this work.

Code availability

The analysis code is included in the same Zenodo record (directory `code/`) and in the GitHub repository above, under an MIT licence. It reproduces the reported results from the supplied data without network access; full re-collection from the public sources is also supported and documented in `README.md`.

Authors' contributions

Kangwei Wang: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing, Project administration.

Ce Shi: Conceptualization, Validation, Writing – review & editing.

All authors read and approved the final manuscript.

Use of large language models

Declared in the manuscript at **Methods §2.9**. In summary: large language models were used to assist with analysis code and with drafting the manuscript text; all code was executed and all outputs verified by the authors; no result was generated by a model; the authors take full responsibility for the final content. The journal's "AI assisted copy editing" exemption was considered and judged not to cover the assistance provided, so the use is declared rather than omitted.

Acknowledgements

[Optional. If none, delete this section or write: "Not applicable." Do not acknowledge anyone whose permission has not been obtained.]

References

arXiv. (2026, June 26). *A year in review: arXiv's 2025 annual report*. https://blog.arxiv.org/2026/06/26/a-year-in-review-arxivs-2025-annual-report/

arXiv. (n.d.). *Monthly submission statistics* [Data set]. https://arxiv.org/stats/get_monthly_submissions

Bao, T., Zhao, Y., Mao, J., & Zhang, C. (2025). Examining linguistic shifts in academic writing before and after the launch of ChatGPT: A study on preprint papers. *Scientometrics, 130*(7), 3597–3627. https://doi.org/10.1007/s11192-025-05341-y

Berens lab. (n.d.). *llm-excess-vocab* [Computer software]. https://github.com/berenslab/llm-excess-vocab

Committee on Publication Ethics. (2023). *Authorship and AI tools: COPE position statement*. https://publicationethics.org/cope-position-statements/ai-author

Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality* (Working Paper 24-013). Harvard Business School. https://www.hbs.edu/faculty/Pages/item.aspx?num=65798

Flanagin, A., Bibbins-Domingo, K., Berkwits, M., & Christiansen, S. L. (2023). Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA, 329*(8), 637. https://doi.org/10.1001/jama.2023.1344

Hugging Face. (n.d.). *arxiv-metadata-snapshot* [Data set]. https://huggingface.co/datasets/librarian-bots/arxiv-metadata-snapshot

International Committee of Medical Journal Editors. (2024). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. https://www.icmje.org/recommendations/

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances, 11*(27), eadt3813. https://doi.org/10.1126/sciadv.adt3813

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). *Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews*. https://arxiv.org/abs/2403.07183

Lopez Bernal, J., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology, 46*(1), 348–355. https://doi.org/10.1093/ije/dyw098

Lopez Bernal, J., Soumerai, S., & Gasparrini, A. (2018). A methodological framework for model selection in interrupted time series studies. *Journal of Clinical Epidemiology, 103*, 82–91. https://doi.org/10.1016/j.jclinepi.2018.05.026

Matsui, K. (2025). Delving into PubMed records: How AI-influenced vocabulary has transformed medical writing since ChatGPT. *Perspectives on Medical Education*. https://doi.org/10.5334/pme.1929

medRxiv & bioRxiv. (n.d.). *Reports and details API* [Data set]. https://api.medrxiv.org/reports/ ; https://api.biorxiv.org/

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

National Center for Biotechnology Information. (n.d.). *E-utilities (Entrez Programming Utilities)*. https://eutils.ncbi.nlm.nih.gov/entrez/eutils/

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708. https://doi.org/10.2307/1913610

NISO. (2026, June). *New report highlights trends and challenges in peer review*. https://www.niso.org/niso-io/2026/06/

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Organization Science. (2026). More versus better: Artificial intelligence, incentives, and the emerging crisis in peer review [Editorial]. *Organization Science, 37*(3). https://doi.org/10.1287/orsc.2026.ed.v37.n3

Paklina, S., Parshakov, P., & Rapoport, E. (2026). LLM-assisted writing and citation advantage: Evidence from scientific publications before and after ChatGPT release. *Scientometrics, 131*(10), 7611–7636. https://doi.org/10.1007/s11192-026-05805-9

Qiu, W., & Pan, F. (2026). Comparing lexical complexity differences between ChatGPT-generated and human-written medical abstracts. *Scientometrics, 131*(8), 5049–5072. https://doi.org/10.1007/s11192-026-05701-2

Scholarly Kitchen. (2026a, May 13). *Is growth always good news? 2026 article submission surges*. https://scholarlykitchen.sspnet.org/2026/05/13/guest-post-is-growth-always-good-news-2026-article-submission-surges/

Scholarly Kitchen. (2026b, September 15). *Peer review triage: Deciding what's worth saving*. https://scholarlykitchen.sspnet.org/2026/09/15/guest-post-peer-review-triage-deciding-whats-worth-saving/

Silverchair. (2026). *Future of peer review report 2026*. https://www.silverchair.com/

The peer-review crisis: How to fix an overloaded system. (2025). *Nature, 644*(8075), 24–27. https://doi.org/10.1038/d41586-025-02457-2

Toner-Rodgers, A. (2024). *Artificial intelligence, scientific discovery, and product innovation* [Working paper]. Massachusetts Institute of Technology.

Xu, X. (n.d.). *CS Conf Stats: Conference acceptance rate and submission statistics* [Data set]. https://csconfstats.xoveexu.com/ ; https://github.com/Xovee/cs-conf-stats

Zielinski, C., Winker, M. A., Aggarwal, R., Ferris, L. E., Heinemann, M., Lapeña, J. F., Pai, S. A., Ing, E., Citrome, L., Alam, M., Voight, M., & Habibzadeh, F. (2024). Chatbots, generative AI, and scholarly

manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Current Medical Research and Opinion, 40*(1), 11–13. https://doi.org/10.1080/03007995.2023.2286102

Figure captions

Fig. 1 Submission volume by ecosystem, indexed to 2021 = 100. Series are annual totals for arXiv `cs.*`, arXiv total, medRxiv, bioRxiv and PubMed (all biomedicine), labelled in the figure as "PubMed (all)" for space. 2021 is the base year because medRxiv launched only in mid-2019, so a 2019 base is undefined for it, and because 2021 precedes ChatGPT for every series. The dashed vertical line marks the public release of ChatGPT (November 2022). By 2025, arXiv `cs.*` reaches an index of 211 while PubMed sits at 105.

Note: the vertical axis is floored at 50 for legibility. Because medRxiv began publishing in June 2019, its 2019 annual total is a partial year and indexes to 6.5; that point is plotted but falls below the axis floor, so the medRxiv series enters the panel from the bottom edge. No series is otherwise truncated.

Fig. 2 CS conference submissions and acceptance rates. (a) Submissions indexed to 2019 = 100, so that meetings of very different size can be compared on one axis; 2019 is the first complete pre-ChatGPT year available for ICLR. (b) Acceptance rate as reported by each meeting. The two panels together show the capacity response: submissions grew without a corresponding fall in acceptance rate. The dashed vertical line marks the public release of ChatGPT (November 2022).

Fig. 3 Review time conditional on acceptance. (a) Median submission-to-acceptance time for a fixed panel of 38 medical journals (each contributing ≥500 fully dated records), with the interquartile range shown as a band; the dashed vertical line marks the ChatGPT release. (b) Distribution of per-journal change between 2019–2022 and 2023–2025, for journals with ≥30 dated records in each window; the dashed vertical line is the median.

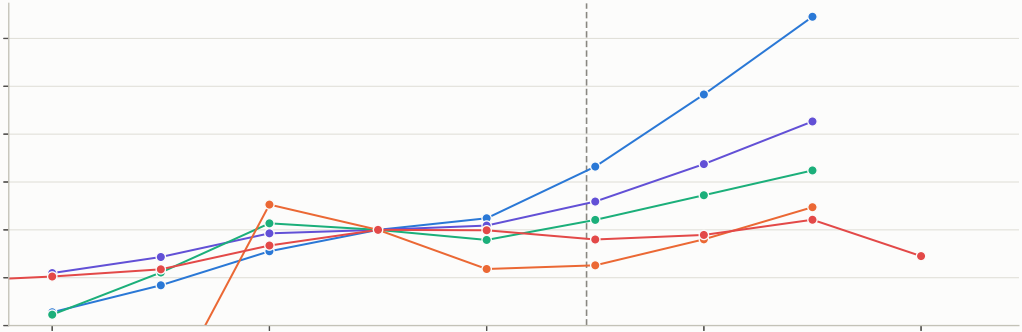
Note: this measure exists only for manuscripts that were accepted. Manuscripts rejected before or during review are absent from the sample by construction, so a journal absorbing pressure through desk rejection will show an unchanged interval for the papers it does accept. The panel result is therefore reported as supporting evidence, and two extremes in panel b are flagged in the text: eLife's 2023 change of review model alters what "accepted" means for that journal, and the very short intervals at Gastroenterology and Journal of Hepatology indicate idiosyncratic editorial routing. No journal-level inference is drawn.

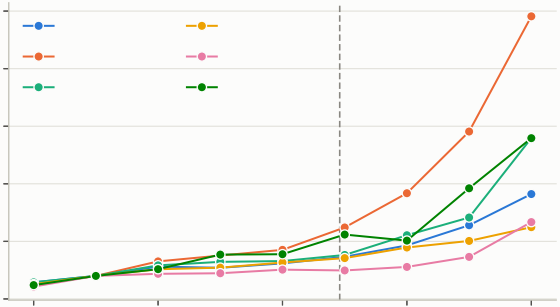
Fig. 4 Prevalence of AI-associated writing markers under the common detector. (a) Share of abstracts containing ≥3 markers. (b) Mean number of distinct markers per abstract. Both panels use the 95-word detector identified as excess in both corpora and are applied identically to 30,000 arXiv `cs.*` abstracts and 13,000–24,000 PubMed abstracts per year. Shading indicates 95% Wilson intervals (narrower than the plotted markers at these sample sizes). The dashed vertical line marks the public release of ChatGPT (November 2022).

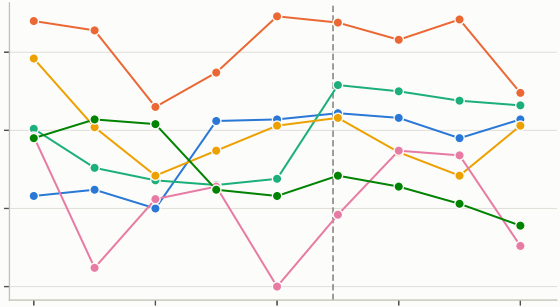
Note: 2026 is shown but should not be interpreted. Coverage of the most recent year is incomplete in both sources — PubMed indexing lag and the arXiv snapshot boundary — and the apparent decline is most likely an artefact of that incompleteness rather than a reversal in writing practice. All inference in the text excludes 2026.

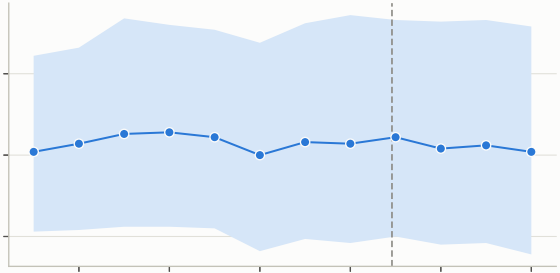
Fig. S1 Subfield robustness of the 2023 break. Small multiples of the share of abstracts containing ≥3 excess-vocabulary markers by year. (a) Eight medical specialties sampled directly by MeSH query (n ≈ 1,500 per specialty-year). (b) The eight best-powered arXiv computer science categories. Panels are ordered by effect size within each block; each panel is titled with the group and the ratio of pooled 2023–24 to 2021–22 prevalence. The dashed vertical line marks the ChatGPT release. Shading is the 95% Wilson interval. Panels are blue where every year clears the n ≥ 500 reliability floor and orange otherwise. The break appears in every panel shown. The three arXiv categories where it is not detected are all below the reliability floor and are listed in Table S4.

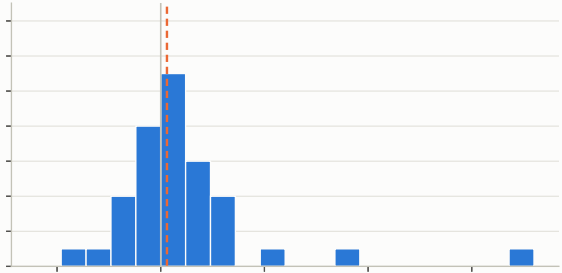
Note: the two-proportion test p-values are omitted from the panel titles for space and are given in Tables S3 and S4. For every panel shown, p < 1e-16 except where noted in those tables. The 2026 value is depressed in every panel; this is the indexing artefact described in the main text, and 2026 is excluded from all tests and from every inference drawn from this figure.

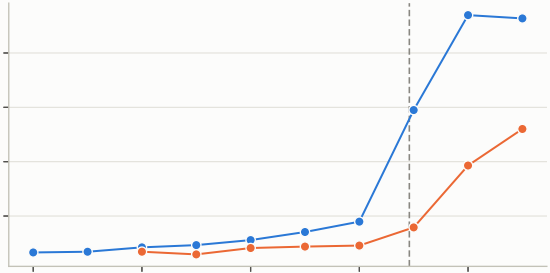


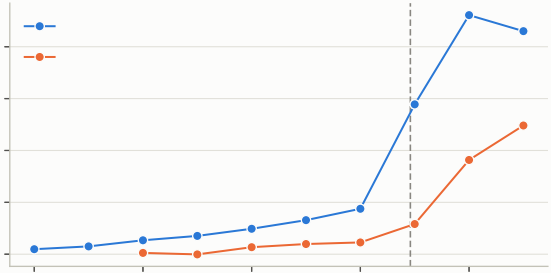


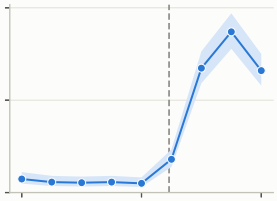
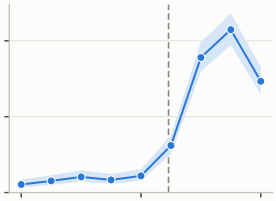
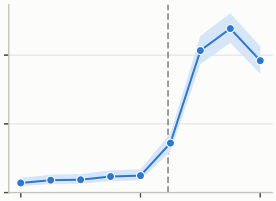
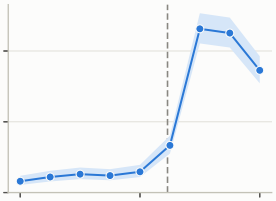
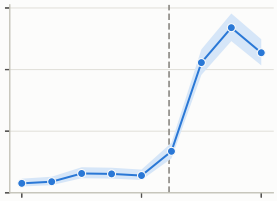
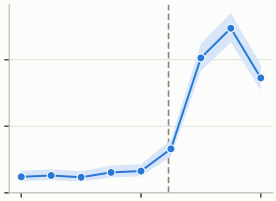
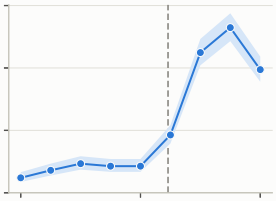
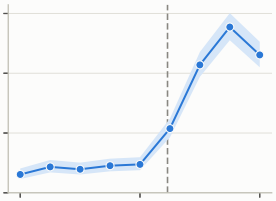
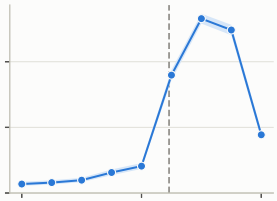
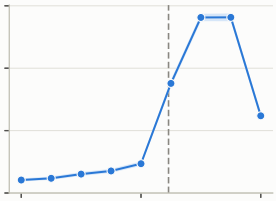
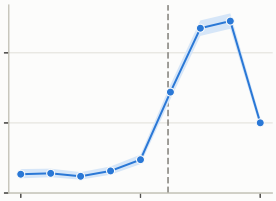
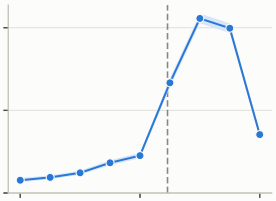
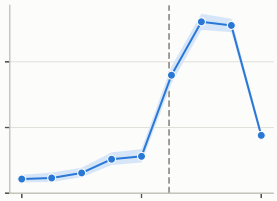
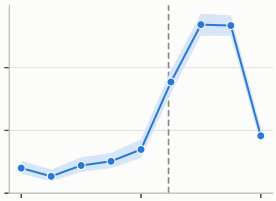
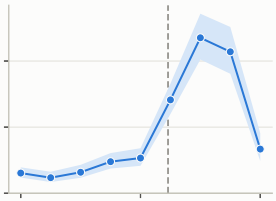
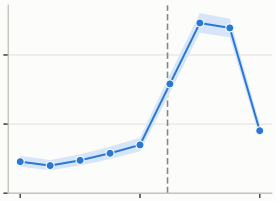










Supplementary Files

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

- SupplementaryTablesS1S5.docx