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.

Artificial Intelligence and Research Integrity in Scholarly Publishing: A Scoping Review of Peer Review, Authorship, and Disclosure Practices

Gabriel Osei Forkuo

gabriel.forkuo@unitbv.ro

Transylvania University of Brașov

Abigail Boatemaa

University of Bradford

Baffour Osei

University of Skills Training and Entrepreneurial Development

Emmanuel Osei-Dwomoh

Controller and Accountant General's Department, Kumasi, Ghana

Research Article

Keywords: AI ethics, large language models, research integrity, scholarly publishing, AI-text detection, disclosure

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11614667/v1

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

Additional Declarations: No competing interests reported.

Artificial intelligence and research integrity in scholarly publishing: a scoping review and ethical analysis

Abstract

Generative artificial intelligence (AI) tools have entered manuscript preparation, peer review, and editorial decision making faster than journals have built policy to govern them, raising ethical questions about fairness, accountability, confidentiality, and honesty. This scoping review maps the published literature on AI and research integrity in scholarly publishing and analyzes its ethical implications. We searched Semantic Scholar, OpenAlex, and CrossRef for records published between 2000 and 2026 and added a supplementary targeted search. We included 77 studies (37 from the database search, 40 from the supplementary search). Empirical or benchmark studies form the largest group (27 studies, 35%), followed by editorials and commentaries (20, 26%). Detection tools show inconsistent accuracy and documented bias against non-native English writing. Authorship guidance converges on excluding AI systems as authors while permitting disclosed assistive use. Guidance for AI used as a writing tool is less developed than guidance for studies that evaluate AI systems. Building on this evidence, we argue that detector output alone is an unfair basis for sanctioning authors, that responsibility for AI-assisted text cannot be delegated to a tool, that reviewers' confidentiality duties require institutional support, and that disclosure requirements should be proportionate and paired with fair verification procedures. The evidence base rests largely on opinion and small studies, and the effectiveness of verification mechanisms remains untested.

Keywords: AI ethics, large language models, research integrity, scholarly publishing, AI-text detection, disclosure

1. Introduction

Generative artificial intelligence (AI) tools, in particular large language models (LLMs), became widely accessible to researchers in late 2022. Within two years, questions about their use in writing, reviewing, and editing scholarly work moved from a niche concern to a standing item on journal editorial agendas. Publishers responded with policy statements. Editorial associations issued guidance. Individual journals wrote disclosure rules into their author instructions. The literature documenting all of this grew just as fast, but it grew unevenly, scattered across editorials, empirical studies, guideline papers, and commentary pieces that rarely cite each other systematically.

This unevenness creates a practical problem for editors, publishers, and research integrity officers who need to set policy now, not once the evidence base matures. A journal deciding whether to require AI-use disclosure has to draw on a literature that has not yet been mapped as a whole. The same is true for a journal deciding whether to rely on AI-text detectors during screening. Individual studies report contradictory findings on detector accuracy [1–4]. Editorial opinion pieces disagree on whether an AI system can meet authorship criteria at all [5–7]. Reporting guidance has proliferated for AI-driven prediction models [8–10] but has not caught up with AI used simply to draft prose.

Research integrity, in the sense this review uses it, covers the practices that keep the scholarly record trustworthy. These practices include honest authorship attribution, accurate reporting of methods and results, transparent disclosure of tools and assistance used, and functioning peer review. Artificial intelligence intersects with all four. A researcher who uses a chatbot to draft a discussion section raises an authorship and disclosure question. A reviewer who pastes a manuscript into a chatbot to generate a review raises a confidentiality and accountability question [11, 12]. A paper mill that uses generative models to produce fabricated results at scale raises a fraud-detection question that may outpace manual screening capacity [13–15].

The stakes of getting this governance right are ethical as well as practical. Peer review and editorial screening function as a filter that the rest of science, and the public reporting built on it, depends on. That filter can be weakened by convincingly fabricated text, by detection systems that misidentify honest authors, or by reviewers who delegate judgment to a tool that cannot be held accountable. When it is weakened, the cost does not fall evenly. It falls on readers who rely on the published record without re-

verifying it, on authors who are wrongly accused, and on researchers whose unpublished work is exposed to third-party services without their consent. The corpus reviewed here includes a proof-of-concept demonstration that generative AI can produce a fraudulent article that looks authentic [16]. It also includes evidence that organized fraud networks operate at scale and studies arguing that generative tools may accelerate their output [13, 14].

This review has two purposes. The first is descriptive: to map the literature systematically. The second is normative: to ask what that literature implies about what authors, reviewers, editors, publishers, and institutions owe one another, and who bears the costs of the current governance gaps. We ask four mapping questions and one ethical question.

First, what does the published literature say about AI use in peer review, and how well do the identified practices align across journals and disciplines? This question matters because peer review is the point in scholarly publishing where a manuscript's confidentiality and a reviewer's independent judgment are both, in principle, protected simultaneously. AI use threatens each in a different way.

Second, what stance has emerged on AI systems as authors or contributors, and where does disagreement remain? Authorship is not only a credit-allocation question. It is also an accountability question, because journals rely on named authors to stand behind a paper's claims, defend it against later challenge, and correct it if it proves wrong.

Third, how reliable are the tools proposed to detect undisclosed AI use, and what tradeoffs do they carry? This question has immediate operational stakes. Journals that consider deploying detection tools in editorial workflows face vendor claims about accuracy, and this review can check those claims against the independent evidence.

Fourth, how is AI itself being used to detect research misconduct and fraudulent publication, including paper mill activity? This question closes the loop the other three questions open. The same technology raising integrity risks in Sections 4.2 through 4.4 also offers new tools for defending against those risks.

Fifth, what ethical obligations follow from this evidence, who bears the costs when governance fails, and which policy positions can be defended on ethical grounds? This question is addressed in Section 5.4, where we state our own normative positions and distinguish them from the findings of the included studies. We treat the ethical argument as a contribution of this paper in its own right, not as an addendum to the evidence map.

We did not restrict inclusion to a single study design. A field this young is documented mostly through editorials, position statements, and small empirical studies rather than large trials. A design-restrictive search would exclude most of what has actually been written.

A scoping review method fits this aim better than a systematic review of effect sizes. Scoping reviews map the extent, range, and nature of research activity on a topic without requiring a single comparable outcome measure across included studies [17]. This literature emerged recently. Much of it takes the form of commentary and guideline development rather than controlled studies. A scoping approach lets us characterize the field's actual shape. It does not force that shape into a meta-analytic structure the evidence cannot yet support.

The remainder of this manuscript proceeds as follows. Section 2 develops a conceptual framework linking AI use contexts to governance mechanisms and outcomes. It also reviews the theoretical basis for treating disclosure, authorship, and detection as related but distinct problems, and it sets out the ethical framework used in the analysis. Section 3 reports our search, screening, and corpus construction methods, including a supplementary search stage that we added after the initial database search under-captured relevant literature. Section 4 presents results organized by theme, with the quantitative characteristics of the included corpus reported first. Section 5 discusses what the evidence supports and where it remains thin (Sections 5.1 to 5.3), develops the ethical analysis (Section 5.4), and then turns to strengths, limitations, and implications for editorial policy (Sections 5.5 to 5.9). Section 6 concludes.

2. Literature Review

2.1 Theoretical framework

Research integrity scholarship traditionally organizes threats to the scholarly record into three categories: fabrication, falsification, and plagiarism, with authorship disputes and undisclosed conflicts of interest treated as adjacent concerns by the International Committee of Medical Journal Editors (ICMJE) [18]. Generative AI does not fit neatly into this taxonomy because it can contribute to all three categories at once. A fabricated dataset produced with AI assistance is still fabrication. Text generated without disclosure raises a plagiarism-adjacent question even when no human author's words were copied, because the reader is misled about who or what produced the content [19]. Authorship frameworks built around the ICMJE criteria, which require the ability to take accountability for a work's content, exclude AI systems by definition. An AI system cannot be held accountable in the way a human co-author can [5, 7, 18].

This review treats three constructs as analytically separate even though they overlap in practice: assistive use, generative use, and disclosure. Assistive use covers tasks like grammar correction, translation, and literature summarization, where the AI system supports a human author's own reasoning and claims. Generative use covers tasks where the AI system produces substantive content that the human author or reviewer then adopts with limited independent verification. An example is drafting a discussion section or generating a peer review report. Disclosure is the governance mechanism that is supposed to make the assistive-generative distinction visible to editors and readers. Several studies in our corpus argue that disclosure alone, without verification, is not sufficient to protect the boundary between these two use types [19–21].

The Committee on Publication Ethics (COPE) addressed this boundary directly in its position statement on authorship and AI tools. Its stance is consistent with the ICMJE position described above: AI tools cannot be credited as authors, and their use should be disclosed in the methods or a similar section of the manuscript [22]. Publishers operationalize the disclosure requirement differently. The author guidelines of the journal AI and Ethics, for example, ask authors to document LLM use in the Methods section but exempt "AI-assisted copy editing" (improvements to human-written text for readability, grammar, and style) from declaration [23]. The pairing of no AI authorship credit with disclosure of substantive AI involvement is close to a cross-organizational consensus position in this literature, although the scope of what must be disclosed is not uniform. The pairing recurs across the disclosure and authorship studies reviewed in Sections 4.3 and 4.4.

2.2 Conceptual framework

Fig. 1 sets out the conceptual framework this review uses to organize its findings. AI use enters the scholarly publishing pipeline through three contexts: manuscript preparation, peer review, and editorial workflow. Within each context, use falls somewhere on a spectrum from assistive to generative. Four governance mechanisms attempt to manage where on that spectrum a given use falls and whether it was disclosed. These are disclosure policies, AI-text detection tools, authorship and accountability norms, and editorial or reviewer guidelines. These mechanisms feed into two outcome domains. The first is research integrity outcomes, meaning the prevalence of undisclosed AI use and integrity violations. The second is trust and quality outcomes, meaning review quality, reproducibility, and the trust that authors, reviewers, and readers place in the published record. A feedback loop runs from outcomes back into policy, because evidence of prevalence and violation rates is what motivates journals and publishers to revise their guidelines.

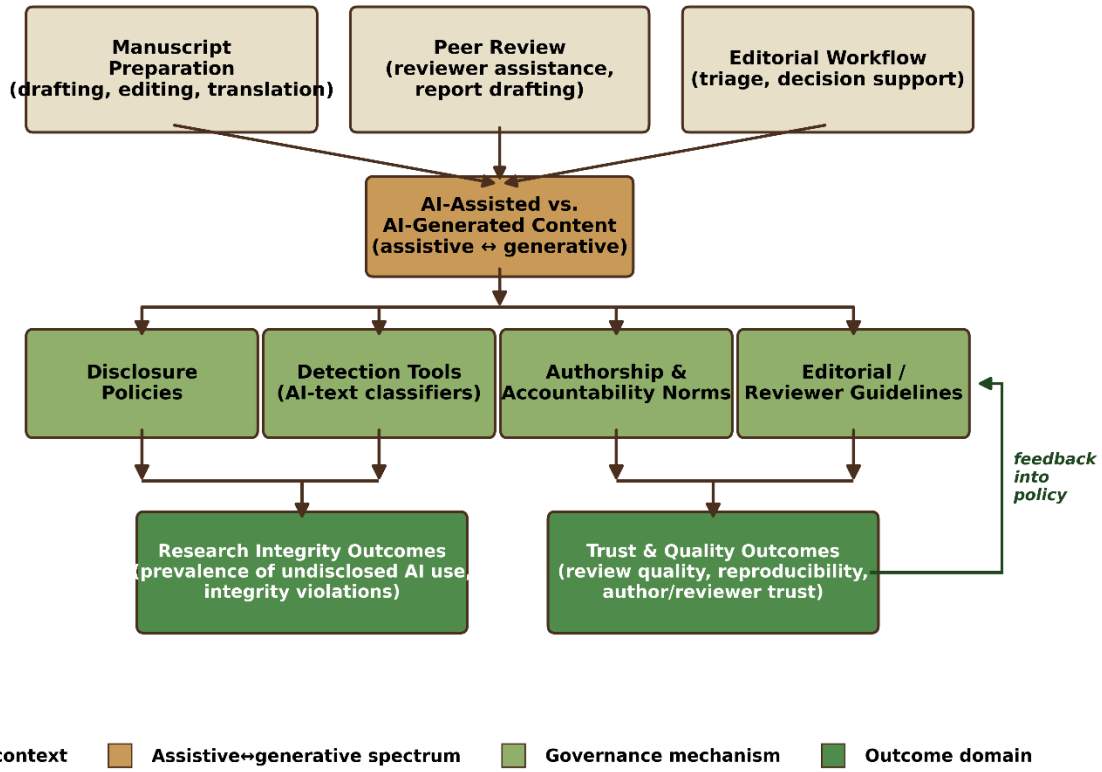


Fig. 1 Conceptual framework linking AI use contexts, governance mechanisms, and outcome domains

This framework does two things for the review that a simple list of themes would not. First, it separates the use context, meaning where in the pipeline AI appears, from the governance mechanism, meaning how that use is managed. That separation lets us ask whether a mechanism proven effective in one context transfers to another. A mechanism proven effective for manuscript-preparation disclosure might not transfer to peer review confidentiality. Second, it treats detection and disclosure as complements rather than substitutes. A detection tool can flag likely AI-generated text, but it cannot establish whether that use was assistive or generative, and it cannot confirm intent. Disclosure, when honest, answers exactly the question detection cannot.

Section 2.1 introduced three constructs this review treats as analytically distinct: assistive use, generative use, and disclosure. Table 1 defines each of these terms, plus two related constructs, as used throughout this manuscript. This keeps the distinction consistent across sections written around different parts of the corpus.

Table 1 Key constructs used throughout this review

Construct	Definition as used in this review	Governance mechanism most directly relevant
Assistive use	AI supports a human author's or reviewer's own reasoning and claims (grammar, translation, summarization) Di	sclosure policy; editorial guidance
Generative use	AI produces substantive content adopted with limited independent verification (drafted sections, generated review reports)	Authorship and accountability norms; disclosure policy
Disclosure	The act of reporting AI involvement to an editor, reader, or both	All four governance mechanisms in Fig. 1

Construct	Definition as used in this review	Governance mechanism most directly relevant
		depend on disclosure being honest
Detection	Algorithmic or human identification of likely AI-generated text without relying on author disclosure	AI-text detection tools
Verification	A process that checks a disclosure statement's accuracy rather than assuming it N	ot yet standardized in any guideline reviewed in this corpus

The verification row in Table 1 is, on the evidence reviewed in Sections 4.4 and 5.3, the least developed governance mechanism in the entire framework. Every other row has at least one guideline, policy statement, or empirical tool associated with it in the corpus. Verification does not, which is itself a gap this review's synthesis identifies rather than one stated directly by any single included study.

2.3 Peer review as a distinct governance problem

Peer review deserves separate theoretical treatment because it combines two integrity risks that manuscript preparation does not: confidentiality and delegated judgment. A reviewer who pastes a submitted manuscript into a commercial chatbot to generate review comments may breach the confidentiality obligation that underlies peer review. The manuscript's content is then transmitted to a third-party service outside the journal's control [11, 12, 24]. Delegated judgment is a separate problem. Even where confidentiality is preserved, a reviewer who substitutes an AI system's assessment for their own expert judgment undermines the accountability structure that gives peer review its epistemic authority [25, 26]. The World Association of Medical Editors (WAME) addressed both concerns directly. It recommended against uploading manuscripts to AI tools that retain or reuse submitted content, while allowing limited assistive use with disclosure to the editor [24].

2.4 Reporting guidelines as a governance response

A cluster of reporting guidelines has emerged specifically for studies that develop or evaluate AI systems, extending the Consolidated Standards of Reporting Trials (CONSORT) and Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) traditions in clinical research methodology to AI-specific reporting items. The CONSORT-AI and SPIRIT-AI extensions cover clinical trials of AI interventions [27, 28]. TRIPOD-AI (now published as TRIPOD+AI), based on the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD), and PROBAST-AI, based on the Prediction model Risk Of Bias ASsessment Tool (PROBAST), cover diagnostic and prognostic prediction models [8, 9]. The most recent addition, TRIPOD-LLM, extends this tradition to studies that develop or evaluate large language models specifically, with 19 main reporting items developed through Delphi consensus [10]. The ChatGPT, Generative Artificial Intelligence, and Natural Large Language Models for Accountable Reporting and Use (CANGARU) initiative takes a broader scope, aiming to produce reporting guidance for generative AI and LLM use across the research process rather than for a single study type [29]. None of these guidelines addresses a narrower and arguably more common scenario. That scenario is an author using a general-purpose chatbot to draft or edit prose, with no AI system under evaluation as the research object itself. That gap is a finding of this review as much as a limitation of the guidelines, and we return to it in Section 5.

2.5 Ethical framework

The literature this review maps is largely written in the language of policy and integrity. Beneath that language sit normative questions about what is fair, what is harmful, and who is answerable. To make these questions explicit, we use two sets of resources. The first is a widely cited synthesis of AI ethics principles: beneficence, non-maleficence, autonomy, justice, and explicability, the last understood to include accountability [30]. These principles adapt the four-principle approach of biomedical ethics [31], and surveys of AI ethics guidelines find convergence on transparency, justice and fairness, non-maleficence, responsibility, and privacy [32]. We use the principles as evaluative criteria, not as a complete moral theory. The second resource is a small set of concepts that bear on specific problems in the corpus. The notion of a responsibility gap, in which the outputs of a learning system cannot be fairly attributed to any human agent [33], helps to frame the authorship question. The notion of testimonial injustice, in which a speaker receives less credibility because of prejudice about their identity [34], helps to frame the consequences of detector bias against non-native English writing. We do not claim that these concepts settle the questions. We use them so that the reasoning behind our positions can be inspected and challenged.

We keep evidence and argument separate. Section 4 reports what the included studies say. Section 5.4 states what we argue on the basis of that evidence. For each position we ask four questions: which ethical issue is at stake and which principle it engages; what evidence in the corpus bears on it; who bears the costs and benefits of the current arrangement; and what position we defend, and what would change our view. The positions in Section 5.4 are ours and are not findings of the included studies.

3. Methods

3.1 Study Design and Protocol Registration

We conducted a scoping review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [17]. We chose a scoping design over a systematic review of a single outcome for two reasons. The literature on AI and research integrity spans editorials, position statements, guideline papers, and empirical studies of varied design. No single comparable effect measure exists across this literature. The review protocol was registered with the Open Science Framework prior to full-text screening. The registration is available at https://osf.io/47nec. Deviations from the registered protocol are reported in Section 3.7.

3.2 Search Strategy

To capture the multidisciplinary literature on AI in scholarly publishing, we employed a structured, two-stage search strategy following PRISMA-ScR principles.

The initial search was conducted across three major academic databases: Semantic Scholar, OpenAlex, and CrossRef. We deployed twelve structured search queries combining terminology related to artificial intelligence and large language models (LLMs) with core publishing domains—including peer review, authorship, academic integrity, research misconduct, manuscript preparation, and editorial workflows. The exact query strings were:

"responsible use large language models scholarly writing"
"AI-assisted literature review evidence synthesis"
"authorship attribution accountability AI-generated text research"
"assistive versus generative AI research workflow"
"artificial intelligence peer review ethics"
"large language models editorial decision-making academic publishing"
"detection AI-generated text research integrity"
"disclosure AI use manuscripts journals policy"
"ChatGPT authorship research misconduct"
"AI text detector accuracy academic writing"
"generative AI prevalence survey researchers"
"LLM-assisted peer review reviewers editors"

Each query was executed across all three databases for records published between 2000 and 2026. The 2000 baseline was chosen so that earlier literature on automated text detection and early AI applications in peer

review could be retrieved, while also covering the acceleration of generative AI research after late 2022. No record published before 2019 was retained in the final corpus (Section 4.1). Deduplication across the three databases yielded 250 unique records. Relevance screening of these records (Section 3.3) yielded an initial corpus of 37 primary database studies.

Because the database queries under-captured relevant literature (Section 3.7), they were supplemented with targeted searches intended to improve coverage of specialized sub-domains. These focused particularly on empirical peer-review evaluations, researcher attitude surveys, paper-mill detection frameworks, and recent reporting standards. Seven targeted query strings were executed:

"AI generative artificial intelligence peer review ethics 2024 2025"
"ChatGPT authorship academic writing ethics scholarly publishing 2023 2024"
"generative AI disclosure policy journals authors manuscripts research misconduct fabrication"
"AI text detector accuracy academic manuscripts arms race detection tools study"
"paper mills AI-generated fraudulent papers retraction scientific literature 2024 2025"
"reporting guideline AI use research CANGARU TRIPOD-LLM ICMJE update 2024 2025"
"survey researchers use of AI writing manuscripts prevalence attitudes 2024"

All candidate records were verified individually. Each record was cross-referenced against original publisher platforms to confirm title, author list, publication year, journal venue, and DOI. This supplementary phase contributed 40 verified studies, resulting in a final synthesis corpus of 77 studies (N = 77). The supplementary stage was not a systematic multi-database search. Its records were identified by targeted queries, so it depends more on the authors' judgment than the database stage does (Section 5.6).

All 77 included studies possess verified primary bibliographic metadata. For studies retrieved through the initial database stage (n = 37), extended programmatic bibliometric metadata, such as automated citation counts, concept tags, and open-access designations, were extracted via the OpenAlex and CrossRef APIs. Consequently, analyses and visualizations that draw specifically on these extended bibliometric fields utilize this 37-study subset, as explicitly noted in the corresponding tables and figures. **Fig. 2** shows the two search stages and the number of records excluded at each stage.

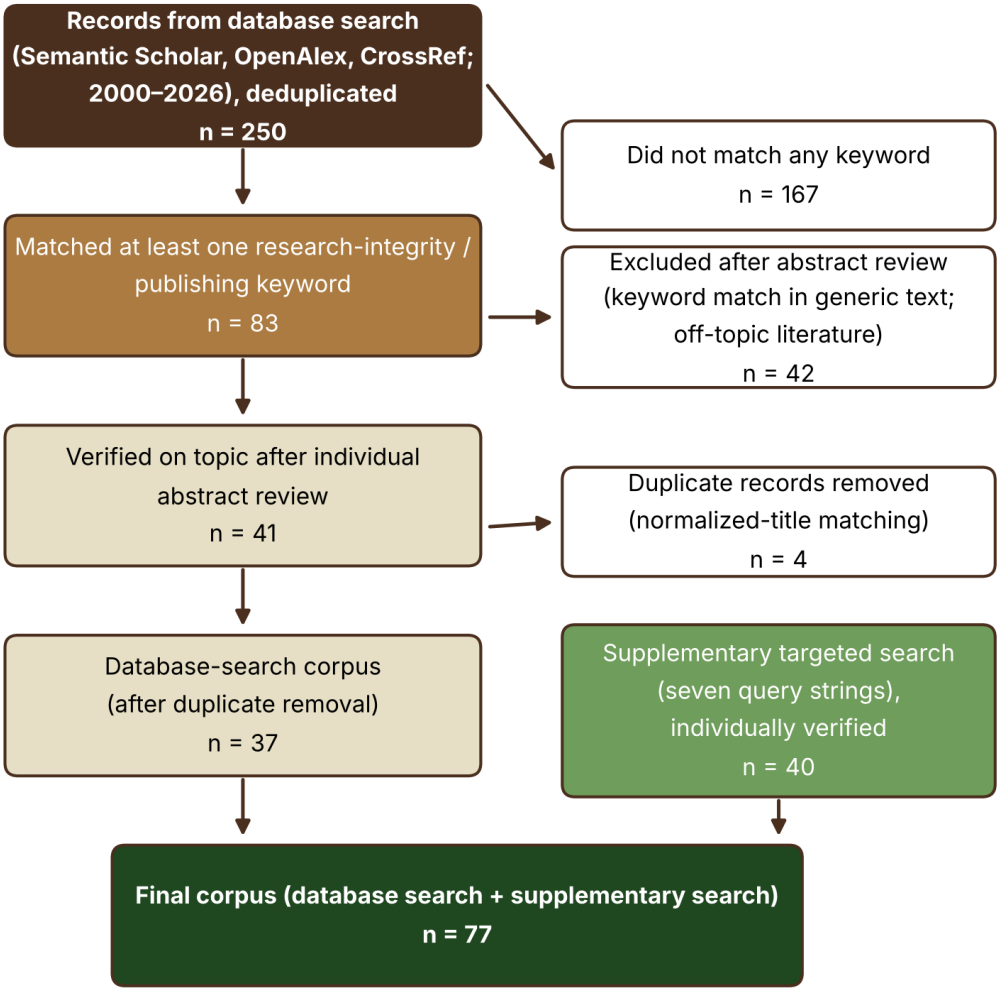


Fig. 2 Corpus selection flow: database search with relevance screening, and supplementary targeted search

3.3 Eligibility Criteria and Relevance Screening

Records were eligible if they (i) were published between 2000 and 2026; (ii) concerned AI, including LLMs, in relation to research integrity or the governance of scholarly publishing, including peer review, authorship, manuscript preparation, editorial workflow, disclosure, detection of AI-generated text, research misconduct, and reporting standards; and (iii) took any study design, including editorials, commentaries, and guideline documents. Records were ineligible if the connection between AI and research integrity appeared only in generic or boilerplate text. Search queries were written in English.

An initial approach to corpus construction stratified the 250-record pool by study design without a topical relevance check. This let studies unrelated to the review's subject into the corpus. Those studies shared generic search-vocabulary overlap with the true queries without addressing AI and research integrity substantively. We discarded this approach. We replaced it with a relevance screen applied to every record in the pool.

The relevance screen worked in three steps. First, we checked each record's title, abstract, and keywords against a keyword pattern. The pattern covered AI use in peer review, authorship, manuscript preparation, editorial workflow, disclosure, and detection of AI-generated text. It also covered research misconduct and reporting-guideline vocabulary associated with this literature (Animal Research: Reporting of In Vivo

Experiments [ARRIVE], SPIRIT-AI, CONSORT-AI, TRIPOD, PROBAST, COPE, ICMJE). This step returned 83 candidate records. Second, we read the abstract of every candidate record individually. We excluded records where the keyword match occurred in generic or boilerplate text rather than in a passage substantively about research integrity. This step excluded 42 records. Most of these were reviews of AI applications in unrelated domains, such as healthcare delivery, drug discovery, or education, whose abstracts happened to contain a matching term. Third, we removed 4 duplicate records identified through normalized-title matching. This left 37 studies from the database search. Four retained records are not primarily about AI. Two concern peer-review transparency and publishing-metric incentives [35, 36], and two are reporting or authorship guidelines that are not AI-specific [37, 38]. We retained them because they bear directly on the integrity mechanisms that AI use puts under pressure, and we identify them in the text where they are used (Sections 4.2, 4.5, and 4.8). We treat their retention as a limitation of keyword-based screening (Section 5.6).

3.4 Study Selection and Data Extraction

Two reviewers screened all records against the relevance criteria described in Section 3.3. We extracted title, authors, publication year, journal or venue, DOI, study design, and abstract for every included study. We classified study design into seven categories: narrative review, empirical or benchmark study, systematic or scoping review, and editorial or commentary. The remaining three categories were user survey or mixed methods, guideline or consensus statement, and theoretical or modeling work. Two included studies are themselves reporting-guideline development papers rather than primary research. We classified these as guideline or consensus statement. We analyzed them alongside primary studies because their content bears directly on the review's questions about governance mechanisms.

We did not extract quantitative effect sizes for statistical synthesis. Most included studies are not structured to report one. We attempted full-text mining for effect sizes, sample sizes, and disclosure statement content across the database-search subset. No manuscript in that subset had machine-readable full text available for extraction. Study design and theme assignment therefore relied principally on titles, abstracts, and bibliographic metadata (Section 5.6). Where a study reports a quantitative result relevant to our four review questions, such as detector accuracy or disclosure rate, we report that result narratively in Section 4. We cite its source rather than pooling it statistically with results from studies that used different outcome definitions.

3.5 Quality and Transparency Assessment

We assessed the 37-study verified subset against five publication-quality and transparency domains. These were indexing across multiple bibliographic databases, open-access availability, presence of a substantive abstract, citation impact relative to the corpus median, and study-design tier within an evidence hierarchy. We built this hierarchy to describe corpus composition, not to weight individual study findings during synthesis. We assigned systematic or scoping reviews, empirical or benchmark studies, and user surveys or mixed-methods studies to a higher tier. We assigned guideline or consensus statements and theoretical or modeling work to a middle tier. We assigned narrative reviews and editorial or commentary pieces to a lower tier. This assessment, modeled on the tiered quality-appraisal approach of the Mixed Methods Appraisal Tool [39], is reported in full in Section 4.9. It is a descriptive characterization of the evidence base, not a risk-of-bias appraisal of individual studies (Section 5.6).

3.6 Synthesis Approach

We synthesized findings thematically rather than statistically. After data extraction, we grouped studies into themes based principally on abstract content. Four themes correspond to our four mapping questions: AI use in peer review, AI authorship and accountability, disclosure and undisclosed-use detection, and AI-assisted detection of research fraud. Two further themes emerged during coding: reporting guidelines for AI-related research (Section 4.5), which cuts across the first four, and AI use in manuscript preparation and scientific writing (Section 4.7), which covers studies not primarily concerned with peer review, authorship, or disclosure. One study on publishing incentives fitted none of these themes and is reported separately

(Section 4.8). A study could contribute to more than one theme. We report the number of studies contributing to each theme. We summarize their findings narratively, using short thematic tables where a tabular format aids comparison across studies. The ethical analysis in Section 5.4 is a separate, argument-based component whose approach is described in Section 2.5.

3.7 Deviations from the Registered Protocol

We made three changes relative to our original plan. First, we added a supplementary targeted search after the initial database search under-captured relevant literature (Section 3.2). Second, we replaced an initial corpus-construction approach that stratified records by study design without a topical relevance check with a relevance screen applied to every record (Section 3.3). Third, in response to editorial review, we added an explicit ethical analysis (Sections 2.5 and 5.4). We did not extract quantitative effect sizes (Section 3.4).

3.8 Use of AI Tools in This Study

During the preparation of this work, the authors used Cursor version 2.4.37 to write and edit the Python scripts used for automatic database search and data analysis. Grammarly for windows (version 1.2.276.1919) was also used to correct all grammatical errors. After using these tools, the authors reviewed and edited the content as needed.

4. Results

4.1 Corpus Characteristics

The final corpus contains 77 studies published between 2019 and 2026. Fig. 3 plots publication volume by year with a fitted linear trend line, while Table 2 reports the counts of publication by study design. Empirical or benchmark studies make up the largest share (27 studies, 35%). Editorial or commentary pieces follow (20, 26%). Narrative reviews account for 13 studies (17%), and guideline or consensus statements account for 11 (14%). User survey or mixed-methods studies number 4 (5%). Theoretical or modeling work and systematic or scoping review each contribute a single study (1% each). Although the search window opened in 2000, the earliest included study was published in 2019, and 34 of the 77 studies (44%) were published in 2023. Annual counts reflect the search design, including year terms in the supplementary queries and the partial coverage of 2026, and should not be read as a measure of publication growth.

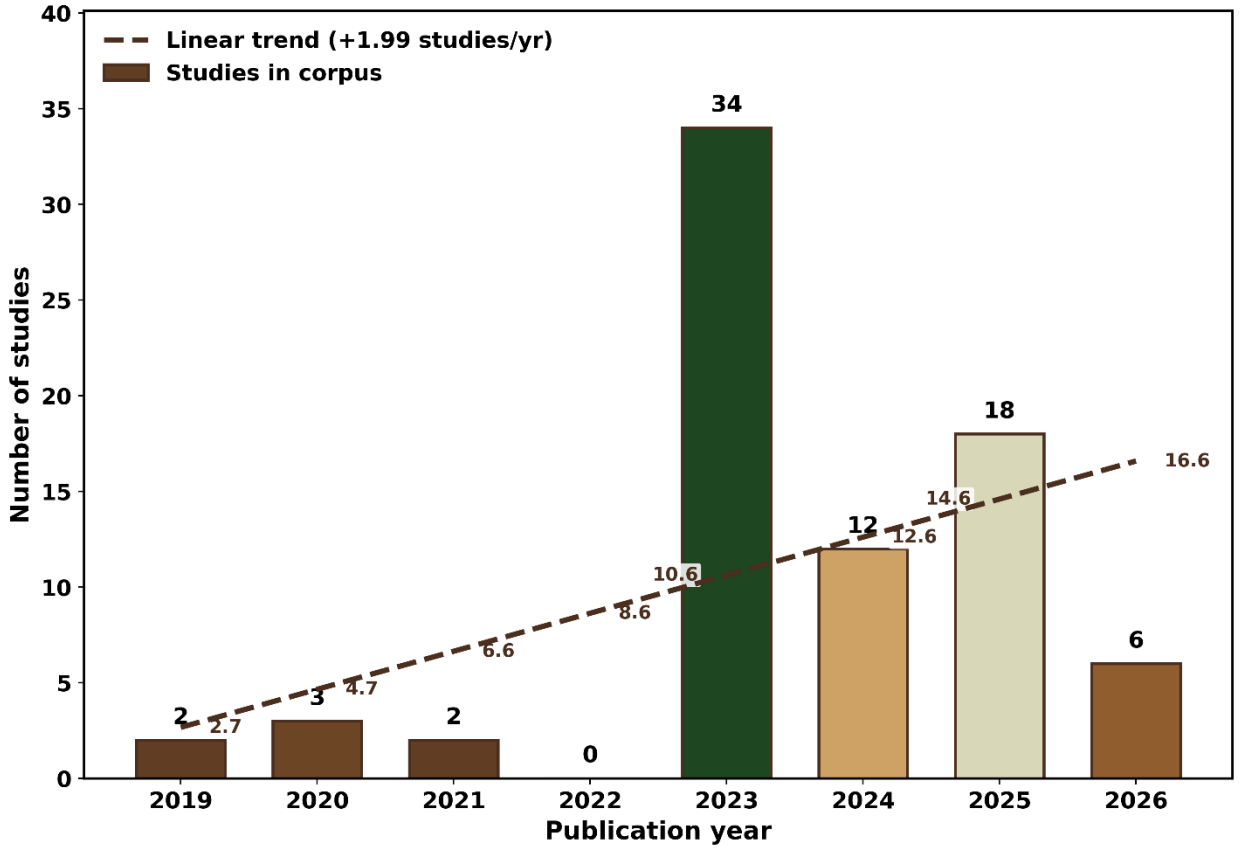


Fig. 3 Publication volume by year with fitted linear trend line, full corpus (n = 77); 2026 is a partial year

Table 2 Corpus composition by study design

Study design	n	% of corpus
Empirical / Benchmark Study	27	35%
Editorial / Commentary	20	26%
Narrative Review	13	17%
Guideline / Consensus Statement	11	14%
User Survey / Mixed Methods	4	5%
Theoretical / Modeling	1	1%
Systematic / Scoping Review	1	1%
Total	**77**	**100%**

The near-absence of systematic or scoping reviews is itself a finding. Only one study in the corpus is a systematic review [13], and it addresses paper mill detection specifically, not AI and research integrity as a whole. We identified no systematic or scoping synthesis in the corpus that covers AI use in peer review, authorship, and disclosure policy as a single integrated topic. This is a statement about our corpus, not a claim that no such synthesis exists elsewhere (Section 5.6). The corpus construction funnel, detailed in Section 3, is summarized quantitatively in Table 3.

Table 3 Corpus construction funnel

Stage	n
Records from database search, deduplicated (2000–2026)	250
Matched ≥ 1 research-integrity/publishing keyword	83
Verified on-topic after individual abstract review	41
Deduplicated database-search corpus	37
Added through supplementary targeted search	40
Final corpus	**77**

Publication volume rose sharply after 2022. This is consistent with the large language model availability timeline described in Section 1. For the 37-study verified subset, median citation count is 282. Maximum citation count is 5,802. A small number of early guideline papers and commentary pieces have accumulated disproportionate citation counts relative to the median study. Fig. 4 cross-tabulates study design against publication year as a heatmap.

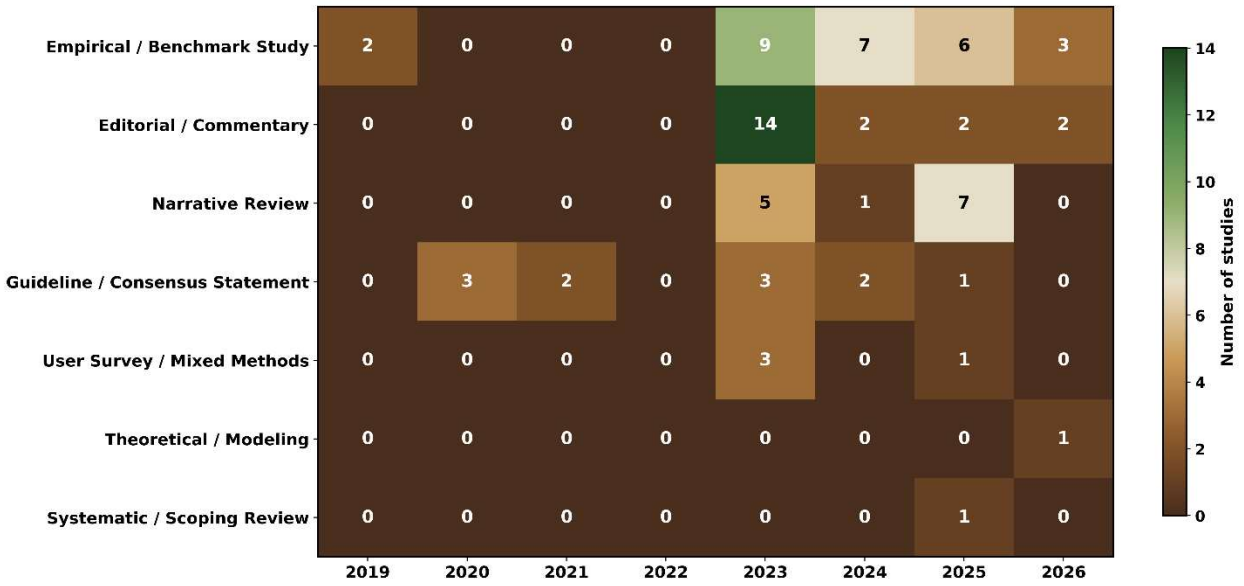


Fig. 4 Study design by publication year (heatmap), full corpus (n = 77)

4.2 AI Use in Peer Review

Table 4 presents selected findings on AI use in peer review. Nine studies in the corpus address AI use in peer review directly [11, 12, 24–26, 40–43]. These studies converge on three findings.

First, reviewers report using generative AI tools to draft review comments, summarize manuscripts, or check for errors. Such use is already reported [11, 41], although the corpus does not measure how widespread it is. Second, uploading a submitted manuscript to a third-party AI service raises a confidentiality problem. This problem is distinct from the quality question of whether AI-assisted review is useful, because the manuscript's content leaves the journal's control either way [12, 24]. Third, editorial associations and individual journals have begun to issue guidance. This guidance is not consistent across publishers in what it permits [42, 43].

One feasibility study tested whether a general-purpose chatbot could produce a usable peer review report. The output required substantial editing before it met the standard of a human-written review. This suggests current assistive use works better as a supplement to reviewer judgment than a replacement for it [40]. A separate commentary argues that generative AI is already present in the peer review pipeline and that formal policy has not kept pace. The authors base this on patterns observed in submitted review reports [11].

Three further studies extend this theme. A narrative review summarizes how AI tools can support the mechanics of the peer review and publishing pipeline. Named functions include reviewer matching, initial quality screening, and language editing. The review cautions that these support functions should not replace substantive expert judgment on a manuscript's scientific merit [44]. A separate narrative treatment of ethical considerations in AI-assisted peer review reaches a similar conclusion from a different angle. It argues that current practice and current guidance have not kept pace with each other [45]. An older empirical study predates generative AI but is relevant to the theoretical framework in Section 2.3. It found that publishing peer review reports publicly changed referee behavior in five scholarly journals. This bears on how transparency mechanisms might interact with AI-assisted review once that practice becomes visible to readers [35].

Table 4 Selected findings on AI use in peer review

Study	Design	Key finding
[25]	Narrative review	AI chatbot use reshapes peer review while raising unresolved confidentiality and accountability risks
[40]	Empirical	Chatbot-generated review reports required substantial editing before meeting human-review standards
[11]	Editorial	Argues that generative AI is already present in submitted peer review reports and that policy has not kept pace
[24]	Guideline	Recommends against uploading manuscripts to AI tools that retain content; permits disclosed assistive use
[44]	Narrative review	AI can support review-pipeline mechanics but should not substitute for expert judgment

4.3 AI Authorship and Accountability

Eleven studies address whether an AI system can be listed as an author or credited for its contribution [5–7, 46–53]. Table 5 shows selected findings on AI authorship. Nearly all of these studies converge on one position. An AI system cannot meet standard authorship criteria, because those criteria require the ability to take accountability for a work's content. An AI system cannot be held accountable in the relevant sense [5, 7]. This position aligns with the ICMJE authorship framework. That framework predates generative AI but was written broadly enough to exclude non-accountable contributors by construction [18]. Policy surveys nonetheless find that the treatment of AI authorship is not uniform across publishers [6].

The disagreement in this literature is not about whether AI can be an author. Consensus on that question is strong. The disagreement concerns how much AI-generated content a work can contain before its human authors' claim to authorship becomes questionable. One study frames this as a spectrum from cocreation to ghostwriting. It argues that existing authorship and ownership frameworks were not built to handle contributions that fall between clearly assistive and clearly generative [46].

A qualitative study of user experience adds a complication. People who used AI to generate text often did not perceive themselves as the owner of that text. This held even when they took public credit for it. The authors term this the AI ghostwriter effect [47]. This finding matters for the authorship debate. It suggests the accountability gap that ICMJE-style criteria are meant to close may persist even when an author discloses AI use and formally takes responsibility.

Table 5 Selected findings on AI authorship

Study	Design	Key finding
[7]	Editorial	NLP systems cannot fulfill authorship criteria requiring accountability for content
[5]	Editorial	An AI chatbot cannot be listed as author of a scholarly article under current norms
[6]	Empirical	Survey of top-journal policies finds inconsistent treatment of AI authorship across publishers
[46]	Editorial	Existing authorship frameworks do not cleanly handle cocreation between human and AI
[47]	User survey	Users often do not perceive ownership of AI-generated text despite claiming credit

4.4 Disclosure and Undisclosed-Use Detection

Fourteen studies address disclosure policy, undisclosed AI use, or detection tools [1–4, 19–21, 54–60]. Table 6 presents selected findings on detection tool reliability. A survey of publisher and journal instructions to authors found that most major publishers had issued some generative AI policy by 2024. The content of these policies varied substantially. They differed in what they required authors to disclose and how disclosure was to be verified [56]. Several studies argue that undisclosed AI-assisted writing should be treated as research misconduct. Their reasoning rests on existing fabrication, falsification, and plagiarism definitions: the reader is misled about the text's origin regardless of whether the underlying claims are accurate [19, 20].

Detection tool accuracy is the most empirically contested topic in the corpus. Two independent studies evaluated commercial AI-text detectors. Both found accuracy rates that varied widely by text sample and detector. Some detectors performed close to chance on adversarially edited text [1, 4]. A more recent analysis found a documented fairness problem in these tools. They flag text written by non-native English speakers as AI-generated at a higher rate than text written by native speakers, independent of whether AI was actually used [2]. A separate analysis reached a stronger conclusion. It argues that, where the true prevalence of AI-generated text is low, most positive detector findings will be false positives. This has direct implications for any journal considering detector-based screening as a primary enforcement mechanism [3].

Empirical estimates of actual undisclosed AI use rates are sparse but point in a consistent direction. One study documented specific cases of undisclosed generative AI use in published academic articles, found through manual and semi-automated review [20]. A longitudinal analysis of open-access medical literature found a rising rate of detected AI-generated text over time. The detected rate exceeded the author-disclosed rate in the same period. The authors caution that detection and disclosure measure related but not identical things [60].

Three further studies add discipline-specific and stakeholder-specific detail to this picture. A cross-sectional survey of medical researchers found generally favorable attitudes toward permitted AI use in the scientific process. It also found strong support for mandatory disclosure. This suggests the disclosure gap documented elsewhere in this theme is not primarily a matter of researchers rejecting disclosure in principle [61]. A discipline-specific analysis of neurosurgical publications found a measurable prevalence of likely AI-generated text using a detection approach. The authors frame the implications around academic integrity and ethical authorship, in a surgical subspecialty literature not previously examined for this issue [62]. A large-scale application of a fine-grained AI-text detector to the biomedical literature characterized where AI-generated writing concentrates, by section and publication type. This offers more granular detection evidence than the single-tool accuracy studies discussed above [63].

Table 6 Selected findings on detection tool reliability

Study	Design	Key finding
[4]	Empirical	AI-text detection tools show inconsistent accuracy across manuscript samples
[1]	Empirical	Detector performance varies substantially by tool and text type tested
[2]	Empirical	Detectors flag non-native English writing as AI-generated at a higher rate
[3]	Empirical	Argues that most positive detector findings will be false where prevalence is low

4.5 Reporting Guidelines for AI-Related Research

Eleven studies in the corpus are themselves guideline or consensus-statement papers [8–10, 24, 27–29, 37, 38, 64, 65]. These guidelines cluster into two groups.

The first group extends the CONSORT and SPIRIT clinical-trial reporting tradition to studies where an AI system is the research object under evaluation. This includes CONSORT-AI and SPIRIT-AI for clinical trials [27, 28], TRIPOD-AI and PROBAST-AI for diagnostic and prognostic prediction models [8, 9], and TRIPOD-LLM for studies evaluating large language models specifically [10]. The ARRIVE 2.0 guideline for animal research is not AI-specific. It is a methodological ancestor of this tradition, and several AI-guideline papers cite it as a template [37]. Likewise, one included consensus statement addresses equitable authorship in international research partnerships rather than AI [38]. It bears on the authorship questions in Section 4.3 and on the global-equity concerns raised in Section 5.4.

The second group addresses AI use in the research and writing process, not AI as a research object. The CANGARU initiative aims to produce reporting guidance for generative AI and LLM use across the research process. It was developed through a living systematic review and Delphi consensus among an international panel [29]. The WAME recommendations address chatbot and generative AI use in scholarly manuscripts specifically. They recommend disclosure to editors and advise against uploading confidential manuscript content to third-party AI services [24]. A university-level framework proposes institutional policy for responsible generative AI use that researchers and departments can adopt directly [65]. A comparative review of this emerging guideline landscape catalogs the competing options, including TRIPOD-LLM, CANGARU, and several smaller proposals. It offers authors and editors practical guidance on which guideline fits which study type [66]. A separate empirical study developed an updated scoring tool to assess how well published manuscripts adhere to the TRIPOD+AI guideline once adopted. It found adherence uneven across the studies it audited. This suggests guideline publication alone does not guarantee guideline uptake [67]. A narrative review takes the researcher and author perspective, rather than the editor perspective that dominates the rest of this theme. It argues that guideline proliferation itself has become a burden. Authors now face the task of identifying which of several overlapping guidelines applies to their specific study before they can even begin reporting [68].

No guideline in either group addresses one common scenario with the same specificity that CONSORT-AI and TRIPOD-LLM apply to AI-driven prediction models. That scenario is a researcher using a general-purpose chatbot for ordinary drafting and editing tasks. This asymmetry, detailed guidance for AI as a research object and thinner guidance for AI as a writing tool, recurs across the disclosure literature in Section 4.4. We return to it in Section 5.

4.6 AI-Assisted Detection of Research Fraud

Eight studies address the use of AI systems to detect research fraud, including paper mill activity and fabricated publications [13–16, 69–72]. Table 7 illustrates selected findings on AI-assisted fraud and paper mill detection. Paper mills, meaning organizations that produce and sell fraudulent manuscripts at scale,

predate generative AI. Several studies in the corpus argue that generative AI has both accelerated paper mill output and created new detection opportunities [13, 14].

A network analysis of retracted and suspected-fraudulent publications identified a small number of editors and brokers who enable fraud at scale. The authors argue that detection efforts focused only on individual manuscripts miss this network structure [14]. Two studies proposed AI-based screening pipelines. One applies a large language model to flag retracted or likely-fraudulent biomedical articles, with an explanation of the flagged features [70]. The other applies AI-based screening across a large corpus of cancer-research publications to identify suspected paper-mill products at a scale manual review cannot match [15].

A separate early study demonstrated that generative AI could itself produce a fraudulent but superficially authentic-looking medical research article [16]. This illustrates a dual-use pattern. The same technology that raises detection opportunities also lowers the cost of producing the fraud being detected.

Table 7 Selected findings on AI-assisted fraud and paper mill detection

Study	Design	Key finding
[14]	Empirical	A small number of editors and brokers enable paper-mill fraud at scale through identifiable networks
[70]	Empirical	An LLM-based system detects retracted biomedical articles with explainable flagged features
[15]	Empirical	AI-based screening identifies suspected paper-mill products across millions of publications
[13]	Systematic review	Synthesizes prevalence, detection methods, and research-integrity impact of paper mills, 2010-2025
[16]	Empirical	Generative AI can produce fraudulent but authentic-looking medical research articles

4.7 AI Use in Manuscript Preparation and Scientific Writing

Fifteen studies address AI use in drafting, editing, or otherwise producing manuscript text directly, independent of the peer review and authorship questions treated in Sections 4.2 and 4.3 [73–87]. This is one of the larger themes in the corpus by study count, which is consistent with manuscript preparation being an early point at which researchers used generative AI.

Two findings recur across this theme. First, several studies report that generative AI can produce prose that is difficult for human readers to distinguish from human-written scientific text. A widely cited study asked blinded reviewers and a detection tool to distinguish AI-generated scientific abstracts from real ones. The reviewers correctly identified AI-generated abstracts only some of the time, and the detection tool performed better than the human reviewers but still made errors in both directions [77]. Second, several studies flag hallucination as the central risk of generative use in scientific writing. Hallucination means fabricated citations, statistics, or claims presented with the same confidence as accurate content. Several authors argue this risk is distinct from, and arguably more consequential than, plagiarism [73, 86]. One classification study proposes a taxonomy of distorted information in AI-generated content for the scientific writing context. It distinguishes factual fabrication from citation fabrication and from subtler forms of unsupported inference [86].

A separate group of studies takes a more applied view, evaluating or advocating specific uses of AI in the writing process rather than assessing risk in the abstract. These include using AI to assist with scientific writing broadly [85], writing review articles specifically [79, 80], and academic writing in general [74, 78, 81]. A narrative synthesis of this applied literature reaches a clear conclusion. AI tools are most reliable for editing, structuring, and language-polishing tasks. They are least reliable when asked to generate novel scientific claims or interpret results. This distinction maps onto the assistive-generative spectrum introduced in Section 2.2 [76, 82].

One theoretical contribution proposes a faceted framework for attributing AI-assisted contributions to a manuscript. It breaks a single disclosure statement into separate facets for ideation, drafting, editing, and verification. The argument is that a single undifferentiated disclosure statement obscures more than it reveals [87]. This proposal has not yet been adopted or tested empirically by any study in the corpus, but it addresses directly the disclosure-granularity problem discussed in Section 5.4.4.

Academic integrity considerations specific to the post-pandemic, post-ChatGPT period appear as a recurring frame across this theme. Several studies situate generative AI as a continuation of longstanding academic integrity concerns, such as contract cheating, plagiarism, and data fabrication, rather than a wholly new category of risk [75, 84]. This framing has implications for policy design. It suggests that institutions with mature academic integrity infrastructure may be better positioned to extend existing processes to AI-related cases, rather than build parallel new ones.

4.8 Publishing Incentives and Systemic Pressure

One study in the corpus addresses the structural publishing incentives that make AI-assisted fabrication attractive in the first place, independent of any specific AI technology. This study documents publish-or-perish metrics losing their value as accurate quality signals once researchers and institutions begin optimizing directly for them, an instance of Goodhart's Law applied to bibliometrics [36]. Although this study predates widespread generative AI availability, it is relevant to the present review because it identifies the incentive structure that paper mills and fabrication schemes exploit. Generative AI lowers the marginal cost of producing publishable-looking output, which sharpens rather than creates the metric-optimization problem this study describes.

4.9 Quality and Transparency of the Evidence Base

We assessed the 37-study verified subset against five publication-quality and transparency domains. Open access availability is universal in this subset: all 37 studies (100%) are openly accessible. Abstract completeness is similarly high, with 35 of 37 studies (94.6%) reporting a substantive abstract of at least 50 characters. These two findings mean the evidence base for this review is, at minimum, easy for other researchers to locate and read in full. That is not a given in research-integrity scholarship generally.

Two other domains show more limited transparency. Only 2 of 37 studies (5.4%) are indexed across more than one bibliographic database, which suggests that much of this literature is discoverable through only one of the indexing sources we used. A researcher searching only one of these sources may therefore miss part of this corpus, even before the topical relevance problem described in Section 3.3 is considered. Citation impact, measured as citation count at or above the corpus median, splits close to evenly (19 of 37 studies, 51.4%). This split is expected by construction given the median-split definition, and it says nothing about outliers, which are visible in Figs. 5 and 8.

The evidence-hierarchy domain is the least favorable finding in this assessment. Only 15 of 37 studies in the verified subset (40.5%) fall into the higher-evidence tier. Section 3.5 defines this tier as systematic or scoping review, empirical or benchmark study, or user survey or mixed-methods study. The remaining 22 studies (59.5%) fall into the middle tier (guideline or consensus statements, theoretical or modeling work) or the lower tier (narrative reviews, editorials, and commentary), which do not carry the same methodological safeguards as primary empirical research. A similar picture holds across the full 77-study corpus (Table 2): 32 studies (41.6%) fall in the higher tier, 12 (15.6%) in the middle tier, and 33 (42.9%) in the lower tier. It is the quantitative basis for the qualitative pattern described throughout Section 4. Much of what is known about AI and research integrity in scholarly publishing rests on expert opinion and position statements rather than controlled measurement. This is not a flaw specific to the studies included here. It reflects how quickly the underlying technology changed relative to the time a rigorous empirical study takes to design, run, and publish.

Fig. 5 shows that citation counts vary substantially by study design in the verified subset. Guideline and consensus statements have the highest median (754) and the highest single-study maximum (5,802), and empirical and benchmark studies show a wide spread (maximum 1,513). This is consistent with a small

number of early, highly visible papers anchoring the field's citation landscape, while later entrants have not yet accumulated comparable citation counts.

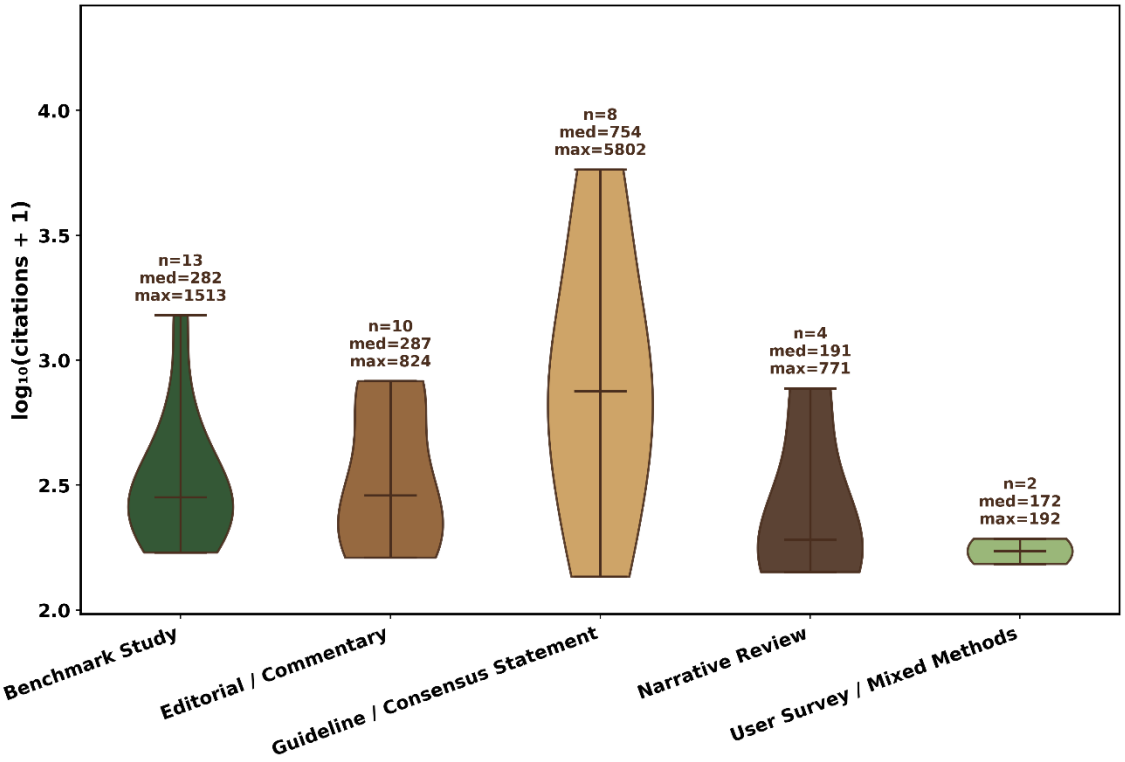


Fig. 5 Citation distribution by study design (violin plot), verified-metadata subset (n=37)

Fig. 6 compares the concept-field profile of the verified subset against the full 250-record metadata pool, as a check on whether relevance screening introduced a systematic topical bias. The two profiles track closely. Computer science and artificial intelligence are the most frequently tagged concepts in both the pool and the verified subset, followed by psychology, engineering, data science, political science, and medicine. This overlap is expected given the review's subject matter. The close match is compatible with the relevance screen removing off-topic records without markedly changing the disciplinary mix of what remained, although concept tags are an imperfect proxy for topic.

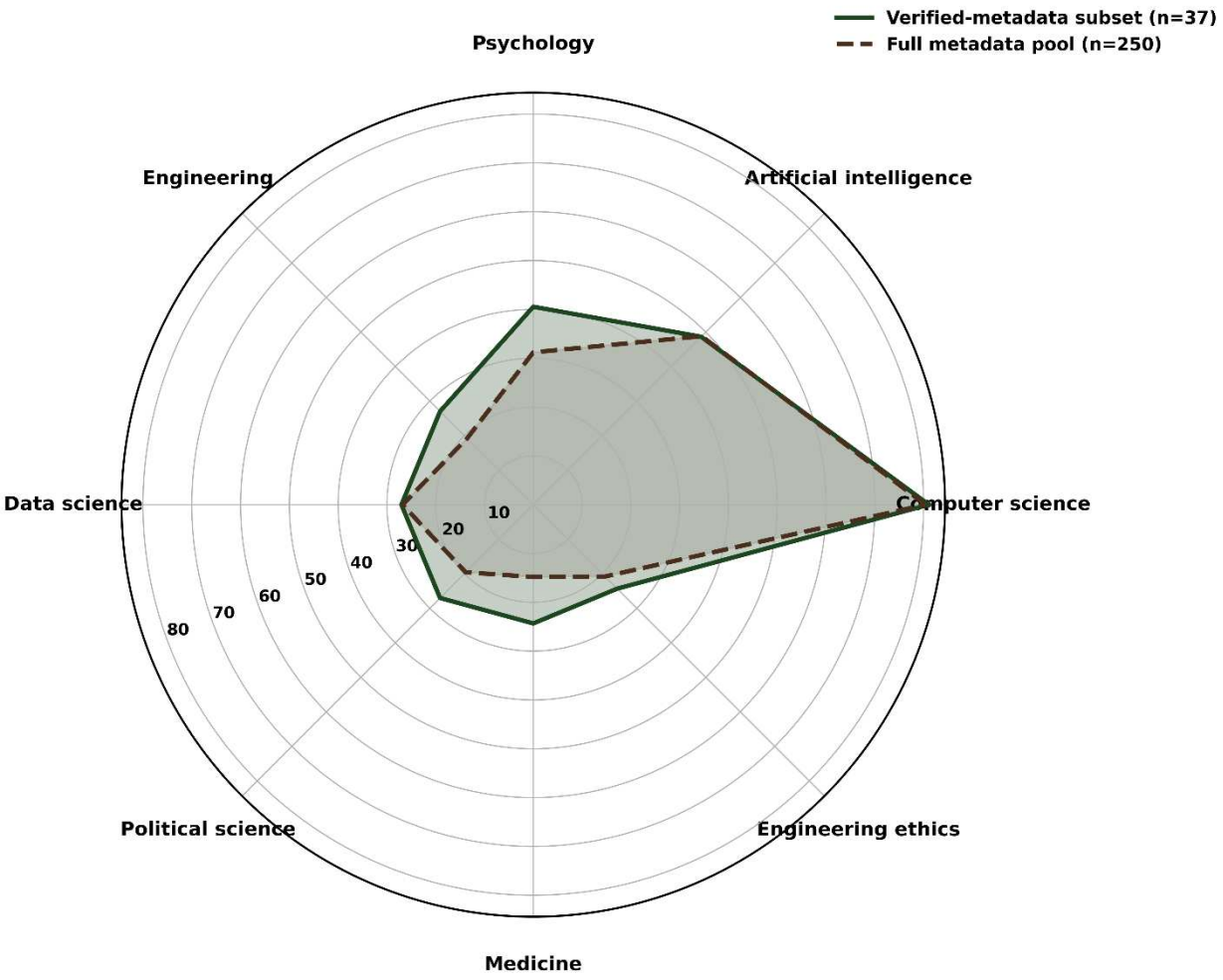


Fig. 6 Concept-field coverage (radar plot), verified subset vs. full metadata pool

Fig. 7 plots citation count against publication year, with bubble size encoding each study's concept-field breadth and color encoding study design. It shows the expected pattern of newer studies clustering at lower citation counts. This reflects citation accumulation time rather than lower quality or relevance. Bubble size does not track year in any visible pattern, which suggests that topical breadth and publication recency are not clearly related in this corpus.

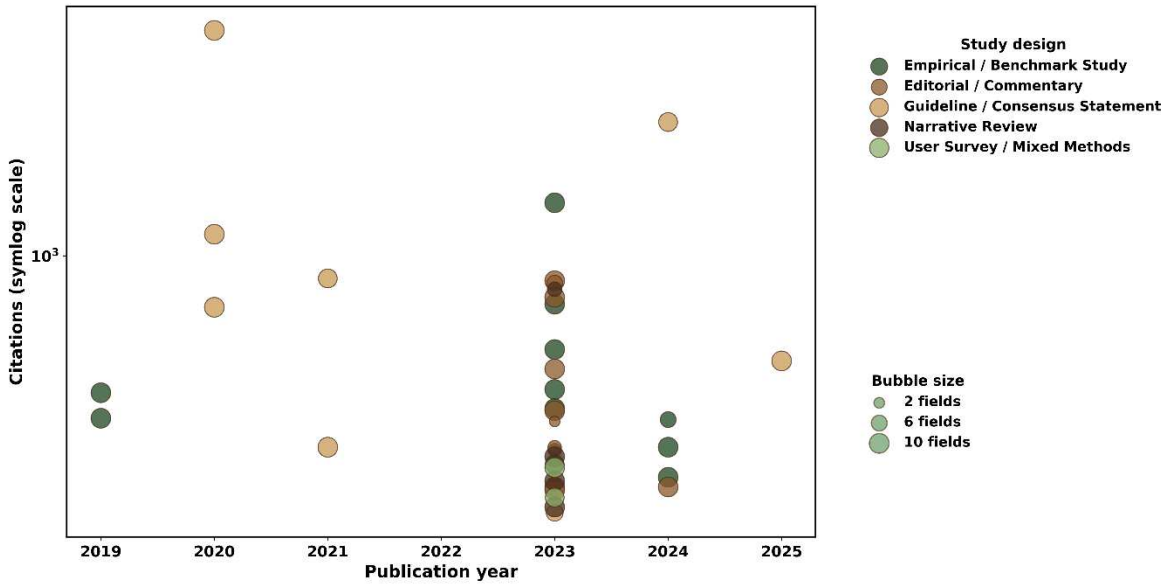


Fig. 7 Citations vs. publication year (bubble plot), bubble size = concept-field breadth, verified subset (n = 37)

Fig. 8 tests that independence more directly with a plain scatter plot of citation count against concept-field breadth alone, fitted with a linear trend line. The fitted slope is positive but the relationship is weak and driven by a small number of high-citation outliers rather than a consistent trend across the full range. A single guideline paper with the maximum concept-field breadth in the subset also holds the highest citation count, which pulls the fitted line upward without indicating a general pattern. We report this finding descriptively rather than testing it for statistical significance, given the small sample (n=37) and the outlier sensitivity visible in the plot itself.

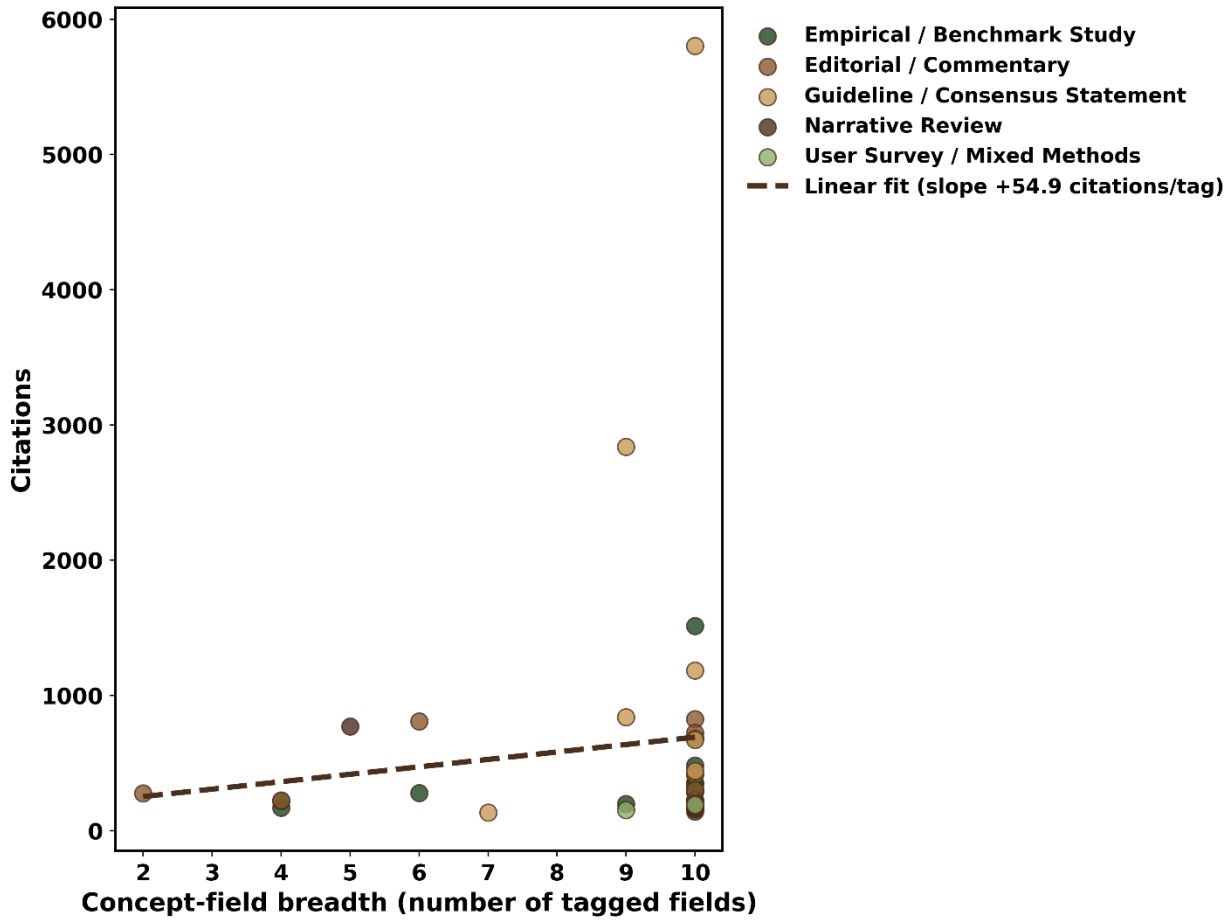


Fig. 8 Citations vs. concept-field breadth (scatter plot with linear fit), verified subset (n = 37)

Fig. 9 reports a citation-based funnel plot restricted to the 12 verified-subset studies that report an extractable sample size, plotting citation impact against a sample-size-derived precision proxy. We report this plot as a descriptive check in the spirit of an Egger's-test funnel plot. It is not a formal test of publication bias. Because citation counts are not effect sizes, the plot cannot support any inference about publication bias, and we include it mainly to document how few studies report a sample size. No included study in the full corpus reports a comparable effect size that would support a proper funnel-plot asymmetry test. The small number of studies with an extractable sample size (12 of 77) is itself a finding. Most of the literature in this area, being editorial, commentary, or narrative-review work, does not report a study population size at all. This limits what any future meta-analysis in this specific area could quantitatively pool without a substantial expansion of primary empirical research.

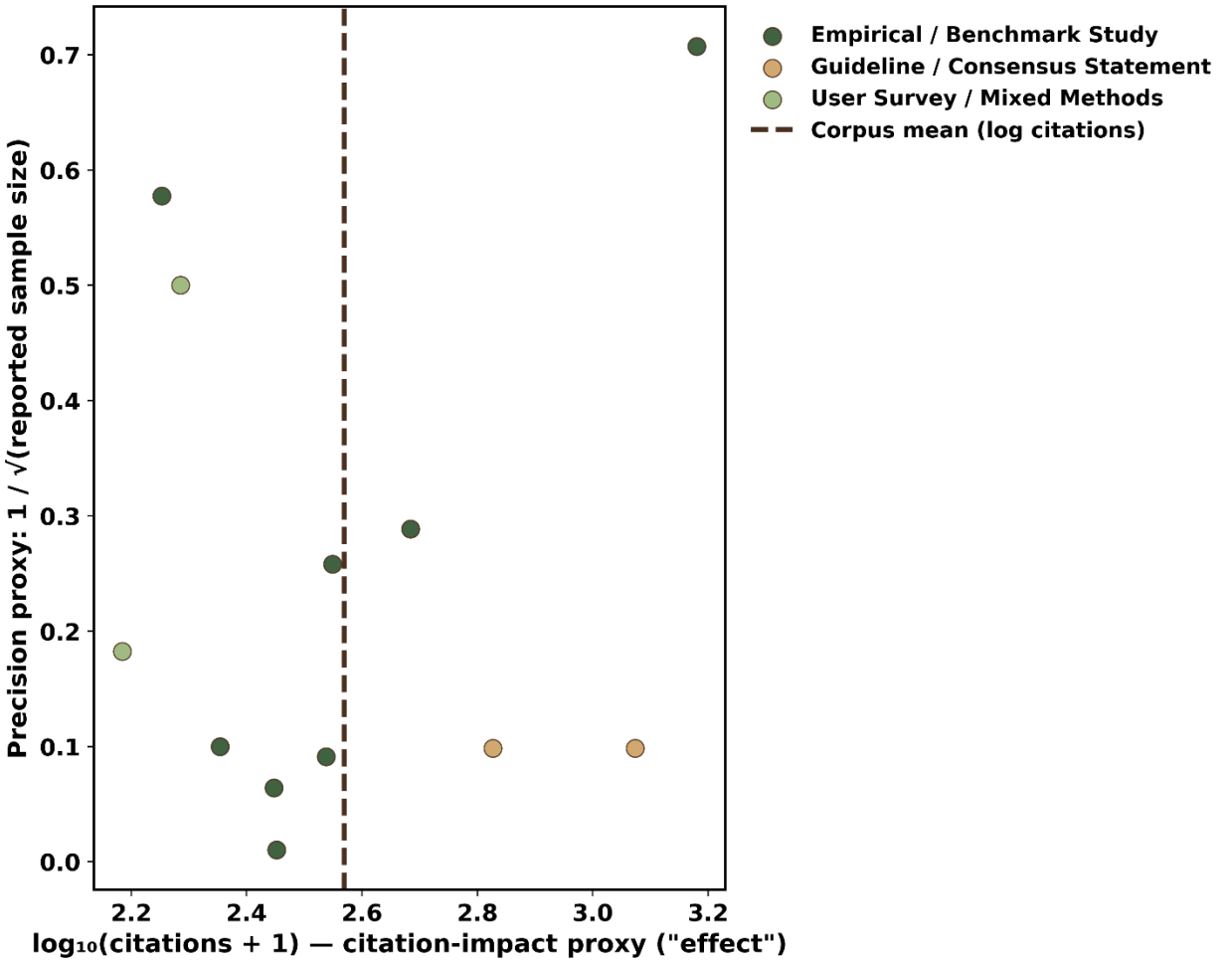


Fig. 9 Citation-based funnel plot (Egger's-test-style descriptive check), studies with reported sample size (n = 12)

Fig. 10 reports the publication-quality and evidence-hierarchy assessment described in Section 3.5, modeled on the tiered appraisal approach of the Mixed Methods Appraisal Tool [39]. Panel A shows the evidence-hierarchy tier by study design across the full 77-study corpus. Panel B shows the five transparency domains for the 37-study verified subset, expressed as the percentage meeting each criterion. The two domains with the lowest percentage meeting criterion, multi-database indexing and higher-tier study design, are the same two domains flagged as the least favorable findings earlier in this section.

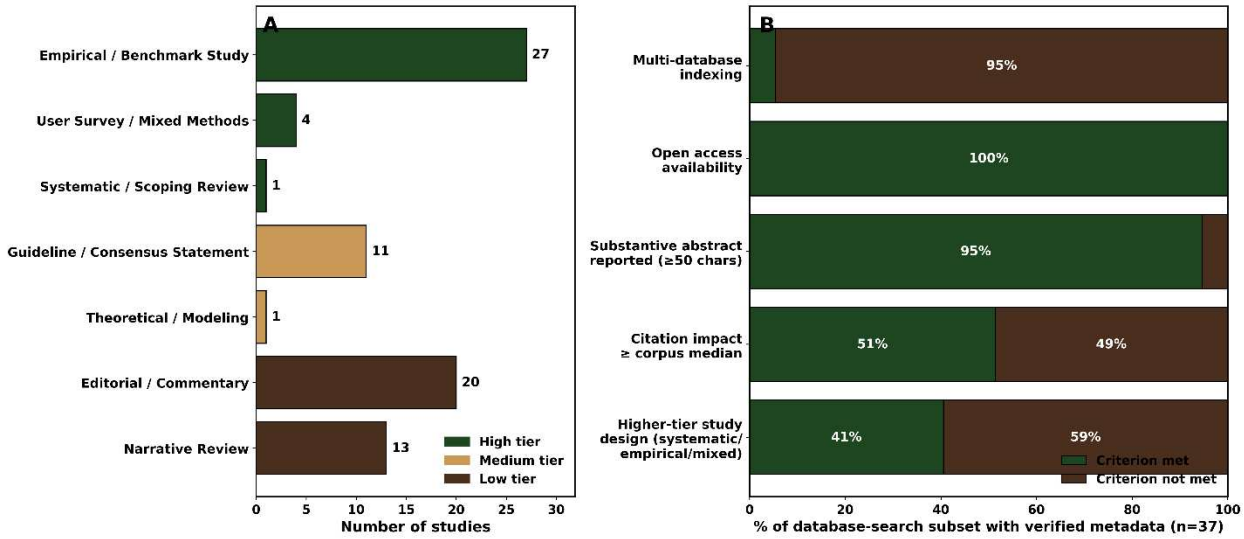


Fig. 10 Publication quality and evidence-hierarchy assessment (Mixed Methods Appraisal Tool-inspired tiering)

5. Discussion

5.1 Summary Relative to Aim

This review set out to map the published literature on AI and research integrity in scholarly publishing, organized around four questions. The evidence answers each question with a different level of confidence. On AI use in peer review, the literature indicates that reviewers are already using AI tools and that confidentiality is the sharpest unresolved risk. It does not yet establish how widespread such use is, or how often it is generative rather than assistive, because the studies available are feasibility tests, surveys, and editorial observation rather than large-scale measurement [40, 41]. On authorship, the literature shows strong convergence. AI systems do not meet accountability-based authorship criteria. This position holds across editorial opinion, empirical policy surveys, and the ICMJE framework itself [5–7, 18]. On detection tool reliability, the independent evidence in the corpus is consistently unfavorable to detector-based enforcement. Multiple independent studies show inconsistent accuracy and bias against non-native English writing. One analysis argues that positive findings will usually be false where prevalence is low [2, 3]. On AI-assisted fraud detection, the literature is newest and most technically oriented. It shows genuine promise for identifying paper mill output at scale, while also demonstrating that the same generative capability lowers the barrier to producing convincing fraud [14, 16]. Section 5.4 asks what follows ethically from these findings.

5.2 The Assistive-Generative Gap in Governance

The clearest pattern across all four themes is a governance gap between two situations. The first is AI as a research object, where an AI system's outputs are what a study evaluates. The second is AI as a writing tool, where an AI system helps produce the manuscript describing an entirely separate piece of research. Reporting guidelines address the first situation in detail. TRIPOD-LLM alone specifies 19 main reporting items for studies that develop or evaluate large language models [10]. Guidance for the second situation remains comparatively thin. The WAME recommendations and the CANGARU initiative are the two most developed efforts. Neither has the specificity or field-testing history that CONSORT-AI has accumulated for clinical AI trials [24, 27, 29].

This gap matters practically because the second situation, not the first, is what most authors and reviewers encounter day to day. A researcher drafting a discussion section with chatbot assistance is not developing a prediction model that needs a TRIPOD-LLM checklist. That researcher needs guidance closer to what WAME and CANGARU are building. That guidance is still under construction while use is already reported across the literature [24, 29].

5.3 Detection Cannot Substitute for Disclosure

A second pattern follows from Section 4.4. The corpus offers converging evidence that AI-text detection tools are not reliable enough to serve as a primary integrity-enforcement mechanism. Accuracy varies by tool and text sample [1, 4]. Detectors show a documented bias against non-native English writing. This means detector-based screening risks disproportionately flagging researchers from non-Anglophone institutions, independent of actual AI use [2]. Where the prevalence of AI-generated text is low, most positive detector flags can be expected to be false [3].

These findings point toward disclosure, not detection, as the more defensible governance mechanism. The ethical reasons for this are developed in Section 5.4. This holds only if disclosure is paired with a verification process rather than treated as self-enforcing. Several studies in the corpus make exactly this point. Disclosure without verification does not close the gap that undisclosed use opens. An author already willing to omit a required disclosure is not deterred by the disclosure requirement itself [19, 20]. What verification should look like remains an open question this review's evidence base cannot answer. No included study tested a verification mechanism against a comparison condition. Section 5.4.4 proposes constraints that any such mechanism should meet.

5.4 Ethical Analysis: What Is Owed, to Whom, and at Whose Cost

The preceding sections describe what the literature reports. This section asks what follows ethically. It states positions that are ours, defends each with reasoning and evidence from the corpus, and identifies who bears the costs of the status quo. These positions are not findings of the included studies, and several go beyond what any single study argued. Table 8 summarizes them.

5.4.1 Fairness and due process in detector-led enforcement

Position. A detector's output should never, on its own, be the basis for rejecting a manuscript, questioning an author's integrity, or imposing a sanction.

Reasoning. Three features of the evidence in Section 4.4 make detector-led enforcement ethically indefensible as a primary mechanism. First, accuracy is inconsistent across tools and texts [1, 4], so any given flag carries an unknown and variable error rate. Second, errors are not randomly distributed. Detectors flag non-native English writing as AI-generated at a higher rate [2], a pattern documented earlier in a widely cited study from outside our corpus [88]. Third, where the true prevalence of AI-generated text is low, most positive flags will be false [3]. A policy of acting on flags would therefore impose the risk of false accusation unequally, and the risk would fall on researchers who already face disadvantage in English-language publishing. This is a question of justice, not only of accuracy: equal treatment requires that people not face a higher chance of an adverse finding for reasons unrelated to their conduct. It also engages non-maleficence, because an accusation of misconduct can damage a career even when it is later withdrawn. The concept of testimonial injustice helps to name a further harm [34]. If an editor treats the prose style of a non-native writer as grounds for doubting their honesty, the writer's word is given less credibility because of who they are, and the detector lends that prejudice the appearance of objectivity.

Costs and benefits. The benefit of detector-led screening is administrative efficiency for publishers and editors. The cost is borne by authors, and disproportionately by non-native English writers. Where one party gains and another bears the risk, the arrangement is hard to justify unless safeguards shift the risk back.

What would change our view. Independent evidence that a specific detector, validated on non-native writing at realistic prevalence, had low and equal error rates across groups would support limited use as one input among several. Even then, due process would require that authors be told of the concern, shown its basis, and given a real opportunity to respond to a human decision-maker.

5.4.2 Accountability and the limits of delegation

Position. Responsibility for the content of a manuscript is non-delegable. Disclosure of AI assistance does not transfer, dilute, or share that responsibility.

Reasoning. The authorship literature converges on excluding AI systems because they cannot be held accountable [5, 7, 18]. The ethical point behind this is that accountability requires an agent who can answer

for a claim, correct it, and bear the consequences of error. The responsibility-gap literature identifies the problem for systems whose outputs no human fully controls [33]. Excluding AI from authorship closes the gap only if human authors in fact accept the responsibility that the exclusion assigns to them. The ghostwriter effect reported in the corpus suggests this cannot be assumed: people who used AI to generate text often did not perceive themselves as its owner while still claiming credit [47]. An author who does not feel ownership of a passage is less likely to check it, and hallucinated citations or unsupported claims [73, 86] can enter the record under a human name. We therefore argue that accountability should be treated as something authors must do (check, verify, and stand behind their text), not only a status that a byline confers. The faceted attribution proposal asks authors to say what they did and not only that they used a tool [87]. It points in a useful direction, although it has not been tested.

Limit. This argument does not say that AI use is wrong. Language polishing and structuring are the uses that the corpus describes as most reliable [76, 82]. The burden of verification falls on authors, which is appropriate because they alone decide what to submit, and journals can support that burden by being clear about what authors are expected to verify.

5.4.3 Confidentiality, consent, and institutional responsibility in peer review

Position. Reviewers must not upload submitted manuscripts to AI services that retain or reuse content, and publishers share responsibility for providing compliant alternatives.

Reasoning. Peer review rests on an implicit agreement: authors share unpublished work on condition that it is read by chosen reviewers and kept confidential. A reviewer who uploads a manuscript to a third-party service transmits the author's work, without consent, to a party the journal does not control [12, 24]. This breaches a duty of confidentiality and undermines authors' autonomy over their own unpublished work. Delegating the assessment itself to a tool also weakens the accountability on which the authority of review depends [25, 26].

The individual duty is clear. The structural question is who should bear the burden of compliance. Reviewers are usually unpaid and under time pressure, a concern the corpus raises through the problem of reviewer fatigue [12]. A policy that prohibits uploading but offers no secure alternative places the whole burden on individuals and invites quiet noncompliance. The guidelines of the journal AI and Ethics illustrate a better structure: they ask reviewers not to upload manuscripts to generative AI tools, state that the publisher is exploring safe tools for reviewers, and ask reviewers to declare any AI support in their reports [23]. We read this as the right shape of obligation, a clear individual duty paired with an institutional commitment. The cost of the transition should be borne by those who organize and benefit from the review system, and not left to reviewers to absorb.

5.4.4 Honesty, disclosure, and proportionate verification

Position. Journals may fairly require disclosure of substantive AI use, but should pair it with verification that is proportionate, non-discriminatory, and open to challenge. The current evidence does not justify blanket surveillance.

Reasoning. Honesty about how a text was produced is an integrity value, and several studies in the corpus treat undisclosed AI use as a misrepresentation to readers about origin [19, 20]. Disclosure without verification, however, is vulnerable to the behavior it aims to deter, because an author willing to omit a required disclosure is unlikely to be deterred by the requirement (Section 5.3). Verification is not ethically free either. Any mechanism that checks disclosure statements, such as auditing the drafting process, requesting version histories, or screening with tools, imposes burdens on authors, may intrude on privacy, and can reproduce the unfairness of detectors if it relies on them. We therefore propose four constraints that any verification mechanism should meet. It should be proportionate, with scrutiny triggered by specific concerns or random audit rather than applied to everyone. It should be non-discriminatory, avoiding signals correlated with language background. It should respect due process, with notice, evidence, and a human decision-maker. It should minimize data, collecting only what is needed and keeping it no longer than necessary. The corpus contains no test of verification mechanisms (Section 5.3), so these are constraints on design and not claims about effectiveness.

Disclosure must also be fair in what it asks. A requirement to disclose every use of AI, including routine language correction, would burden most heavily the authors who rely on such help to write in a second language, and it would invite stigma. Some publishers exempt assistive copy editing from declaration, as the guidelines of the journal AI and Ethics do [23]. We think this line is ethically defensible because it tracks the reader's interest: readers need to know when a tool contributed substantive content or analysis, not when it corrected grammar. The assistive-generative distinction (Section 2.1) is imperfect, but disclosure rules keyed to it are more proportionate than rules that treat all use alike. Disclosure is fair only if it does not itself attract a penalty. If authors who disclose are treated less favorably than those who do not, the policy rewards concealment, so editors should state explicitly that compliant disclosure is not grounds for adverse treatment.

5.4.5 Dual use, incentives, and the distribution of costs

Position. Responsibility for the integrity risks created by generative AI is shared by developers, publishers, institutions, and authors. Defensive burdens should not fall mainly on the least resourced actors.

Reasoning. The corpus shows the dual-use character of the technology. It can produce a convincing fraudulent article [16], and it can help detect paper-mill output at scale [15, 70]. The benefit of detection, which is protection of the scholarly record, must be weighed against the fairness constraints of Section 5.4.1, because screening models that flag articles can also produce false positives with serious consequences for individual authors. Work on publishing metrics shows that systems driven by a measurable target invite optimization that erodes the target's meaning [36], and generative AI lowers the cost of such optimization. Treating the problem only as one of finding individual bad actors leaves the incentive structure intact and places the burden on editors and honest authors. Evidence that a small number of editors and brokers enable fraud at scale points to institutional responsibility [14]. Institutions and funders that reward publication counts are part of the problem, and publishers whose workflows are exploited are part of the solution.

Global equity. Several points bear on global justice. Detector bias falls on non-native English writers [2]. The corpus draws mainly on English-language, largely biomedical, venues. A consensus statement in the corpus addresses equitable authorship in international research partnerships [38]. Access to compliant tools may also differ across institutions, a point the corpus does not test. A governance regime designed around the assumptions of well-resourced, English-speaking research settings risks exporting its costs to researchers elsewhere. This is a reason to involve a wider range of researchers in policy design and to monitor the differential impact of any enforcement tool.

5.4.6 Summary of the ethical analysis

Table 8 Ethical issues, evidence, and positions defended in Section 5.4

Ethical issue	Principle(s) engaged	Evidence in the corpus	Who bears the cost now	Position defended
Detector-led enforcement	Justice; non-maleficence	Inconsistent accuracy; bias against non-native English writing; false positives where prevalence is low (Section 4.4)	Authors, especially non-native English writers	Detector output alone must not ground adverse action; due process is required
Delegation of responsibility for AI-assisted text	Explicability (accountability); non-maleficence	Consensus that AI cannot be an author; ghostwriter effect; hallucination risk (Sections 4.3, 4.7)	Readers and the scholarly record	Responsibility is non-delegable; disclosure does not transfer it
Uploading manuscripts to AI services in peer review	Autonomy; non-maleficence; explicability	Confidentiality and accountability concerns; inconsistent guidance (Section 4.2)	Authors (exposure of unpublished work); reviewers	Reviewers must not upload to services that retain content; publishers must

Ethical issue	Principle(s) engaged	Evidence in the corpus	Who bears the cost now	Position defended
			(compliance without support)	provide compliant tools
Disclosure and verification	Justice; autonomy (privacy); honesty	Disclosure policies vary; no tested verification mechanism (Sections 4.4, 5.3)	Authors, especially those who rely on AI for language support	Proportionate disclosure rules; verification that is proportionate, non-discriminatory, and open to challenge
Fraud, incentives, and global equity	Beneficence; justice	Dual-use evidence; fraud networks; incentive distortion (Sections 4.6, 4.8)	Honest authors, editors, and readers; researchers outside well-resourced English-language settings	Shared responsibility across developers, publishers, and institutions; monitor differential impact

5.5 Strengths

This review has methodological strengths that should be read alongside the limitations in Section 5.6. First, the corpus construction is documented, with the number of records at each stage reported (Fig. 2 and Table 3), so readers can see how the final corpus was assembled. Second, when an initial stratified-by-design sampling approach produced a corpus with serious topical validity problems (Section 3.3), we discarded it and rebuilt the corpus with a relevance screen, and we report that change instead of omitting it. Third, the supplementary search stage in Section 3.2 followed a stated procedure with seven query strings, and every added record was individually verified against its source. Fourth, the explicit separation of evidence from ethical argument (Sections 2.5 and 5.4) lets readers accept the evidence map while contesting our normative positions.

5.6 Limitations

Search and coverage. We searched three aggregators (Semantic Scholar, OpenAlex, and CrossRef) rather than subject-specific databases such as PubMed, Scopus, or Web of Science, and we did not exhaustively search gray literature or institutional repositories. Policy documents, publisher guidelines, and regional reports that are not indexed in academic catalogs may therefore be missing. Our queries were written in English, so literature in other languages is probably under-represented. Given the equity concerns raised in Section 5.4.5, this is a limitation of the review as well as a feature of the evidence base. Many included studies appear in biomedical or clinical venues, and the reporting guidelines discussed are largely clinical, so our conclusions may not generalize to the humanities, social sciences, or engineering.

Corpus construction. The supplementary targeted search contributed 40 of the 77 studies, more than the database search itself (37). Targeted searching depends more on the authors' judgment than systematic database searching does, and the bibliometric analyses in Section 4.9 apply only to the 37 database-search studies. Annual publication counts (Fig. 3) are shaped by the search design and the partial coverage of 2026 and should not be read as a publication trend. Keyword-based relevance screening also retained four records that are not primarily about AI (Section 3.3).

Data and synthesis. Full-text extraction was not possible for the database-search subset (Section 3.4), so study design and theme assignment relied principally on titles, abstracts, and metadata. Misclassification is possible, and abstracts may omit nuance relevant to our interpretation. Theme assignment was an interpretive judgment. The assistive-generative spectrum and the governance gap we describe are our

analytic constructs and are not terms used by all included studies. The heterogeneity of formats prevents quantitative meta-analysis or direct statistical comparison across the corpus.

Quality appraisal. We did not perform a formal risk-of-bias appraisal. Most included papers are editorials, position statements, and guideline documents, for which standard appraisal tools do not apply, and this is consistent with scoping-review methodology [17]. The evidence-tier classification in Section 4.9 describes corpus composition and should not be read as a judgment of any study's quality. Seven included records are preprints (arXiv or bioRxiv), whose findings may change. The citation-based analyses (Figs. 5 to 9) are descriptive, are confounded by publication age, and cannot support inferences about study quality or publication bias.

The ethical analysis. Our analysis draws on a limited set of frameworks: principle-based AI ethics, the responsibility gap, and testimonial injustice. Other traditions, such as virtue ethics, care ethics, or non-Western ethical frameworks, might weigh the same evidence differently. The positions in Section 5.4 are arguments, not empirical findings, and they rest on an evidence base that is changing quickly.

5.7 Implications for Editorial Policy

Four implications follow from this evidence base and the ethical analysis for editors and publishers setting AI-related policy now.

First, journals should treat disclosure requirements and detection tools as separate mechanisms serving separate purposes, not as substitutes for each other. Detection tools are not accurate enough to anchor an enforcement policy on their own [2, 3], and any use of detector output in an editorial decision should follow the due-process safeguards in Section 5.4.1. Second, peer review confidentiality policy should address AI use explicitly, and publishers should provide compliant tools alongside the prohibition. Reviewers should not have to infer whether uploading a manuscript to a chatbot is permitted. The WAME recommendations offer a model journals can adapt without waiting for field-specific guidance to mature [24]. Third, disclosure policies should state what must be disclosed and what need not be (Section 5.4.4), and should be paired with verification that meets the constraints set out there. Fourth, journals publishing AI-related primary research should adopt the appropriate existing reporting guideline (CONSORT-AI, SPIRIT-AI, TRIPOD-AI, TRIPOD-LLM) where the study design falls within its scope, rather than developing a new one. These guidelines were built and validated recently, and duplicating that effort wastes the consensus they already represent [9, 10, 27, 28].

5.8 Stakeholder-Specific Implications

The implications set out in Section 5.7 apply mainly to editors and publishers. Four other stakeholder groups appear throughout the corpus with distinct, sometimes conflicting interests, and the evidence supports separate implications for each.

Authors. Several studies in the corpus treat undisclosed AI use, rather than disclosed AI use, as the misconduct concern [19, 20]. An author who discloses substantive AI use in accordance with a journal's policy is acting transparently, in a way that is consistent with the authorship frameworks represented in this corpus. An author who does not disclose has no comparable position if their text is later flagged by a detector prone to false positives, even when the flag itself is wrong [2, 3]. Given Section 5.4.1, authors should also be able to expect that a flag alone will not be treated as proof. We suggest that authors disclose substantive use by default, keep a record of how AI was used, and ask the journal when its policy is silent.

Peer reviewers. The confidentiality risk identified in Section 4.2 falls specifically on reviewers, who handle confidential manuscript content that authors have not yet made public. A reviewer who uploads a manuscript to a third-party AI tool exposes another researcher's unpublished work without that researcher's consent [12, 24]. The WAME guidance, that reviewers should not upload manuscripts to AI tools that retain submitted content, is the most direct evidence-based recommendation this review can offer to individual reviewers [24]. The institutional counterpart is described in Section 5.4.3.

Editors and editorial boards. Beyond the policy implications in Section 5.7, editors face a coordination problem that this review's evidence base does not resolve. Guidance varies across publishers and journals [42, 43, 56], so an author or reviewer working across multiple journals encounters inconsistent rules for

functionally similar situations. No study in the corpus tested whether harmonized cross-publisher guidance improves compliance relative to journal-specific guidance.

Research institutions and funders. Institutions appear less often in this corpus than journals do. The studies that do address institutional policy point toward a specific recommendation: extending existing academic integrity infrastructure to AI-related cases rather than building parallel new processes [65, 84]. Research integrity offices, ethics committees, and academic integrity panels already have established procedures for handling fabrication, falsification, and plagiarism allegations. The evidence reviewed in Section 4.7 suggests that undisclosed AI use fits within, rather than outside, that existing procedural framework, which may lower the institutional cost of responding to it. Institutions and funders also share responsibility for the incentive structures discussed in Section 5.4.5.

5.9 Future Research

The evidence base would benefit most from four kinds of studies that are largely absent from the current corpus. First, prospective studies should measure actual disclosure rates against actual AI use rates in the same manuscript sample, rather than measuring each separately. This would let the field quantify the disclosure gap directly rather than infer it. Second, studies should test a verification mechanism against a no-verification control, and should report its differential impact on authors by language background and region. This would give journals evidence to design disclosure policy, rather than adopt policy on the strength of editorial argument alone. Third, empirical work should examine whether disclosure itself carries reputational or editorial penalties, since the fairness of disclosure policy depends on the answer. Fourth, future studies in this area should report sample sizes and quantitative outcomes in a form that supports meta-analytic synthesis. The current literature's heavy reliance on editorial and commentary formats limits what any future systematic review can quantitatively pool.

6. Conclusions

This scoping review synthesized 77 studies on AI and research integrity in scholarly publishing and analyzed their ethical implications. Four conclusions follow from the evidence. First, reviewers are using AI tools in peer review, and uploading manuscripts to third-party services raises confidentiality and accountability problems. Second, AI systems do not meet the accountability-based authorship criteria of the major frameworks reviewed. Third, the independent evidence in the corpus indicates that AI-text detectors are unreliable as a basis for enforcement and show bias against non-native English writing. Fourth, AI-based screening shows promise for uncovering paper-mill output, while generative tools also lower the cost of producing convincing fraud. Reporting guidelines for studies that evaluate AI systems are well developed within the clinical domains they target, whereas guidance for AI used as a writing tool is less mature.

On ethical grounds, we argued that detector output alone is an unfair basis for adverse action, particularly against non-native English writers; that responsibility for a manuscript's content cannot be delegated to AI; that the duty of reviewer confidentiality is matched by an institutional duty to provide compliant tools; and that disclosure requirements should be proportionate to the reader's interest and paired with verification that is proportionate, non-discriminatory, and open to challenge. These are normative positions, and the effectiveness of verification mechanisms remains untested. Future research should measure disclosure against actual use, evaluate verification and its differential impact on researchers, and support shared reporting standards for AI as a writing aid, with researchers from diverse regions and disciplines involved.

References

1. Elkhatat, A.M., Elsaid, K., Al-Meer, S.: Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity* 19, 17 (2023). https://doi.org/10.1007/s40979-023-00140-5

2. Pratama, A.R.: The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science* 11, e2953 (2025). https://doi.org/10.7717/peerj-cs.2953
3. Tsigaris, P., Teixeira da Silva, J.A.: AI detecting AI in academic writing: Why most AI detector findings are false. *Next Research* 7, 101396 (2026). https://doi.org/10.1016/j.nexres.2026.101396
4. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., Waddington, L.: Testing of detection tools for AI-generated text. *International Journal for Educational Integrity* 19, 26 (2023). https://doi.org/10.1007/s40979-023-00146-z
5. Lee, J.Y.: Can an artificial intelligence chatbot be the author of a scholarly article? *Journal of Educational Evaluation for Health Professions* 20, 6 (2023). https://doi.org/10.3352/jeehp.2023.20.6
6. Lund, B.D., Orhan, Z., Mannuru, N.R., Bevara, R.V.K., Pochineni, M., Mah, J., Wang, Z.: Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals. *Learned Publishing* 37, 13–23 (2024). https://doi.org/10.1002/leap.1582
7. Yeo-Teh, N.S.L., Tang, B.L.: NLP systems such as ChatGPT cannot be listed as an author because these cannot fulfill widely adopted authorship criteria. *Accountability in Research* 30, 533–536 (2023). https://doi.org/10.1080/08989621.2023.2177160
8. Collins, G.S., Dhiman, P., Andaur Navarro, C.L., Ma, J., Hooft, L., Reitsma, J.B., Logullo, P., Beam, A.L., Peng, L., Van Calster, B., van Smeden, M., Riley, R.D., Moons, K.G.M.: Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 11, e048008 (2021). https://doi.org/10.1136/bmjopen-2020-048008
9. Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J.B., van Smeden, M., Boulware, L.E., Logullo, P.: TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385, e078378 (2024). https://doi.org/10.1136/bmj-2023-078378
10. Gallifant, J., et al.: The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine* 31, 60–69 (2025). https://doi.org/10.1038/s41591-024-03425-5
11. Cheng, K., Sun, Z., Liu, X., Wu, H., Li, C.: Generative artificial intelligence is infiltrating peer review process. *Critical Care* 28, 149 (2024). https://doi.org/10.1186/s13054-024-04933-z
12. Hosseini, M., Horbach, S.P.J.M.: Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review* 8, 4 (2023). https://doi.org/10.1186/s41073-023-00133-5
13. Kamel, R., Barghi, N.G.: Paper mills in scholarly publishing: A systematic review of their prevalence, detection, and impact on research integrity. *Trends in Scholarly Publishing* 4, 81–87 (2025). https://doi.org/10.21124/tsp.2025.81.87
14. Richardson, R.A.K., Hong, S., Byrne, J.A., Stoeger, T., Amaral, L.A.N., de Waard, A.: The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. *Proceedings of the National Academy of Sciences* 122, e2420092122 (2025). https://doi.org/10.1073/pnas.2420092122
15. Scancar, B., Byrne, J.A., Causeur, D., Barnett, A.G.: Revealing the paper mill iceberg: AI-based screening of cancer research publications. *bioRxiv* (2025). https://doi.org/10.1101/2025.08.29.673016
16. Májovský, M., Černý, M., Kasal, M., Komarc, M., Netuka, D.: Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of Medical Internet Research* 25, e46924 (2023). https://doi.org/10.2196/46924
17. Tricco, A.C., et al.: PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine* 169, 467–473 (2018). https://doi.org/10.7326/M18-0850
18. International Committee of Medical Journal Editors: Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. https://www.icmje.org (2024)

19. Tang, B.L.: Undeclared AI-assisted academic writing as a form of research misconduct. *Science Editor* 48, 1–7 (2025). https://doi.org/10.36591/SE-4804-02
20. Glynn, A.: Academ-AI: Documenting the undisclosed use of generative artificial intelligence in academic publishing. *arXiv* (2024). https://doi.org/10.48550/arXiv.2411.15218
21. Hosseini, M., Resnik, D.B., Holmes, K.: The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics* 19, 449–465 (2023a). https://doi.org/10.1177/17470161231180449
22. Committee on Publication Ethics: Authorship and AI tools: COPE position statement. https://publicationethics.org/cope-position-statements/ai-author (2023)
23. Springer Nature: AI and Ethics: Submission guidelines. https://link.springer.com/journal/43681/submission-guidelines (2026). Accessed 5 November 2026
24. Zielinski, C., Winker, M.A., Aggarwal, R., Ferris, L.E., Heinemann, M., Lapeña, J.F., Jr., Pai, S.A., Ing, E., Citrome, L., Alam, M., Voight, M., Habibzadeh, F.: Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Colombia Médica* 54, e1015622 (2023). https://doi.org/10.25100/cm.v54i3.5622
25. Ben Saad, H., Dergaa, I., Ghouili, H., Ceylan, H. İ., Chamari, K., Dhahbi, W.: The assisted technology dilemma: A reflection on AI chatbots use and risks while reshaping the peer review process in scientific research. *AI & Society* 40, 5649–5656 (2025). https://doi.org/10.1007/s00146-025-02299-6
26. Chauhan, C., Currie, G.: The impact of generative artificial intelligence on the external review of scientific manuscripts and editorial peer review processes. *American Journal of Pathology* 194, 1829–1832 (2024). https://doi.org/10.1016/j.ajpath.2024.08.002
27. Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., Denniston, A.K., SPIRIT-AI and CONSORT-AI Working Group: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine* 26, 1364–1374 (2020). https://doi.org/10.1038/s41591-020-1034-x
28. Rivera, S.C., Liu, X., Chan, A.-W., Denniston, A.K., Calvert, M.J., SPIRIT-AI and CONSORT-AI Working Group: Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine* 26, 1351–1363 (2020). https://doi.org/10.1038/s41591-020-1037-7
29. Cacciamani, G.E., Eppler, M.B., Ganjavi, C., Pekcan, A., Biedermann, B., Collins, G.S., Gill, I.S.: Development of the ChatGPT, generative artificial intelligence and natural large language models for accountable reporting and use (CANGARU) guidelines. *arXiv* (2023). https://doi.org/10.48550/arXiv.2307.08974
30. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. *Harvard Data Science Review* 1(1) (2019). https://doi.org/10.1162/99608f92.8cd550d1
31. Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics, 8th edn. Oxford University Press, New York (2019)
32. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
33. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6, 175–183 (2004). https://doi.org/10.1007/s10676-004-3422-1
34. Fricker, M.: Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press, Oxford (2007)
35. Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., Squazzoni, F.: The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications* 10, 322 (2019). https://doi.org/10.1038/s41467-018-08250-2

36. Fire, M., Guestrin, C.: Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. *GigaScience* 8, giz053 (2019). https://doi.org/10.1093/gigascience/giz053
37. Percie du Sert, N., et al.: The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology* 18, e3000410 (2020). https://doi.org/10.1371/journal.pbio.3000410
38. Morton, B., Vercueil, A., Masekela, R., Heinz, E., Reimer, L., Saleh, S., Kalinga, C., Seekles, M., Biccard, B., Chakaya, J., Abimbola, S., Obasi, A., Oriyo, N.: Consensus statement on measures to promote equitable authorship in the publication of research from international partnerships. *Anaesthesia* 77, 264–276 (2021). https://doi.org/10.1111/anae.15597
39. Hong, Q.N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., Pluye, P.: The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information* 34, 285–291 (2018). https://doi.org/10.3233/EFI-180221
40. Biswas, S., Dobaria, D., Cohen, H.L.: ChatGPT and the future of journal reviews: A feasibility study. *The Yale Journal of Biology and Medicine* 96, 415–420 (2023). https://doi.org/10.59249/SKDH9286
41. Ebadi, S., Nejadghanbar, H., Salman, A.R., Khosravi, H.: Exploring the impact of generative AI on peer review: Insights from journal reviewers. *Journal of Academic Ethics* 23, 1383–1397 (2025). https://doi.org/10.1007/s10805-025-09604-4
42. Leimeister, J.M., Aaltonen, A., Currie, W., Schlagwein, D.: From 'ChatGPT et al.' to generative artificial intelligence (GenAI) in peer review: Safeguarding trust while modernising scholarly publishing. *Journal of Information Technology* 41, 1–15 (2026). https://doi.org/10.1177/02683962261458829
43. Peer review in the time of artificial intelligence. *Nature Nanotechnology* 21, 479 (2026). https://doi.org/10.1038/s41565-026-02177-2
44. Kousha, K., Thelwall, M.: Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing* 37, 4–12 (2023). https://doi.org/10.1002/leap.1570
45. Bozkurt, A.: Artificial intelligence in scholarly peer review: Ethical considerations, current practices, and future implications. *The International Review of Research in Open and Distributed Learning* 27, 11–30 (2026). https://doi.org/10.19173/irrodl.v27i3.10081
46. Bozkurt, A.: GenAI et al.: Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. *Open Praxis* 16, 1–10 (2024). https://doi.org/10.55982/openpraxis.16.1.654
47. Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., Welsch, R.: The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction* 31, 1–40 (2023). https://doi.org/10.1145/3637875
48. Flanagin, A., Bibbins-Domingo, K., Berkwits, M., Christiansen, S.L.: Nonhuman "authors" and implications for the integrity of scientific publication and medical knowledge. *JAMA* 329, 637–639 (2023a). https://doi.org/10.1001/jama.2023.1344
49. Grimes, M., von Krogh, G., Feuerriegel, S., Rink, F., Gruber, M.: From scarcity to abundance: Scholars and scholarship in an age of generative artificial intelligence. *Academy of Management Journal* 66, 1617–1625 (2023). https://doi.org/10.5465/amj.2023.4006
50. Homolak, J.: Opportunities and risks of ChatGPT in medicine, science, and academic publishing: A modern Promethean dilemma. *Croatian Medical Journal* 64, 1–3 (2023). https://doi.org/10.3325/cmj.2023.64.1
51. Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., Smith, A.: Generating scholarly content with ChatGPT: Ethical challenges for medical publishing. *The Lancet Digital Health* 5, e105–e106 (2023). https://doi.org/10.1016/S2589-7500(23)00019-5
52. Lin, Z.: Why and how to embrace AI such as ChatGPT in your academic life. *Royal Society Open Science* 10, 230658 (2023). https://doi.org/10.1098/rsos.230658
53. Lubowitz, J.H.: ChatGPT, an artificial intelligence chatbot, is impacting medical literature. *Arthroscopy* 39, 1121–1122 (2023). https://doi.org/10.1016/j.arthro.2023.01.015

54. Anderson, N., Belavý, D.L., Perle, S.M., Hendricks, S., Hespanhol, L., Verhagen, E., Memon, A.R.: AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sport & Exercise Medicine* 9, e001568 (2023). https://doi.org/10.1136/bmjsem-2023-001568
55. Chemaya, N., Martin, D.: Perceptions and detection of AI use in manuscript preparation for academic journals. *PLOS ONE* 19, e0304807 (2024). https://doi.org/10.1371/journal.pone.0304807
56. Ganjavi, C., Eppler, M.B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G.S., Gill, I.S., Cacciamani, G.E.: Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: Bibliometric analysis. *BMJ* 384, e077192 (2024). https://doi.org/10.1136/bmj-2023-077192
57. Gotoman, J.E.J., Luna, H.L.T., Sangria, J.C.S., Santiago, C.S., Jr., Barbuco, D.D.V.: Accuracy and reliability of AI-generated text detection tools: A literature review. *American Journal of IR 4.0 and Beyond* 4, 1–9 (2025). https://doi.org/10.54536/ajirb.v4i1.3795
58. Jarrah, A.M., Wardat, Y., Fidalgo, P.: Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say? *Online Journal of Communication and Media Technologies* 13, e202346 (2023). https://doi.org/10.30935/ojcmt/13572
59. Pellegrina, D., Helmy, M.: AI for scientific integrity: Detecting ethical breaches, errors, and misconduct in manuscripts. *Frontiers in Artificial Intelligence* 8, 1644098 (2025). https://doi.org/10.3389/frai.2025.1644098
60. Wolfrath, N., Patel, S., Flitcroft, M., Banerjee, A., Somai, M., Crotty, B.H., Kothari, A.N.: Rising prevalence of detected AI-generated text in medical literature: Longitudinal analysis in open access articles. *arXiv* (2026). https://doi.org/10.48550/arXiv.2603.19316
61. Ng, J.Y., Maduranayagam, S.G., Suthakar, N., Li, A., Lokker, C., Iorio, A., Moher, D., Chow, R.: Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: An international cross-sectional survey. *The Lancet Digital Health* 7, e48–e58 (2025). https://doi.org/10.1016/j.landig.2024.10.007
62. Schneider, D.M., Mishra, A., Gluski, J., Shah, H., Ward, M., Brown, E.D., Sciubba, D.M., Lo, S.-F.L.: Prevalence of artificial intelligence-generated text in neurosurgical publications: Implications for academic integrity and ethical authorship. *Cureus* 17, e79086 (2025). https://doi.org/10.7759/cureus.79086
63. She, R.: Fine-grained detection of AI-generated writing in the biomedical literature. *bioRxiv* (2026). https://doi.org/10.64898/2026.01.01.697311
64. Flanagin, A., Kendall-Taylor, J., Bibbins-Domingo, K.: Guidance for authors, peer reviewers, and editors on use of AI, language models, and chatbots. *JAMA* 330, 702–703 (2023b). https://doi.org/10.1001/jama.2023.12500
65. Smith, S.M., Tate, M., Freeman, K., Walsh, A., Ballsun-Stanton, B., Lane, M.: A university framework for the responsible use of generative AI in research. *Journal of Higher Education Policy and Management* 47, 17–36 (2025). https://doi.org/10.1080/1360080X.2025.2509187
66. Huo, B., Collins, G.S., Cacciamani, G.E., Guyatt, G.: Reporting guidelines for studies involving generative artificial intelligence applications: What do I use, and when? *npj Digital Medicine* 8, 646 (2025). https://doi.org/10.1038/s41746-025-02113-z
67. de Kanter, E., Kaul, T., Heus, P., de Groot, T.M., Kuijten, R.H., Reitsma, J.B., Collins, G.S., Hooft, L., Moons, K.G.M., Damen, J.A.A.: Adherence to TRIPOD+AI guideline: An updated reporting assessment tool. *Journal of Clinical Epidemiology* 191, 112118 (2026). https://doi.org/10.1016/j.jclinepi.2025.112118
68. Suh, D.H., Han, D.S.: Proper utilization of generative AI in scholarly writing: The researcher and author perspective. *Journal of Korean Medical Science* 41, e304 (2026). https://doi.org/10.3346/jkms.2026.41.e304

69. Bernard, C.: AI-generated scientific papers: Crisis? What crisis? *eNeuro* 13, ENEURO.0470-25.2025 (2026). https://doi.org/10.1523/ENEURO.0470-25.2025
70. Chen, L., Fu, S., Freedman, G., Zor, C., Martin, G., Kinross, J., Vaghela, U., Serban, O., Toni, F.: Pub-Guard-LLM: Detecting retracted biomedical articles with reliable explanations. *arXiv* (2025). https://doi.org/10.48550/arXiv.2502.15429
71. Elali, F.R., Rachid, L.N.: AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns* 4, 100706 (2023). https://doi.org/10.1016/j.patter.2023.100706
72. Sabel, B.A., Knaack, E., Gigerenzer, G., Bilc, M.-I.: Fake publications in biomedical science: Red-flagging method indicates mass production. *Naunyn-Schmiedeberg's Archives of Pharmacology* 399, 2943–2955 (2025). https://doi.org/10.1007/s00210-025-04275-9
73. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* 15, e35179 (2023). https://doi.org/10.7759/cureus.35179
74. Ateriya, N., Sonwani, N.S., Thakur, K.S., Kumar, A., Verma, S.K.: Exploring the ethical landscape of AI in academic writing. *Egyptian Journal of Forensic Sciences* 15, 36 (2025). https://doi.org/10.1186/s41935-025-00453-1
75. Crothers, E., Japkowicz, N., Viktor, H.L.: Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access* 11, 7163–7183 (2023). https://doi.org/10.1109/ACCESS.2023.3294090
76. Dergaa, I., Chamari, K., Żmijewski, P., Saad, H.B.: From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport* 40, 615–622 (2023). https://doi.org/10.5114/biolsport.2023.125623
77. Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., Pearson, A.T.: Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine* 6, 75 (2023). https://doi.org/10.1038/s41746-023-00819-6
78. Hosseini, M., Rasmussen, L.M., Resnik, D.B.: Using AI to write scholarly publications. *Accountability in Research* 31, 715–733 (2023b). https://doi.org/10.1080/08989621.2023.2168535
79. Huang, J., Tan, M.: The role of ChatGPT in scientific communication: Writing better scientific review articles. *American Journal of Cancer Research* 13, 1148–1154 (2023).
80. Kacena, M.A., Plotkin, L.I., Fehrenbacher, J.C.: The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports* 22, 115–121 (2024). https://doi.org/10.1007/s11914-023-00852-0
81. Khlaif, Z.N., Mousa, A., Hattab, M.K., Itmazi, J., Hassan, A.A., Sanmugam, M., Ayyoub, A.: The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *JMIR Medical Education* 9, e47049 (2023). https://doi.org/10.2196/47049
82. Kooli, C.: Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability* 15, 5614 (2023). https://doi.org/10.3390/su15075614
83. Lund, B., Wang, T., Mannuru, N.R., Nie, B., Shimray, S., Wang, Z.: ChatGPT and a new academic reality: Artificial intelligence–written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* 74, 570–581 (2023). https://doi.org/10.1002/asi.24750
84. Perkins, M.: Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice* 20, 1–24 (2023). https://doi.org/10.53761/1.20.02.07
85. Salvagno, M., Taccone, F.S., Gerli, A.G.: Can artificial intelligence help for scientific writing? *Critical Care* 27, 75 (2023). https://doi.org/10.1186/s13054-023-04380-2
86. Sun, Y., Sheng, D., Zhou, Z., Wu, Y.: AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications* 11, 1298 (2024). https://doi.org/10.1057/s41599-024-03811-x
87. Xexéo, G.: A faceted proposal for transparent attribution of AI-assisted text production. *arXiv* (2026). https://doi.org/10.48550/arXiv.2604.25346

88. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., Zou, J.: GPT detectors are biased against non-native English writers. *Patterns* 4, 100779 (2023). https://doi.org/10.1016/j.patter.2023.100779