applied sciences

MDPI

Review

Detecting AI-Generated Text and Assessing Citation Integrity in Scholarly Communication: A Scoping Review of Methods, Limitations, and Manipulation

Karol Wieczorek * and Daniel Kaczmarski *

Department of Information Systems, Kielce University of Technology, 7 Tysiąclecia Państwa Polskiego Ave., 25-314 Kielce, Poland
* Correspondence: k.wieczorek@tu.kielce.pl (K.W.); d.kaczmarski@tu.kielce.pl (D.K.)

Abstract

Large language models complicate text-origin assessment and increase risks from nonexistent, inaccurate, or strategically used citations. This scoping review maps methods for detecting AI-generated or substantially AI-modified text and for evaluating reference integrity and citation manipulation in scholarly communication. We searched Scopus and Web of Science Core Collection on 31 August 2026, and IEEE Xplore on 2 September 2026. Four thematic search streams yielded 1633 records after filters. Following deduplication, removal of five 2027 records, citation searching, title-and-abstract screening, and full-text assessment, 71 peer-reviewed publications were retained; 69 formed the main synthesis and two provided conceptual background. Detection approaches included statistical signals, conventional machine learning, transformer classifiers, stylometry, watermarking, retrieval, and process-based provenance. Performance varied across generators, domains, languages, text lengths, paraphrases, and mixed human–AI texts, limiting the evidentiary value of a single detector score. Citation-integrity studies addressed fabricated or inaccurate references, claim–source agreement, citation function, coercion, cartels, stacking, self-citation, paper mills, and metadata manipulation. On the basis of this mapped evidence, we propose a layered conceptual audit combining provenance, targeted detection, bibliographic verification, claim-level checking, and network or metadata analysis. The integrated workflow itself has not yet been empirically validated, and consequential decisions should remain subject to documented human review.

Keywords: AI-generated text; large language models; citation integrity; reference hallucination; stylometry; bibliometric manipulation; scoping review; PRISMA-ScR

Check for updates

Academic Editor: Rui Araújo

Received: 22 September 2026
Revised: 15 October 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Generative artificial intelligence has changed how academic texts are drafted, revised, and evaluated. Large language models can support editing and summarization, but they can also produce complete passages, imitate an author's style, and return references that look plausible even when they are wrong. A reliable assessment therefore requires two separate questions: how the text was produced, and whether its evidence and citations are sound. We examined these issues in a scoping review reported in accordance with PRISMA-ScR [1].

AI-text detection is probabilistic. A detector does not observe the writing process. It compares a submitted passage with patterns learned during development, including token probabilities, vocabulary, syntax, sentence rhythm, punctuation, semantic representations,

or similarity to an author's previous writing [2–21]. Its output is therefore a statistical judgment, not a direct observation of authorship.

This distinction matters in higher education and scholarly publishing. Academic documents are translated, edited, and revised, and many are created through mixed workflows in which a model drafts, rewrites, or polishes only selected passages. Comparative studies show that detector performance changes with the generating model, discipline, text length, author language background, and classification threshold [22–35]. Paraphrasing can sharply reduce detectability [36–39], while writers using English as an additional language may face a higher false-positive risk [22,40].

Text origin does not determine evidentiary quality. Human-written work may contain inaccurate or strategically selected citations, while AI-assisted work may contain references that the author has verified. Nevertheless, evaluations of chatbots show that they can generate convincing combinations of titles, author names, journals, and digital object identifiers (DOIs) for publications that do not exist or do not support the response [41–43]. Finding a document is therefore only the first step of verification.

A second problem is claim–source agreement. A reference can be bibliographically correct yet omit a qualification, strengthen a tentative conclusion, or point to a paper that is merely related to the topic [44–48]. Citation function also matters: a source may provide background, describe a method, support a result, mark a limitation, enable comparison, or serve as the object of criticism [49–55]. A citation count cannot represent these different roles.

Manipulation introduces another level of risk. The literature describes coerced citations, strategic self-citation, citation exchange between journals, citation cartels, paper mills, purchased citations, and references inserted into metadata even though they are absent from the published article [56–68]. Network and bibliometric methods can identify unusual concentration, reciprocity, or timing, but they rarely reveal intent. An algorithmic alert therefore requires examination of the full text, metadata, and editorial context.

AI-text detection and citation auditing have developed in different research traditions. The former draws mainly on natural language processing, machine learning, stylometry, and process traces. The latter combines metadata verification, claim–source checking, citation-function classification, and bibliometric network analysis. In editorial practice, however, both are part of the same integrity assessment. Treating them together helps prevent a detector score from replacing a direct examination of evidence.

The contribution of this review is not a new detector or evidence that text-generation signals cause citation problems. It is the methodological integration of previously separate evidence streams within one decision process. The proposed audit keeps five questions distinct—writing-process provenance, text-origin signals, bibliographic existence and metadata, claim–source support and citation function, and network or metadata anomalies—so that a weak signal in one layer cannot substitute for direct evidence in another. This five-layer structure is an author-developed conceptual synthesis of the mapped literature, not an empirically validated diagnostic instrument.

A scoping review is appropriate because the field includes technical benchmarks, comparative experiments, linguistic analyses, case studies, citation-network research, and publication-ethics scholarship. These designs do not estimate one common effect and should not be pooled in a single meta-analysis. The purpose is to map available methods and findings, identify the conditions under which they remain informative, and show where evidence is still insufficient.

This review organizes methods for detecting text generated or substantially modified by language models and methods for assessing citation integrity in scholarly communication. It addresses five questions:

1. Which methods are used to detect AI-generated or AI-modified text in academic communication?
2. How does detector reliability vary by model, discipline, language, text length, mixed authorship, and paraphrasing?
3. Which reference errors and hallucinations occur in AI-assisted content, and how can they be verified?
4. How are natural language processing, network analysis, and bibliometrics used to identify citation function, distortion, and manipulation?
5. Which components should be included in an integrated audit of text origin and citation integrity?

The remainder of the article follows the review process. Section 2 describes the review design, information sources, search strategies, eligibility criteria, deduplication, data charting, and descriptive quality appraisal. Section 3 reports the selection process and the characteristics of the 71 included publications. Section 4 synthesizes evidence on text-origin identification, the limitations of detection, reference integrity, citation function, and bibliometric manipulation. Section 5 discusses the findings, practical implications, research gaps, and limitations of the review, and Section 6 presents the conclusions. Appendix A connects each analyzed publication with its corresponding reference-list entry, while Appendix B reproduces the complete search strategies.

2. Materials and Methods

The methods were designed to make identification, screening, coding, and synthesis reproducible from the retained search log and selection table. Because the review spans heterogeneous study designs, we used a scoping-review approach with qualitative thematic synthesis. Reporting follows the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [1]. PRISMA-ScR is used as a reporting framework, not as the analytical method itself.

2.1. Review Design

The review was organized into four streams: (A) detection of AI-generated or AI-modified text; (B) reference existence, accuracy, and hallucination; (C1) citation function and context; and (C2) citation and bibliometric manipulation. These streams structured both the search and the subsequent coding. The protocol was not prospectively registered. Search strings, counts, and decisions were retained in a review workbook to provide an auditable record.

2.2. Information Sources and Search Dates

Formal searches were conducted in Scopus and Web of Science Core Collection on 31 August 2026, and in IEEE Xplore on 2 September 2026. These sources were selected because Scopus and Web of Science provide broad interdisciplinary indexing across computing, education, medicine, scholarly communication, and bibliometrics, while IEEE Xplore adds technical and conference literature that may be underrepresented in multidisciplinary indexes. The combination was intended to support a broad, reproducible map rather than to claim exhaustive coverage. An exploratory search in the ACM Digital Library Basic Edition returned 114 records after refinement, but the available interface did not permit the required filtering and batch export. ACM Digital Library was therefore not included in the formal record set. No separate, documented search was conducted in ACL Anthology or PubMed. Publications indexed only in those sources may consequently have been missed. Citation searching and overlap with the three formal databases may have recovered some ACL conference papers and medical AI studies, but the extent of that recovery was not measured. This limitation is considered when interpreting the results and should be addressed in a future update.

For stream A, the date range was 2019–2026. Streams B, C1, and C2 had no lower year limit and ended on the search date. Records assigned a 2027 publication year by a database were removed before screening because they fell beyond the prespecified cutoff. English-language articles, conference papers, and reviews were included. Early-access records were retained when otherwise eligible. The complete database-specific strings and filters are reproduced in Appendix B.

2.3. Search Strategy

Each query combined terms for the phenomenon with terms for the task or scholarly context. Scopus searches used TITLE-ABS-KEY, Web of Science used Topic (TS), and IEEE Xplore used All Metadata/Command Search. Wildcards were adjusted to the syntax supported by each platform. After filtering, 875 records were exported from Scopus, 569 from Web of Science, and 189 from IEEE Xplore. Stream totals were 733 for AI-text detection, 351 for reference accuracy and hallucination, 456 for citation function and context, and 93 for citation manipulation.

The primary contextual query for stream C2 returned no results in IEEE Xplore. A sensitivity search that removed the contextual block returned eight results, five of which met the document filters and were exported. Scopus also displayed 155 preprints in its separate preprint collection. These records were not merged with the peer-reviewed search results.

Stream C1 was deliberately framed as a targeted search for computational methods that identify citation function or context because research question 4 asked specifically how natural language processing, machine learning, and related automated methods are used. The computational term block increased specificity and made the search manageable, but it also narrows coverage. Noncomputational studies of rhetorical roles, citation practice, and scholarly communication that do not use this terminology may be underrepresented. Citation searching partly broadened the evidence base, but it does not allow the size of this potential omission to be quantified. Accordingly, the C1 results should be read as a map of computational citation-function research rather than a comprehensive history of citation analysis.

2.4. Eligibility Criteria

Records were eligible when they presented or evaluated a method for text-origin detection; compared automated detectors with human judgment; examined stylometry, authorship, paraphrasing, bias, or process traces; assessed reference existence, metadata, or claim relevance; classified citation function or context; or studied coercive citation, cartels, citation exchange, paper mills, metadata manipulation, or strategic self-citation. The application had to concern academic text, scholarly publishing, or a method directly transferable to these settings.

Records were excluded when search terms were only incidental, the topic did not address any review stream, the item was a news report, erratum, or abstract without substantive content, or the full text could not be obtained. Conceptual papers without original data or a method could be retained as background but were not treated as empirical evidence. Preprints were treated as supplementary search material; when a peer-reviewed version was available, that published version was used in the analysis and reference list.

2.5. Deduplication and Screening

Duplicates were removed hierarchically. Records with the same DOI were merged first. When no DOI was available, exact matching was performed on normalized titles after standardizing case, punctuation, and spacing. Of 1633 exported records, 615 duplicates were removed, leaving 1018 unique database records. Five records assigned a 2027 publication

year were then removed. Citation searching identified 24 additional records. In total, 1037 titles and abstracts were screened.

At this stage, 960 records were excluded and 77 reports were sought for full-text assessment. Six full texts could not be obtained. They were not assessed from abstracts alone and were not included in the synthesis. The final set comprised 71 publications. One reviewer performed screening and appraisal with computational support for record organization and preliminary relevance tagging. A second reviewer conducted targeted cross-checks of questioned records and the final included set, but did not independently rescreen the records excluded at title-and-abstract screening. When a cross-check raised uncertainty, the record was returned to the prespecified eligibility or appraisal criteria and discussed until a reasoned decision was reached; the selection table retained the final decision and its basis. Because these checks were targeted rather than a prespecified, blinded duplicate sample, neither a disagreement count nor an inter-reviewer agreement coefficient would provide a valid estimate of reproducibility and none was calculated. Computational support was limited to maintaining the structured review workbook, normalizing DOI and title fields, flagging exact potential duplicates, and assigning preliminary keyword-based relevance tags. It made no inclusion, exclusion, category, or appraisal decision. The absence of complete independent duplicate screening and appraisal is a material limitation of this review.

2.6. Data Charting and Coding

For each included publication, we recorded the title, authors, year, DOI or another persistent publication identifier, source, document type, discovery route, synthesis role, dominant topic, and descriptive appraisal. Seven thematic categories were used: A1, detection methods and models; A2, detector and human-performance comparisons; A3, linguistic features, stylometry, and authorship verification; A4, robustness, evasion, bias, and provenance; B, reference accuracy and hallucination; C1, citation function and context; and C2, citation and metric manipulation. Each publication was assigned to one dominant category while connections to other topics were retained in the narrative synthesis. The dominant category was selected from the publication's primary stated objective and principal unit of analysis. When more than one stream applied, the category corresponding to the method and results with the main analytical role was used; secondary connections were retained in the narrative synthesis. The category counts are therefore mutually exclusive, and no fractional allocation was used.

The extraction form captured the research problem, data type, unit of analysis, method, validation approach, main findings, and reported limitations. Accuracy, precision, recall, F1, and area under the receiver operating characteristic curve (AUROC) values were preserved as defined by the original authors. They were not converted into a common effect because the underlying tasks, class distributions, and test conditions differed substantially. Screening and coding were conducted using Microsoft Excel 2019 MSO, Version 16.0.10417.20228 (64-bit; Microsoft Corporation, Redmond, WA, USA). Bibliographic records were managed using Zotero 10.0.6 (64-bit; Corporation for Digital Scholarship, USA). No laboratory materials, measuring devices, or other physical equipment were used.

2.7. Descriptive Quality Appraisal and Synthesis

Quality was appraised descriptively across six areas: clarity of the objective, description of the data or sample, reproducibility of key methodological steps, alignment between evaluation and objective, proportionality of conclusions to the reported results, and discussion of limitations. A domain was considered adequate when: (1) an explicit objective or question was stated; (2) the source, composition or sampling frame, and unit of analysis were described; (3) the main analytical steps could be reconstructed; (4) outcomes and comparators

matched the stated objective; (5) conclusions did not exceed the results; and (6) material threats to validity were discussed. A high rating required at least five adequate domains, including the objective, data/sample description, and evaluation–objective alignment, with no critical inconsistency or material reporting gap. A moderate rating required three or four adequate domains, or five adequate domains with a material reporting gap, provided that the principal result remained interpretable. A low rating was assigned when no more than two domains were adequate or when a critical inconsistency prevented interpretation of the principal result. Domains were not weighted or summed into a continuous score, and the rating was used to qualify interpretation rather than to exclude a study. A single formal risk-of-bias instrument was deliberately avoided because the review combined benchmark studies, linguistic analyses, citation-network studies, case studies, and conceptual work for which the same instrument would not be valid. The two background papers were not formally appraised; the other 69 peer-reviewed publications received a descriptive appraisal.

Because study designs, languages, models, data, and outcome measures varied widely, the evidence was synthesized thematically. Sixty-nine peer-reviewed publications were included in the main synthesis, and two papers provided conceptual background [68,69]. When a work first circulated as a preprint but was later published in a journal, the peer-reviewed version was used; this applied to the study by Sadasivan and colleagues [39]. Main conclusions were based on convergence across multiple studies rather than on a single reported metric.

3. Results

The results first describe how the evidence was selected and how the 71 included documents were distributed across years, document types, databases, topics, and quality levels. This descriptive overview provides the context needed to interpret the thematic findings in Section 4.

The complete set of 71 publications analyzed in this review comprises references [2–72]. Appendix A provides, for every citation number in that set, the dominant category, role in the synthesis, and descriptive appraisal. References [73–75] are used only to compare the present review with recent reviews and surveys, and reference [76] only to describe current publication ethics guidance; they are not counted among the 71 analyzed publications. Full bibliographic information appears only in the reference list.

3.1. Study Selection

The three formal databases yielded 1633 records after filters: 875 from Scopus, 569 from Web of Science, and 189 from IEEE Xplore. After removal of 615 duplicates, 1018 unique database records remained. Five records dated 2027 were removed before screening. Citation searching added 24 records, bringing the total screened to 1037. After title and abstract screening, 77 reports were sought, six were not retrieved, and 71 full texts were assessed and retained. All 71 were peer-reviewed publications: 69 formed the main synthesis and two were used as conceptual background. Table 1 gives the detailed distribution of records across databases and search streams, while Figure 1 summarizes the complete identification and selection process.

Table 1. Records after filtering by search stream and database.

Search Stream	Scopus	Web of Science	IEEE Xplore	Total
AI-generated text detection	421	195	117	**733**
Reference accuracy and hallucination	162	158	31	**351**

Table 1. *Cont.*

Search Stream	Scopus	Web of Science	IEEE Xplore	Total
Citation function and context	252	168	36	**456**
Citation manipulation	40	48	5	**93**
Total	**875**	**569**	**189**	**1633**

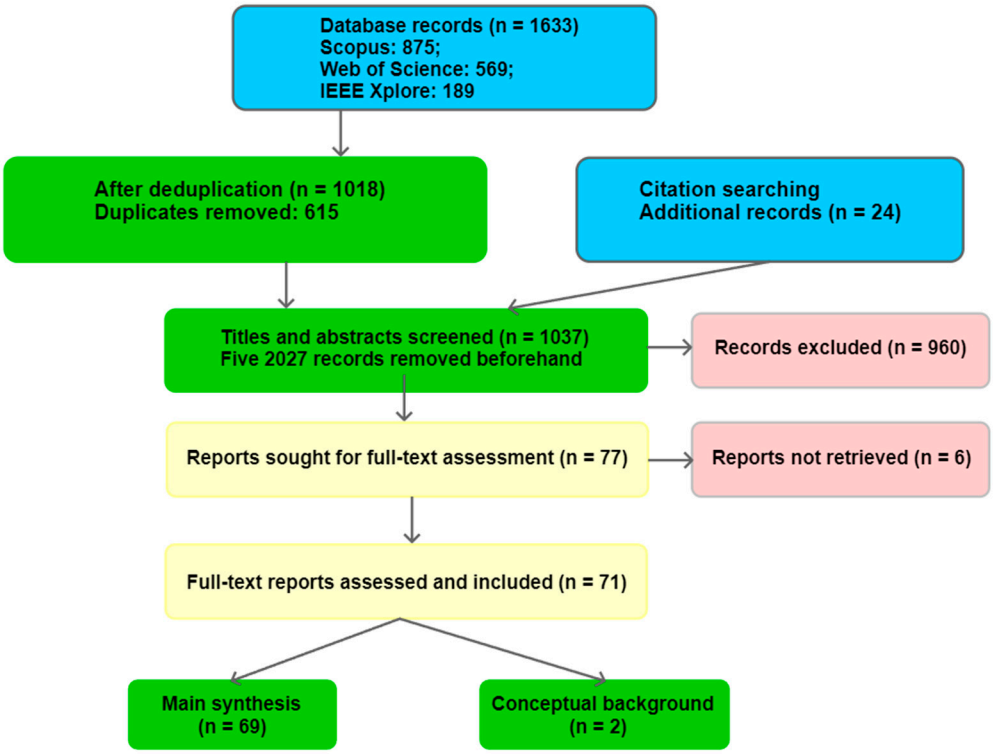


Figure 1. Flow of records through the scoping review.

3.2. Publication Years and Document Types

The evidence is recent. Fifty-eight of the 71 included publications (81.7%) were published from 2023 to 2026. The largest annual count occurred in 2024 ($n = 18$), followed by 2023 ($n = 16$), 2025 ($n = 15$), and 2026 ($n = 9$). Earlier work, beginning in 2005, focused mainly on authorship, stylometry, citation context, network anomalies, and bibliometric misconduct. These publications provide methodological background for the more recent AI-related studies. Table 2 reports the exact count for every year represented in the final set.

Table 2. Included publications by year.

Year	Number (n)	Share [%]
2005	1	1.4
2009	1	1.4
2012	1	1.4
2016	1	1.4
2017	1	1.4
2018	1	1.4
2019	1	1.4

Table 2. *Cont.*

Year	Number (*n*)	Share [%]
2020	2	2.8
2021	4	5.6
2023	16	22.5
2024	18	25.4
2025	15	21.1
2026	9	12.7
Total	**71**	**100**

Research articles were the most common document type ($n = 46$; 64.8%), followed by conference papers ($n = 19$; 26.8%). Two reviews and four other peer-reviewed analytical papers were also retained. All 71 items in the final set were peer-reviewed publications. Table 3 reports the complete distribution of document types.

Table 3. Document types in the final set.

Document Type	*n*	Share [%]
Research articles	46	64.8
Conference papers	19	26.8
Other peer-reviewed analytical papers	4	5.6
Reviews	2	2.8
Total	**71**	**100**

3.3. Discovery Routes and Thematic Distribution

Thirty-four included publications appeared in both Scopus and Web of Science. Fifteen were found only in Scopus, four only in Web of Science, three in both Scopus and IEEE Xplore, and two only in IEEE Xplore. Thirteen of the final included publications came from citation searching. These mutually exclusive routes show that the databases were complementary and that no single source recovered the entire final set. The mutually exclusive discovery routes are summarized in Table 4.

Table 4. Discovery routes for publications included in the full-text analysis.

Discovery Route	*N*	Share [%]
Scopus and Web of Science	34	47.9
Scopus only	15	21.1
Citation searching	13	18.3
Web of Science only	4	5.6
Scopus and IEEE Xplore	3	4.2
IEEE Xplore only	2	2.8
Total	**71**	**100**

The routes were mutually exclusive. Overall coverage was 52 publications in Scopus, 38 in Web of Science, five in IEEE Xplore, and 13 identified through citation searching.

The 69 publications in the main synthesis were assigned to seven categories. Forty-one addressed AI-text detection and authorship (A1–A4), while 28 addressed reference integrity, citation context, or manipulation (B, C1, and C2). The two background papers were kept outside these thematic counts. Table 5 reports the size of each category and identifies where it is discussed in the thematic synthesis.

Table 5. Thematic categories in the main synthesis.

Code and Category	*N*	Main Section
A1. AI-text detection methods and models	15	Section 4.1
A2. Detectors compared with human judgment	14	Section 4.2
A3. Stylometry and authorship verification	5	Section 4.1
A4. Robustness, evasion, bias, and provenance	7	Section 4.2
B. Reference accuracy and hallucination	9	Section 4.3
C1. Citation function and context	7	Section 4.4
C2. Citation and metric manipulation	12	Section 4.5

3.4. Descriptive Quality Appraisal

Among the 69 publications in the main synthesis, 44 (63.8%) received a high rating and 25 (36.2%) received a moderate rating. None received a low rating. The two background papers were not formally appraised because they did not form part of the analytical evidence. This distribution describes the retained studies only and should not be interpreted as evidence that the field is uniformly of high quality. It is likely influenced by database indexing, the eligibility criteria, full-text availability, and the appraisal thresholds. Records excluded before full-text synthesis were not appraised. A future review that searches ACL Anthology and PubMed directly may recover a more heterogeneous set of designs and reporting standards.

The quality categories should not be interpreted as a common performance ranking. The included studies addressed different tasks and used different units of analysis, benchmarks, and outcome measures. Appraisal was used to identify where stronger validation supported a conclusion and where interpretation required greater caution; it was not designed to estimate the quality distribution of the entire field.

4. Thematic Synthesis

The thematic synthesis connects technical detection research with work on reference accuracy, citation meaning, and bibliometric manipulation. The five subsections move from signals within the text to checks on sources and, finally, to patterns that emerge across publication and citation networks.

4.1. Methods for Identifying Text Origin

The 20 publications assigned to this section do not define detection as one task. Some classify an entire document as human- or machine-written. Others identify the source model, locate AI-generated sentences in mixed text, or compare a new submission with the established writing profile of an individual author [2–21]. These distinctions matter because a document-level label cannot show which passages were generated or how much human revision occurred.

Three broad information sources recur. Probability-based methods use token likelihood, perplexity, or changes in likelihood after perturbation. Feature-based methods use vocabulary, syntax, sentence and paragraph length, punctuation, function words, and

discourse structure. Learned approaches use neural and transformer representations, often combined with statistical or stylistic features [2–5,16]. Newer systems increasingly combine these signals instead of relying on a single indicator. Table 6 compares these approaches and summarizes their main sources of uncertainty.

Table 6. Detection approaches, evaluation setting, validation, and principal limitations.

Approach	Dataset/Domain and Language	Unit of Analysis	Generator(s) or Comparison	Evaluation and Principal Metric	Validation	Main Limitation and Studies
Probability- or likelihood-based detection	General, news, and scientific text; mainly English	Passage or document	Candidate/source language models and human text	Likelihood curvature, perplexity-related scores; AUROC or classification performance	Mostly internal; limited cross-dataset evidence	Model and domain sensitivity [12,15]
Engineered features with conventional machine learning	Academic abstracts, essays, examination answers, and related text; English and selected non-English settings	Sentence, abstract, essay, or document	ChatGPT and other generators versus human text	Accuracy, precision, recall, F1, or AUROC	Usually held-out internal data; some cross-domain tests	Feature transfer across genres, languages, and models [6–8,14,16]
Neural, transformer, and hybrid classifiers	Academic and mixed benchmark datasets; mainly English, with selected multilingual studies	Sentence, segment, or document	Multiple language models versus human text	Accuracy, F1, AUROC, and related classification measures	Predominantly internal; external validation uneven	Dependence on the training distribution and generator version [2–5,9–11,16]
Mixed-text and local detection	Human–AI collaborative and academic documents; mainly English	Sentence, local segment, or document boundary	Alternating human and model-generated passages	Local classification or boundary-detection performance, often F1	Mostly internal or limited cross-dataset testing	Segment length, editing, and frequency of authorship changes [4,5,11,13]
Stylometry and authorship verification	Articles, literature, software-education writing, and student essays; language and genre vary	Author profile, essay, document, or document pair	Writer's prior texts compared with questioned writing and/or generated text	Classification performance and linguistic-feature contrasts	Mainly domain-specific internal validation	Requires suitable prior writing; legitimate development, editing, and translation can shift style [17–21]

DetectGPT illustrates a zero-shot approach. It compares the likelihood of an observed passage with lightly perturbed versions and uses probability curvature rather than a separately trained classifier [15]. LLMDet instead builds likelihood profiles for candidate source models and attempts both human-versus-machine classification and model attribution [12]. Supervised systems include bidirectional long short-term memory (BiLSTM), convolutional neural network (CNN), transformer, and hybrid architectures. Several studies report high results within their own data and validation settings [2–11,14,16], but these values describe specific tasks rather than universal detector accuracy.

Mixed authorship has shifted attention from whole-document labels to local analysis. Gritsay and colleagues segmented academic documents and classified shorter passages [11]. Zeng and colleagues used a two-stage procedure that first detected consistent authorship spans and then classified those spans in human–AI collaborative writing [13]. Other frameworks combined token, semantic, and linguistic features to locate generated sentences and to remain useful after rewriting [4,5]. These methods better match real writing workflows, but their performance still depends on segment length and on how often authorship changes within a document.

Stylometry changes the reference point. Instead of asking whether a text resembles an average model output, authorship verification asks whether it resembles the writer's previous work. Studies used lexical diversity, syntactic complexity, punctuation, modality,

discourse markers, and neural similarity between writing samples [17–21]. The evidence shows that useful features vary by genre, language, and model version. A characteristic that separates one set of student essays may have little value in news articles or scientific abstracts. Individual profiles can add context, but they require suitable earlier writing and can be affected by legitimate development, editing, or translation.

Overall, detection methods answer different questions. Probability-based tools seek traces of the generation mechanism, supervised classifiers learn differences present in development data, and authorship verification measures departure from an individual style. A defensible academic procedure may combine document-level triage, local passage analysis, linguistic review, and process evidence, but no method provides a direct percentage of AI contribution.

4.2. Reliability and Limitations of Detection

The 21 publications synthesized here show that detector performance cannot be reduced to one percentage. Results change with the generator, text domain and length, class balance, threshold, human revision, and definition of an AI-positive case [22–40,70,71]. A high score in a controlled benchmark does not establish comparable performance on a new model or on authentic academic work.

Human reviewers and automated detectors make different errors. People may rely on fluency, tone, repetition, or factual plausibility, while software responds to statistical patterns that users cannot see [23,26–28,32,35]. In several studies, neither group was uniformly reliable. Performance varied across abstract types, essays, and medical writing, and confidence did not always track correctness. The most useful comparison is therefore not a contest for a single winner, but an analysis of when each source of judgment fails.

False positives are a major concern. Human scientific manuscripts have been labeled as AI-generated [31], and second-language academic writing can receive elevated scores because it is more predictable to a model [22,40]. This is not a minor statistical issue when the output is used in disciplinary proceedings. The relevant quantity is performance in a population similar to the one being assessed, including its languages, genres, and legitimate editing practices.

False negatives are equally important. Paraphrasing, deliberate errors, lexical substitution, translation, and sentence-level rewriting can reduce many detector scores [36–39]. Retrieval-based defenses performed better against controlled paraphrases in one study, but they require access to a very large history of previous model outputs and raise questions about privacy, retention, and infrastructure [38]. The peer-reviewed study published in Transactions on Machine Learning Research (TMLR) by Sadasivan and colleagues reached the same general conclusion: recursive paraphrasing sharply reduced several detector and watermark signals, reinforcing the need to test robustness under realistic transformations [39].

Research on hybrid text further weakens binary labels. Tools performed differently when a model only edited human writing, when humans revised model output, and when documents contained alternating passages [28,32,33]. A two-class decision erases these distinctions. Reporting should therefore specify the unit of analysis, model version, threshold, class distribution, and definition of assistance. Without this information, an accuracy or F1 value is difficult to interpret.

Four levels of evidence should therefore be distinguished. First, a publication may show only that a method has been investigated. Second, it may show good performance under a particular benchmark and threshold. Third, it may provide external evidence that the method generalizes to authentic scholarly documents from other models, disciplines, languages, or time periods. Fourth, it may demonstrate that the evidence is reliable and

proportionate enough to support an editorial or institutional decision. The reviewed literature commonly supports the first two levels, but evidence for broad generalization and decision-level use remains limited.

Process traces offer a different form of evidence. Large-scale distributional methods can estimate changes across collections of peer reviews without accusing individual reviewers [70]. Revision identifiers stored in DOCX files can indicate progressive writing, pasting, or later editing, although they do not prove misconduct by themselves [71]. Draft histories, source notes, data, code, and the author's explanation are more directly related to the production process than a style score. Their use should remain proportionate and respect privacy [22–40,70,71].

4.3. Reference Hallucination and Claim–Source Agreement

Reference integrity requires several separate checks. A record may be entirely invented, may combine metadata from different publications, may point to an outdated version, or may exist but fail to support the claim. The nine publications in this section show why a complete-looking reference list cannot be accepted without verification [41–48,72].

Experiments with chatbots found frequent errors in existence, DOI, authorship, title, journal, year, and relevance [41–43]. Civelekler and Citirik evaluated ChatGPT (GPT-3.5 free version), Copilot (Microsoft GitHub integrated version), DeepSeek (publicly accessible medical version), and Gemini (free web-based version) using 35 standardized clinical paragraphs in the specialized domain of ocular inflammation and uveal disease and checked the generated citations against PubMed. Under the study's citation-accuracy criterion, DeepSeek achieved the highest proportion (65.7%; 23/35), followed by Copilot (42.9%; 15/35) and ChatGPT (37.1%; 13/35), while Gemini achieved 5.7% (2/35) [41]. These figures describe the tested systems, prompts, domain, sample, and PubMed-based procedure; they are not general estimates of model factuality or claim-level support. In another study, which tested ChatGPT-4.0, only 10% of references generated for general prompts could be confirmed, although narrower prompts produced better results [42]. A six-field Reference Hallucination Score applied to 500 references generated by six medical AI chatbots found relevance errors in 61.6%, along with substantial metadata errors [43]. These studies used different prompts, systems, definitions, and verification procedures, so their percentages should not be compared as a common ranking. Table 7 brings together the successive checks for reference integrity, citation function, and citation context.

The hardest step is often semantic rather than bibliographic. Automated work on biomedical citation integrity found that retrieving the relevant evidence passage from the full text limited classification performance [46]. Systems may identify a related paper but still miss whether a sentence overstates, contradicts, or only partially reflects the source [44–46]. Earlier citation-network case studies show how selective repetition and the loss of qualifications can turn tentative or anecdotal evidence into apparent authority [47,48].

Version matching also requires care. PreprintResolver compared titles, authors, and identifiers across several scholarly databases to identify journal versions of arXiv papers [72]. The method was useful for supervised queries but could return multiple or partly incorrect candidates. Automated replacement should therefore end with human verification.

A defensible workflow begins with resolving the DOI or title, matching all bibliographic fields, and then reading the relevant full-text passage. Complex sentences should be divided into atomic claims, each linked to evidence. If support cannot be located, topical similarity alone should not be treated as confirmation. Automated systems can prioritize checks and retrieve candidate passages, but current evidence does not support fully automatic adjudication.

Table 7. Reference-integrity checks and analysis of citation function and context.

Analytical Scope	Level or Object of Review	Verification Question, Application, and Typical Problem	Studies
Reference integrity	Document existence	Does the title or DOI resolve to a real publication? Typical problem: an entirely fabricated reference.	[41–43]
Reference integrity	Metadata agreement	Do the authors, year, journal, pages, and DOI match? Typical problem: a record assembled from several publications.	[41,43,72]
Reference integrity	Claim–source agreement	Does the source support the specific claim? Typical problem: overstatement, an omitted qualification, or an irrelevant citation.	[44–48]
Reference integrity	Document version	Is the appropriate and current version cited? Typical problem: retaining a preprint citation after final publication.	[72]
Reference integrity	Citation history	Has a claim acquired apparent authority through repetition? Verification requires returning to the primary evidence.	[45,47,48]
Citation function	Sentence	Does the claim require a source? Application: identifying potential gaps in the argument.	[53]
Citation function	Local context	What function and stance does the citation express? Application: distinguishing background, support, comparison, and criticism.	[50,52,54]
Citation function	Article section	Does the meaning depend on citation location? Application: deciding when the full text must be examined.	[49,51,53]
Citation function	Full publication	Which passage in the source supports the claim? Application: direct claim–evidence checking.	[49,51]
Citation function	Time and network	How does a paper's role change, and does an error spread to later work? Application: monitoring the propagation of unreliability.	[49,55]

4.4. Citation Function and Context

Citation counts show how often a publication was mentioned, not why it was mentioned. The seven studies in this section examine the need for a citation, its rhetorical role, local context, and change in meaning across later publications [49–55]. These layers are necessary both for assessing argument quality and for interpreting bibliometric patterns.

Citation-worthiness models identify sentences likely to require a reference [53]. Their performance changes across disciplines and data sources, and they cannot determine which source is best or whether a statement counts as common knowledge. Citation-function models classify roles such as background, method, use, comparison, support, or criticism [50,52,54,55]. Full-text exploration improves these tasks because section location and surrounding argument often change the meaning of an in-text reference [51].

Research on unreliability propagation adds a further requirement. A citation to a corrected, retracted, or weak paper is not automatically wrong, because the author may discuss it critically. Conversely, an apparently neutral citation may repeat an unsupported claim. Reliable analysis should therefore combine citation status with the local sentence, nearby context, section, and—when available—the evidence passage in the cited work [49].

These studies support a layered interpretation. First determine whether a sentence needs a source. Then identify which claim the citation attaches to, classify its function, and evaluate claim–source agreement. Only after these steps should citation counts or network position be interpreted as evidence of influence.

4.5. Citation and Bibliometric Manipulation

Manipulation includes practices that operate at different levels: the manuscript, peer review, journal, metadata infrastructure, and wider citation network. The 12 main-synthesis studies and one background paper describe coercion, citation exchange, stacking, paper mills, purchased citations, and metadata discrepancies [56–68]. No single signal identifies all of these behaviors.

Coercive citation occurs when an editor or reviewer pressures an author to add references that are not needed for the argument. Survey and conceptual work shows that authors may comply because editorial power is asymmetric and refusal may appear to endanger acceptance [60,67,68]. A count-based system cannot distinguish a justified recommendation from coercion. Relevant evidence may include review reports, the relationship between recommended citations and the reviewer, and whether the request explains the scientific relevance of each source.

Strategic self-citation and citation stacking can arise from incentives tied to thresholds and rankings [63–66]. Self-citation is not inherently improper: it can represent continuity in a research program. Concern increases when citation concentration, reciprocity, or timing changes abruptly without a corresponding intellectual explanation. Country-level and journal-level studies show that metric-based evaluation can alter behavior across entire communities [64–66].

Network models seek these unusual patterns. CIDRE identifies groups of journals with citation relationships that depart from expected patterns and separates within-group concentration from external citation behavior [62]. Other methods disturb node embeddings or model directed citation networks to detect manipulated clusters [56,61]. Their value lies in prioritizing cases, not proving intent. Small fields, special issues, shared methods, and legitimate collaboration can also create dense networks.

Recent studies show how generative AI and metadata systems can lower the cost of manipulation. Citation-mill research demonstrated that fabricated authors and AI-generated preprints can be used to create mutually citing records and that citations can be purchased [58]. A case study placed an AI-generated paper with targeted references on an academic networking platform, where it could affect visible metrics [57]. Besançon and colleagues found references present in DOI metadata but absent from the published files; these "sneaked references" altered citation counts and could be detected only by comparing publication files with metadata and downstream platforms [59]. Table 8 summarizes the main forms of manipulation, their warning signals, and the verification required before drawing conclusions.

Table 8. Forms of citation manipulation, warning signals, and required verification.

Practice	Level	Warning Signal	Required Verification	Studies
Coercive citation	Review and editorial process	Repeated requests to add weakly related sources	Review reports, scholarly relevance, and conflicts of interest	[60,67,68]
Strategic self-citation or citation stacking	Author or journal	Sudden concentration of citations within an unusual time window	Field norms and the actual intellectual relationship	[63–66]

Table 8. *Cont.*

Practice	Level	Warning Signal	Required Verification	Studies
Citation cartel or exchange	Journal network	Excessive reciprocity and closed citation flows	Network structure, timing, journal scope, and editorial links	[61,62]
Paper mill or citation mill	External market and platform	Repetitive content, fabricated identities, or mass linking	Identity, full text, provenance, and available transaction evidence	[56,58]
Metadata manipulation	Publisher and indexing infrastructure	References present in metadata but absent from the article	Comparison of PDF, HTML, Crossref, and platform records	[59]
AI-generated content used to inflate metrics	Repository and citation network	Weak content combined with mass or reciprocal citation	Content, chronology, network relations, and evidence of intent	[57,58]

Automated detection is strongest when it can point to a specific, checkable anomaly: an unmatched DOI, a reference missing from the PDF, a sudden reciprocal citation burst, or a tightly connected group. It is weakest when intent is the deciding factor.

5. Discussion

Section 4 established what the reviewed methods can and cannot measure. The Discussion therefore focuses on how those outputs should be used in editorial and institutional decisions, where the consequences of a false accusation or a missed integrity problem are materially different from benchmark error.

5.1. Position Relative to Previous Reviews and Integration of the Domains

Recent reviews have examined AI-text detection methods, datasets, applications, and technical limitations [73,74], while a broader taxonomy of AI hallucination has included bibliographic errors among several forms of distorted information [75]. These contributions are important, but they do not jointly map text-origin assessment, reference existence and metadata, claim–source agreement, citation function, and bibliometric manipulation within one audit pathway. Table 9 makes the difference in scope and review design explicit.

The novelty of the present review is therefore the integration of evidence that has usually been treated separately and the explicit separation of evidentiary levels within a five-layer audit. This integration is methodological rather than causal: a text-origin alert does not predict a citation defect, and a citation anomaly does not establish AI use. The advance lies in routing each concern to the evidence capable of answering it and preventing one type of signal from being used as a substitute for another. The framework remains a conceptual synthesis to be tested prospectively.

Text-origin assessment and citation integrity answer different questions and carry different error costs. A false-positive detector result can unfairly implicate an author, whereas an undetected fabricated or distorted reference can leave an unsupported claim in the literature. The two assessments should therefore remain separate during evidence collection and be combined only when a documented decision requires both [41–48,72]. Text-origin tools examine statistical or stylistic signals and process traces [2–40,70,71], whereas citation auditing relies on metadata, full-text evidence, rhetorical function, and network structure [44,46,49,56,59,62]. A decision record should state which question was examined, what evidence was available, and what uncertainty remains.

Table 9. Comparison with recent reviews and surveys.

Review	Design, Sources, Period, and n	AI-Text Detection	Reference Hallucination	Citation Function	Bibliometric Manipulation	Integrated Audit
Wu et al. (2025) [73]	Technical survey; database search, period, and included-study flow not reported.	Yes	Incidental examples only	No	No	No
Kehkashan et al. (2025) [74]	Broad review; database search, period, and included-study flow not reported.	Yes	No dedicated synthesis	No	No	No
Sun et al. (2024) [75]	Taxonomy and content analysis of 243 distorted-information cases; not a systematic review.	No dedicated detector review	Included within a broader hallucination taxonomy	No	No	No
Present review	Scoping review; Scopus, Web of Science, IEEE Xplore, and citation searching; searches in 2026; evidence published in 2005–2026; $n = 71$	Yes	Yes	Yes	Yes	Yes—conceptual, not yet validated

5.2. Why AI Detection Is Not Proof of Authorship

Detector output should be interpreted as conditional evidence, not as proof of identity or intent. The reviewed studies show sensitivity to generator version, discipline, language, text length, editing, threshold choice, and deliberate evasion [22–40,69–71]. The relevant practical question is therefore whether the tool was validated on a population and task comparable to the manuscript under review, not whether the interface displays a high percentage. In terms of the four levels distinguished in Section 4.2, most detector evidence reaches the second level, whereas editorial and institutional use require the third and fourth.

The evidentiary standard should rise with the consequence of the decision. In a low-stakes editorial check, a detector alert may justify asking for clarification. In an allegation of misconduct, the same alert is insufficient without process evidence such as drafts, version history, research notes, data or code, and an opportunity for the author to explain the writing workflow [38–40,69–71]. Table 10 distinguishes conclusions that are strongly tied to changing models and datasets from methodological principles that are more stable.

Table 10 separates numerical conclusions that can become obsolete quickly from methodological principles that are more stable. Detector rankings, thresholds, and subgroup error rates should be treated as time-stamped evidence tied to specified versions and test data. By contrast, the need to verify references directly, distinguish bibliographic correctness from claim support, and avoid inferring intent from an anomaly remains valid even as particular tools change.

Table 10. Dependence of the review's conclusions on changing models, datasets, and systems.

Conclusion	Model/Version Dependence	Interpretation and Update Need
Exact detector accuracy or ranking	High	Specific to the tested model, detector version, prompt or decoding setting, domain, class balance, and threshold; revalidate after deployment or version changes.
Cross-domain and cross-language performance	High	Dataset- and population-dependent; external validation is required for each intended scholarly setting.
Editing and paraphrasing can reduce detectability	Moderate to high	The mechanism recurs across studies, but the size of the effect changes with the attack, generator, detector, and text length; repeat robustness tests.
Reference existence and metadata require direct verification	Low at the methodological level	The principle is stable, although databases, resolver coverage, and model behavior change.
Citation existence does not establish claim–source agreement	Low at the methodological level	The distinction is stable; implementation still requires access to the cited full text and subject expertise.
A network anomaly does not establish intent	Low at the methodological level	The principle is stable, but thresholds and expected network structure remain field- and platform-dependent.

5.3. A Proposed Conceptual Multilayer Publication-Integrity Audit

On the basis of the scoping synthesis, the authors propose a proportionate conceptual audit organized in five layers: documented writing process and disclosure of assistance; targeted assessment of text-origin signals [2–40,70,71]; verification of reference existence and metadata; claim-to-source checking in the cited full text, together with interpretation of citation function and context; and examination of temporal, network, or metadata patterns that may indicate manipulation [44,46,49–67]. This integrated workflow has not been empirically validated as a whole. It should be understood as a structured synthesis and a hypothesis for prospective evaluation, not as a diagnostic instrument with established accuracy. The layers should be activated according to the concern rather than applied automatically to every paper.

Each layer should retain a separate result and evidentiary basis. A missing DOI may be a correctable metadata error, claim–source disagreement requires subject expertise, and a suspected citation cartel requires longitudinal network evidence. Keeping these outcomes separate prevents one ambiguous signal from becoming a general verdict on the manuscript.

Automation is most useful for triage and documentation: it can flag a nonexistent identifier, retrieve candidate evidence passages, identify an uncited claim, or prioritize an unusual network for examination. Final decisions should be made by a qualified person who can inspect the full text, explain the basis of the concern, and consider the author, reviewer, editor, or publisher response.

5.4. Implications for Authors, Journals, and Institutions

Evidence-based implications. Across the reviewed studies, detector outputs are conditional on the test setting, bibliographic existence is distinct from claim support, and unusual citation patterns do not by themselves establish intent. These recurring findings support

direct verification, preservation of the underlying evidence, and proportionate human review before any consequential decision [22–48,56–71].

Recommendations proposed in this review. For authors, the starting point is the publication ethics consensus that AI tools cannot be listed as authors because they cannot take responsibility for the submitted work, and that authors remain fully responsible for all content, including passages drafted, translated, or edited with AI assistance [76]. Authors may use AI tools for translation or language and style editing if that is permitted by the target journal. The author should compare the edited text with the source; verify that technical meaning, numerical values, uncertainty, and disciplinary terminology have not changed; and independently check every reference. Reference verification should also reflect disciplinary practice: clinical claims require checking current source texts and guidance, computational claims require matching methods with code or data when available, and interpretive claims require attention to citation context. Confidential or unpublished material should not be entered into an external service without authorization. The tool, purpose, and extent of assistance should be documented and disclosed whenever the journal policy requires it. Language polishing must not be used to generate or conceal data, analyses, results, or references. These recommendations are derived from the evidence mapped above but have not been tested as a complete intervention. They are particularly important for authors writing in an additional language, because predictable second-language patterns and extensive editing can both affect detector scores [22,30,33,36–40].

Journals should state separately what is permitted for translation, grammar and style editing, drafting, data analysis, image generation, and reference discovery. Reviewer forms should require a scientific justification for requested citations and disclosure of a reviewer's relevant self-citations. A detector alert should trigger a documented question rather than automatic rejection; editors should first establish whether the tool was validated for the submission's language, genre, and discipline [46,59,60,67,68].

Institutional integrity officers should preserve the tool name, version, threshold, input text, and specific alert; examine drafts, version histories, notes, data, and code where proportionate; and obtain disciplinary review before drawing a conclusion. Authors should receive the evidence underlying the concern and a meaningful opportunity to respond. Bibliographic databases and metric platforms should expose corrections, retractions, citation provenance, and discrepancies between publication files and deposited metadata [58,59,62,64,65].

5.5. Research Gaps

First, evaluation should move from direct model outputs toward realistic scholarly workflows that include translation, language editing, partial assistance, and human revision. Multilingual and cross-disciplinary benchmarks should preregister generator versions, thresholds, class distributions, and subgroup analyses, and should report false-positive rates for writers using English as an additional language [22–40,69–71]. External validation should be repeated when major generator or detector versions change.

Second, claim–source validation needs shared benchmarks with independently annotated claim boundaries, degrees of support, citation functions, secondary citations, and changes in meaning over time. Systems should return an evidence passage and calibrated uncertainty so that retrieval failure can be separated from reasoning failure. Annotation studies should use at least two independent coders and transparent adjudication [44–46,49–55].

Third, citation-manipulation research needs longitudinal labels that distinguish confirmed misconduct, corrected error, legitimate specialty structure, and unresolved anomaly.

Evaluations should combine full texts, versioned metadata, editorial roles, review histories, and temporal networks, and should report not only detection accuracy but also the burden of false alerts and the effect of appeal or correction procedures [56–62].

Fourth, prospective studies should test complete audit workflows in real editorial and institutional settings. Useful outcomes include confirmed corrections, time required from experts and authors, missed problems, appeal outcomes, privacy costs, and harm caused by false accusations. This would show whether a technically strong method improves decisions rather than merely producing a high benchmark score [46,49,56,61,62].

5.6. Limitations of This Review

The formal search covered Scopus, Web of Science, and IEEE Xplore. These databases were selected for complementary interdisciplinary and technical coverage, but they were not sufficient to establish exhaustive retrieval. ACM Digital Library could not be formally included because the available Basic Edition blocked the required filtering and batch export. ACL Anthology and PubMed were not searched separately, and the review did not measure how many publications indexed only in those sources were recovered through citation searching or database overlap. The review may therefore underrepresent NLP conference literature and medical AI studies. In addition, the computational term block in stream C1 preferentially retrieved automated citation-function and citation-context research and may have missed noncomputational scholarship on rhetorical roles and citation practice. A future update should search ACL Anthology and PubMed directly, broaden the noncomputational citation-analysis strategy, and rerun deduplication and screening.

Screening and quality appraisal were performed primarily by one reviewer with computational support and targeted cross-checks by a second reviewer, not by two independent reviewers across the complete set. The targeted checks were not a prespecified, blinded duplicate sample; a post hoc agreement coefficient or disagreement rate would therefore be misleading and was not calculated. Although questioned records were reassessed against the stated criteria and discussed, reviewer subjectivity may still have influenced eligibility, categorization, and appraisal. This is a material limitation. A future update should use independent duplicate title-and-abstract screening, full-text assessment, and appraisal with prospectively defined adjudication. Six potentially relevant reports were not assessed because their full texts could not be obtained. The absence of low appraisal ratings applies only to the retained publications and may reflect indexing, eligibility, and full-text availability as well as study reporting. It should not be generalized to the entire field. The review also combines heterogeneous tasks, data types, and outcome measures, so accuracy, F1, recall, and AUROC values should not be read as a common ranking.

Finally, both generators and detectors change quickly. Exact accuracy estimates, rankings, thresholds, and subgroup error rates are highly dependent on model, detector, dataset, language, and version. They describe the technology and test conditions at the time of a study and require renewed validation after material changes. Indexing rules, repository controls, and metric calculations can also change. Table 10 makes this distinction explicit so that transient numerical findings are not mistaken for permanent properties of current tools.

6. Conclusions

This review shows that text-origin detection, reference verification, citation-context analysis, and bibliometric anomaly detection are complementary but distinct forms of evidence. Their outputs should be interpreted in relation to a defined decision, population, and consequence rather than merged into a single trust score.

The practical contribution is a proposed conceptual audit pathway: use automated methods to identify a specific question, preserve the underlying evidence, verify references

and claims directly, and escalate only when the concern remains after human review. The pathway is derived from the mapped evidence but has not yet been prospectively validated as an integrated workflow. Legitimate AI assistance for translation or language editing should be assessed under the journal's policy and the author's documented workflow, not inferred from fluency or a detector percentage.

Algorithms that detect statistical signs of AI-assisted writing and algorithms that detect anomalous citation, network, or metadata patterns address different phenomena. Each can identify cases for examination, but neither establishes intent or misconduct. Similar signals may have legitimate explanations, so every consequential decision should rely on checkable evidence, qualified human review, and an opportunity for the author to respond [2–40,56–71].

The main research priority is therefore not another universal detector score. It is prospective validation of complete audit workflows, including multilingual subgroup error rates, explanation quality, reviewer workload, privacy safeguards, corrections, appeals, and harms from false accusations.

Author Contributions: Conceptualization, K.W. and D.K.; methodology, K.W. and D.K.; software, D.K.; validation, K.W. and D.K.; formal analysis, D.K.; investigation, D.K.; resources, D.K.; data curation, D.K.; writing—original draft preparation, K.W. and D.K.; writing—review and editing, K.W. and D.K.; visualization, D.K.; supervision, K.W.; project administration, K.W. and D.K.; funding acquisition, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The study reviewed published literature and did not involve human participants or identifiable personal data.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this review are included in this article and its Appendices A and B.

Acknowledgments: During the preparation of this manuscript, the authors used Gemini 3.1 Pro (Google) for the purposes of checking textual consistency and typographical errors. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACL	Association for Computational Linguistics
ACM	Association for Computing Machinery
AI	artificial intelligence
AUROC	area under the receiver operating characteristic curve
BiLSTM	bidirectional long short-term memory
CNN	convolutional neural network
DOCX	Microsoft Word document format
DOI	digital object identifier
HTML	HyperText Markup Language
IEEE	Institute of Electrical and Electronics Engineers
LLM	large language model
NLI	natural language inference
NLP	natural language processing
PDF	Portable Document Format
PRISMA-ScR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
TMLR	Transactions on Machine Learning Research

Appendix A. Characteristics of the 71 Analyzed Publications

Table A1 links all 71 analyzed publications to their reference-list numbers and reports only the dominant category, synthesis role, and descriptive appraisal. Full bibliographic records are not repeated in the appendix.

Table A1. Coding summary for all included publications.

Citation	Dominant Category	Role in Synthesis	Descriptive Appraisal
[2]	A1	Main study	High
[3]	A1	Main study	Moderate
[4]	A1	Main study	High
[5]	A1	Main study	High
[6]	A1	Main study	Moderate
[7]	A1	Main study	High
[8]	A1	Main study	High
[9]	A1	Main study	Moderate
[10]	A1	Main study	Moderate
[11]	A1	Main study	Moderate
[12]	A1	Main study	High
[13]	A1	Main study	High
[14]	A1	Main study	High
[15]	A1	Main study	High
[16]	A1	Main study	High
[17]	A3	Supporting study	Moderate
[18]	A3	Supporting study	High
[19]	A3	Supporting study	High
[20]	A3	Supporting study	High
[21]	A3	Supporting study	High
[22]	A2	Main study	High
[23]	A2	Main study	High
[24]	A2	Main study	High
[25]	A2	Main study	High
[26]	A2	Main study	Moderate
[27]	A2	Main study	High
[28]	A2	Main study	High
[29]	A2	Main study	High
[30]	A2	Main study	High
[31]	A2	Main study	High
[32]	A2	Main study	High
[33]	A2	Main study	High
[34]	A2	Main study	High
[35]	A2	Main study	Moderate
[36]	A4	Main study	Moderate
[37]	A4	Main study	High
[38]	A4	Main study	Moderate
[39]	A4	Main study	Moderate
[40]	A4	Main study	Moderate
[41]	B	Main study	High
[42]	B	Main study	High
[43]	B	Main study	High
[44]	B	Main study	High
[45]	B	Main study	High
[46]	B	Main study	High
[47]	B	Supporting study	Moderate
[48]	B	Main study	High

Table A1. *Cont.*

Citation	Dominant Category	Role in Synthesis	Descriptive Appraisal
[49]	C1	Supporting study	High
[50]	C1	Supporting study	Moderate
[51]	C1	Supporting study	Moderate
[52]	C1	Supporting study	Moderate
[53]	C1	Supporting study	High
[54]	C1	Supporting study	High
[55]	C1	Supporting study	Moderate
[56]	C2	Main study	High
[57]	C2	Main study	Moderate
[58]	C2	Main study	High
[59]	C2	Main study	Moderate
[60]	C2	Main study	High
[61]	C2	Main study	Moderate
[62]	C2	Main study	High
[63]	C2	Main study	Moderate
[64]	C2	Main study	Moderate
[65]	C2	Main study	High
[66]	C2	Main study	High
[67]	C2	Main study	Moderate
[68]	Background	Conceptual background	Not formally appraised—background text
[69]	Background	Conceptual background	Not formally appraised—background text
[70]	A4	Main study	Moderate
[71]	A4	Supporting study	High
[72]	B	Supporting study	Moderate

Appendix B

This appendix reproduces the strings retained in the search log. Scopus used TITLE-ABS-KEY, Web of Science used TS=, and IEEE Xplore used All Metadata/Command Search. Wildcards were retained only when supported by the platform.

Appendix B.1. Scopus—31 August 2026

Filters: English; Article, Conference Paper, and Review. Stream A: 2019–2026; streams B, C1, and C2: from the beginning of database coverage through the search date.

A. A total of 1347 broad-search results; 474 after adding the academic-context block; and 421 after filters.

("AI-generated text" OR "machine-generated text" OR "LLM-generated text" OR "large language model generated text" OR "ChatGPT-generated text" OR "generative AI text") AND (detect* OR identif* OR classif* OR authorship OR attribution OR stylometr* OR watermark* OR provenance) AND (scientific OR scholarly OR academic OR manuscript* OR publication* OR essay* OR student*)

B. A total of 220 before filters; 162 after filters.

("citation verification" OR "citation validity" OR "citation correctness" OR "citation accuracy" OR "citation consistency" OR "citation quality" OR "citation hallucination*" OR "fabricated citation*" OR "fake citation*" OR "ghost citation*" OR miscitation OR "mis-citation" OR "reference hallucination*") AND (scientific OR scholarly OR academic OR publication* OR manuscript* OR research)

C1. A total of 300 before filters; 252 after filters.

("citation intent" OR "citation intents" OR "citation function" OR "citation functions" OR "citation context" OR "citation contexts" OR "citation classification" OR "citation role" OR "citation roles" OR "citation entailment") AND ("natural language processing" OR

NLP OR "natural language inference" OR NLI OR "machine learning" OR "deep learning" OR "large language model*" OR LLM* OR automatic OR computational)

C2. A total of 266 before filters; 40 after filters. An earlier broad query returned 1394 results and was rejected as insufficiently specific.

("citation manipulation" OR "citation manipulations" OR "citation cartel" OR "citation cartels" OR "citation coercion" OR "coercive citation" OR "coercive citations" OR "citation stacking" OR "citation mill" OR "citation mills" OR "citation farm" OR "citation farms" OR "citation ring" OR "citation rings" OR "bibliometric manipulation" OR "citation gaming" OR "citation fraud") AND ("research integrity" OR "publication ethics" OR "academic publishing" OR "scholarly publishing" OR bibliometric* OR scientometric*)

Appendix B.2. Web of Science Core Collection—31 August 2026

Filters: English; Article, Review, and Proceedings Paper; Early Access retained. Stream A: 2019–2026.

A. A total of 202 before filters; 195 after filters.

TS=(("AI-generated text" OR "machine-generated text" OR "LLM-generated text" OR "large language model generated text" OR "ChatGPT-generated text" OR "generative AI text") AND (detect* OR identif* OR classif* OR authorship OR attribution OR stylometr* OR watermark* OR provenance) AND (scientific OR scholarly OR academic OR manuscript* OR publication* OR essay* OR student*))

B. A total of 171 before filters; 158 after filters.

TS=(("citation verification" OR "citation validity" OR "citation correctness" OR "citation accuracy" OR "citation consistency" OR "citation quality" OR "citation hallucination*" OR "fabricated citation*" OR "fake citation*" OR "ghost citation*" OR miscitation OR "miscitation" OR "reference hallucination*") AND (scientific OR scholarly OR academic OR publication* OR manuscript* OR research))

C1. A total of 168 before and after filters.

TS=(("citation intent" OR "citation intents" OR "citation function" OR "citation functions" OR "citation context" OR "citation contexts" OR "citation classification" OR "citation role" OR "citation roles" OR "citation entailment") AND ("natural language processing" OR NLP OR "natural language inference" OR NLI OR "machine learning" OR "deep learning" OR "large language model*" OR LLM* OR automatic OR computational))

C2. A total of 58 before filters; 48 after filters.

TS=(("citation manipulation" OR "citation manipulations" OR "citation cartel" OR "citation cartels" OR "citation coercion" OR "coercive citation" OR "coercive citations" OR "citation stacking" OR "citation mill" OR "citation mills" OR "citation farm" OR "citation farms" OR "citation ring" OR "citation rings" OR "bibliometric manipulation" OR "citation gaming" OR "citation fraud") AND ("research integrity" OR "publication ethics" OR "academic publishing" OR "scholarly publishing" OR bibliometric* OR scientometric*))

Appendix B.3. IEEE Xplore—2 September 2026

Filters: Journals, Conferences, and Early Access Articles; English when available. Stream A: 2019–2026. Unsupported wildcards were removed in IEEE Xplore.

A. A total of 118 before filters; 117 after filters.

("AI-generated text" OR "machine-generated text" OR "LLM-generated text" OR "large language model generated text" OR "ChatGPT-generated text" OR "generative AI text") AND (detect OR identif OR classif OR authorship OR attribution OR stylometr OR watermark OR provenance) AND (scientific OR scholarly OR academic OR manuscript OR publication OR essay OR student)

B. A total of 31 before and after filters.

("citation verification" OR "citation validity" OR "citation correctness" OR "citation accuracy" OR "citation consistency" OR "citation quality" OR "citation hallucination" OR "fabricated citation" OR "fake citation" OR "ghost citation" OR miscitation OR "mis-citation" OR "reference hallucination") AND (scientific OR scholarly OR academic OR publication OR manuscript OR research)

C1. A total of 36 before and after filters.

("citation intent" OR "citation intents" OR "citation function" OR "citation functions" OR "citation context" OR "citation contexts" OR "citation classification" OR "citation role" OR "citation roles" OR "citation entailment") AND ("natural language processing" OR NLP OR "natural language inference" OR NLI OR "machine learning" OR "deep learning" OR "large language model" OR LLM OR automatic OR computational)

C2—primary. A total of 0 results.

("citation manipulation" OR "citation manipulations" OR "citation cartel" OR "citation cartels" OR "citation coercion" OR "coercive citation" OR "coercive citations" OR "citation stacking" OR "citation mill" OR "citation mills" OR "citation farm" OR "citation farms" OR "citation ring" OR "citation rings" OR "bibliometric manipulation" OR "citation gaming" OR "citation fraud") AND ("research integrity" OR "publication ethics" OR "academic publishing" OR "scholarly publishing" OR bibliometric OR scientometric)

C2—sensitivity search. A total of 8 results; 5 after filters and exported.

"citation manipulation" OR "citation manipulations" OR "citation cartel" OR "citation cartels" OR "citation coercion" OR "coercive citation" OR "coercive citations" OR "citation stacking" OR "citation mill" OR "citation mills" OR "citation farm" OR "citation farms" OR "citation ring" OR "citation rings" OR "bibliometric manipulation" OR "citation gaming" OR "citation fraud"

Appendix B.4. Preprints and Sources Not Formally Included

Scopus displayed 155 preprints from 2017 to 2026 in a separate collection. They were not merged with peer-reviewed exports. The ACM Digital Library Basic Edition returned 114 records after refinement of stream A, but the lack of required filters and batch export prevented formal inclusion. No separate searches were conducted in ACL Anthology or PubMed.

References

1. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
2. Wang, J.; Li, Y.; You, T.; Liu, L.; Wang, P. DSAD: A Dual-Stream Aligned Detection Framework for LLM-Generated Exam Answers. *Expert Syst. Appl.* **2026**, *313*, 131602. [CrossRef]
3. Hazim, L.R.; Ata, O. HQML-NLP: A Hybrid Quantum Machine Learning Framework for Scholarly AI-Text Detection. *Appl. Soft Comput.* **2026**, *191*, 114634. [CrossRef]
4. Wang, Z.; Wu, L.; An, W.; Zhang, H.; Wang, Y. Multi-Level Style Preference Optimization: An Adaptive Detection Framework for Human-Machine Hybrid Text. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 33746–33754. [CrossRef]
5. Xu, W.; Tao, Z.; Chen, R.; Xi, D.; Li, Z. Unveiling Large Language Models Generated Texts: A Multi-Level Fine-Grained Detection Framework. *Decis. Support Syst.* **2026**, *205*, 114655. [CrossRef]
6. Sharimbayev, B.; Kazin, A.; Kadyrov, S. Multi-Class Detection of Humanized AI Text Using Machine Learning and Transformer Models. In *2025 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2025; pp. 127–131.
7. Popescu-Apreutesei, L.-E.; Iosupescu, M.-S.; Necula, S.C.; Păvăloaia, V.-D. Upholding Academic Integrity amidst Advanced Language Models: Evaluating BiLSTM Networks with GloVe Embeddings for Detecting AI-Generated Scientific Abstracts. *Comput. Mater. Contin.* **2025**, *84*, 2605–2644. [CrossRef]

8. Nikolaev, K. Development of a Neural Network Model for Recognizing Russian-Language Generated Texts. In *2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024; pp. 396–400. [CrossRef]
9. Noheria, A.; Ratna, S.; Ram, P.K. Enhancing AI Text Detection: Embracing Bi-LSTM and Transformer Block Fusion. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*; IEEE: New York, NY, USA, 2024; pp. 1–6.
10. Boeva, G.; Gritsai, G.; Grabovoy, A. Team ap-team at PAN: LLM Adapters for Various Datasets. In *Proceedings of the Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9–12 September 2024*; CEUR-WS: Aachen, Germany, 2024; Volume 3740, pp. 2527–2535. Available online: https://ceur-ws.org/Vol-3740/ (accessed on 15 October 2026).
11. Gritsay, G.M.; Grabovoy, A.V.; Kildyakov, A.S.; Chekhovich, Y.V. Artificially Generated Text Fragments Search in Academic Documents. *Dokl. Math.* **2023**, *108*, S434–S442. [CrossRef]
12. Wu, K.; Pang, L.; Shen, H.; Cheng, X.; Chua, T.-S. LLMDet: A Third Party Large Language Models Generated Text Detection Tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics (ACL): Singapore, 2023; pp. 2113–2133.
13. Zeng, Z.; Liu, S.; Sha, L.; Li, Z.; Yang, K.; Liu, S.; Gašević, D.; Chen, G. *Detecting AI-Generated Sentences in Human-AI Collaborative Hybrid Texts: Challenges, Strategies, and Insights*; IJCAI Organization: Marina del Rey, CA, USA, 2024; pp. 7545–7553.
14. Desaire, H.; Chua, A.E.; Isom, M.; Jarosova, R.; Hua, D. Distinguishing Academic Science Writing from Humans or ChatGPT with over 99% Accuracy Using Off-the-Shelf Machine Learning Tools. *Cell Rep. Phys. Sci.* **2023**, *4*, 101426. [CrossRef] [PubMed]
15. Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C.D.; Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023), Honolulu, HI, USA, 23–29 July 2023*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 24950–24962. Available online: https://proceedings.mlr.press/v202/mitchell23a.html (accessed on 15 October 2026).
16. Masih, A.; Afzal, B.; Firdoos, S.; Mahmood, J.; Ali, A.; Abdulnabi, M.S.; Balungu, D.M. Classifying Human vs. AI Text with Machine Learning and Explainable Transformer Models. *Sci. Rep.* **2025**, *15*, 43310. [CrossRef] [PubMed]
17. Sui, M. Writing with Emotion? Assessing Emotional Valence and Appeals in AI-Generated vs. Human-Written Articles. *AI Soc.* **2026**, *41*, 1521–1535. [CrossRef]
18. Amirjalili, F.; Neysani, M.; Nikbakht, A. Exploring the Boundaries of Authorship: A Comparative Analysis of AI-Generated Text and Human Academic Writing in English Literature. *Front. Educ.* **2024**, *9*, 1347421. [CrossRef]
19. Mizumoto, A.; Yasuda, S.; Tamura, Y. Identifying ChatGPT-Generated Texts in EFL Students' Writing: Through Comparative Analysis of Linguistic Fingerprints. *Appl. Corpus Linguist.* **2024**, *4*, 100106. [CrossRef]
20. Rios, S.; Zhang, Y.; Oliveira, E.A. *Authorship Verification in Software Engineering Education: Forget ChatGPT and Focus on Students' Academic Writing Profiles*; Australasian Society for Computers in Learning in Tertiary Education (ASCILITE): Tugun, Australia, 2023; pp. 195–204.
21. Herbold, S.; Hautli-Janisz, A.; Heuer, U.; Kikteva, Z.; Trautsch, A. A Large-Scale Comparison of Human-Written versus ChatGPT-Generated Essays. *Sci. Rep.* **2023**, *13*, 18617. [CrossRef] [PubMed]
22. Za'rour, R.; Altakhaineh, A.R.M. False Positive Risk in AI Detection of L2 Academic Writing: A Case Study. *Discourses Cult.* **2026**, *25*, 37–78. [CrossRef]
23. AL-Rawas, M.; Abdul Qader, O.A.J.; Lin, G.S.S.; Beh, Y.H.; Sabarudin, M.A.; Ang, Y.; Low, J.F.; Abdullah, J.Y.; Noorani, T.Y. Human versus Artificial Intelligence: Investigating Ability of Young Academics from Research and Non-Research Institutions to Identify ChatGPT-Generated Dental Research Abstracts. *Sci. Rep.* **2026**, *16*, 12287. [CrossRef] [PubMed]
24. Erol, G.; Ergen, A.; Gülşen Erol, B.; Kaya Ergen, Ş.; Bora, T.S.; Çölgeçen, A.D.; Araz, B.; Şahin, C.; Bostancı, G.; Kılıç, İ.; et al. Can We Trust Academic AI Detective? Accuracy and Limitations of AI-Output Detectors. *Acta Neurochir.* **2025**, *167*, 214. [CrossRef] [PubMed]
25. Doru, B.; Maier, C.; Busse, J.S.; Lücke, T.; Schönhoff, J.; Enax-Krumova, E.; Hessler, S.; Berger, M.; Tokic, M. Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study. *JMIR Med. Educ.* **2025**, *11*, e62779. [CrossRef] [PubMed]
26. Fiedler, A.; Döpke, J. Do Humans Identify AI-Generated Text Better than Machines? Evidence Based on Excerpts from German Theses. *Int. Rev. Econ. Educ.* **2025**, *49*, 100321. [CrossRef]
27. Waltzer, T.; Pilegard, C.; Heyman, G.D. Can You Spot the Bot? Identifying AI-Generated Writing in College Essays. *Int. J. Educ. Integr.* **2024**, *20*, 11. [CrossRef]
28. Liu, J.Q.J.; Hui, K.T.K.; Al Zoubi, F.; Zhou, Z.Z.X.; Samartzis, D.; Yu, C.C.H.; Chang, J.R.; Wong, A.Y.L. The Great Detectives: Humans versus AI Detectors in Catching Large Language Model-Generated Medical Writing. *Int. J. Educ. Integr.* **2024**, *20*, 8. [CrossRef]
29. Elkhatat, A.M.; Elsaid, K.; Almeer, S. Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text. *Int. J. Educ. Integr.* **2023**, *19*, 17. [CrossRef]

30. Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltynek, T.; Guerrero-Dib, J.; Popoola, O.; Sigut, P.; Waddington, L. Testing of Detection Tools for AI-Generated Text. *Int. J. Educ. Integr.* **2023**, *19*, 26. [CrossRef]
31. Rashidi, H.H.; Fennell, B.D.; Albahra, S.; Hu, B.; Gorbett, T. The ChatGPT Conundrum: Human-Generated Scientific Manuscripts Misidentified as AI Creations by AI Text Detection Tool. *J. Pathol. Inform.* **2023**, *14*, 100342. [CrossRef] [PubMed]
32. Cheng, A.; Lin, Y.; Reedy, G.; Joseph, C.; Wirkowski, S.; Mallette, V.; Nagesh, V.; Krieser, D.; Calhoun, A. Ability of AI Detection Tools and Humans to Accurately Identify Different Forms of AI-Generated Written Content. *Adv. Simul.* **2025**, *10*, 66. [CrossRef] [PubMed]
33. Hadra, M.; Cambridge, K.; Mesbah, M. Evaluating the Accuracy and Reliability of AI Content Detectors in Academic Contexts. *Int. J. Educ. Integr.* **2026**, *22*, 4. [CrossRef]
34. Walters, W.H. The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Inf. Sci.* **2023**, *7*, 20220158. [CrossRef]
35. 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 Digit. Med.* **2023**, *6*, 75. [CrossRef] [PubMed]
36. Antipova, K.; Horban, H. *A Metric-Driven Evaluation of Rephrased and Generated Texts*; CEUR-WS: Aachen, Germany, 2025; Volume 4158, pp. 51–60.
37. Perkins, M.; Roe, J.; Vu, B.H.; Postma, D.; Hickerson, D.; McGaughran, J.; Khuat, H.Q. Simple Techniques to Bypass GenAI Text Detectors: Implications for Inclusive Education. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 53. [CrossRef]
38. Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; Iyyer, M. Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 27469–27500. [CrossRef]
39. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-Generated Text Be Reliably Detected? Stress Testing AI Text Detectors Under Various Attacks. *Trans. Mach. Learn. Res.* **2025**. Available online: https://openreview.net/forum?id=OOgsAZdFOt (accessed on 15 October 2026).
40. Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; Zou, J. GPT Detectors Are Biased against Non-Native English Writers. *Patterns* **2023**, *4*, 100779. [CrossRef] [PubMed]
41. Civelekler, M.; Citirik, M. Exploring Artificial Intelligence's Role in Citation Generation for Ocular Inflammation and Uveal Diseases Research: A Comparative Evaluation Across Four Models. *Ocul. Immunol. Inflamm.* **2026**, *34*, 397–402. [CrossRef] [PubMed]
42. Acut, D.P.; Malabago, N.K.; Malicoban, E.V.; Galamiton, N.S.; Garcia, M.B. "ChatGPT 4.0 Ghosted Us While Conducting Literature Search:" Modeling the Chatbot's Generated Non-Existent References Using Regression Analysis. *Internet Ref. Serv. Q.* **2025**, *29*, 27–54. [CrossRef]
43. Aljamaan, F.; Temsah, M.-H.; Altamimi, I.; Al-Eyadhy, A.; Jamal, A.; Alhasan, K.; Mesallam, T.A.; Farahat, M.; Malki, K.H. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. *JMIR Med. Inform.* **2024**, *12*, e54345. [CrossRef] [PubMed]
44. Xu, Y.; Gao, J.; Yu, X.; Bi, B.; Shen, H.; Cheng, X. ALiiCE: Evaluating Positional Fine-Grained Citation Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Association for Computational Linguistics (ACL): Singapore, 2025; Volume 1, pp. 545–561.
45. Sarol, M.J.; Schneider, J.; Kilicoglu, H. Automatic Identification of Citation Distortions in Biomedical Literature: A Case Study. *Proc. Assoc. Inf. Sci. Technol.* **2025**, *62*, 595–606. [CrossRef]
46. Sarol, M.; Ming, S.; Radhakrishna, S.; Schneider, J.; Kilicoglu, H. Assessing Citation Integrity in Biomedical Publications: Corpus Annotation and NLP Models. *Bioinformatics* **2024**, *40*, btae420. [CrossRef] [PubMed]
47. Sarringhaus, L.; McGrew, W.; Marchant, L. Misuse of Anecdotes in Primatology: Lessons from Citation Analysis. *Am. J. Primatol.* **2005**, *65*, 283–288. [CrossRef] [PubMed]
48. Greenberg, S.A. How Citation Distortions Create Unfounded Authority: Analysis of a Citation Network. *BMJ* **2009**, *339*, 210–213. [CrossRef] [PubMed]
49. Zheng, H.; Fu, Y.; Sarol, M.; Sarraf, I.; Schneider, J. Addressing Unreliability Propagation in Scientific Digital Libraries. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2024.
50. Bordignon, F.; Gambette, P. A Corpus of Critical Citations Contexts. *J. Open Humanit. Data* **2024**, *10*, 39. [CrossRef]
51. La Quatra, M.; Cagliero, L.; Baralis, E. Leveraging Full-Text Article Exploration for Citation Analysis. *Scientometrics* **2021**, *126*, 8275–8293. [CrossRef]
52. Sibaroni, Y.; Widyantoro, D.H.; Khodra, M.L. Automatic Identification of Compare Paper Relations. *Int. J. Electr. Eng. Inform.* **2020**, *12*, 141–154. [CrossRef]
53. Zeng, T.; Acuna, D.E. Modeling Citation Worthiness by Using Attention-Based Bidirectional Long Short-Term Memory Networks and Interpretable Models. *Scientometrics* **2020**, *124*, 399–428. [CrossRef]
54. Lauscher, A.; Glavaš, G.; Ponzetto, S.P.; Eckert, K. Investigating the Role of Argumentation in the Rhetorical Analysis of Scientific Publications with Neural Multi-Task Learning Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Singapore, 2018; pp. 3326–3338.

55. He, J.; Chen, C. *Understanding the Changing Roles of Scientific Publications via Citation Embeddings*; CEUR-WS: Aachen, Germany, 2017; Volume 2004, pp. 42–48.
56. Zhang, J.; Liu, J.; Haihong, E.; Hu, T.; Qiao, X.; Tang, Z. A Paper Mill Detection Model Based on Citation Manipulation Paradigm. *J. Data Inf. Sci.* **2025**, *10*, 167–187. [CrossRef]
57. Al-Sinani, H.S.; Mitchell, C.J. AI-Generated Papers and Manipulation of Academic Metrics: A Case Study. In *Proceedings of the 8th IEEE Congress on Information Science and Technology (CiSt 2025), Marrakech, Morocco*; IEEE: New York, NY, USA, 2025; pp. 583–588. [CrossRef]
58. Ibrahim, H.; Liu, F.; Zaki, Y.; Rahwan, T. Citation Manipulation through Citation Mills and Pre-Print Servers. *Sci. Rep.* **2025**, *15*, 5480. [CrossRef] [PubMed]
59. Besançon, L.; Cabanac, G.; Labbé, C.; Magazinov, A. Sneaked References: Fabricated Reference Metadata Distort Citation Counts. *J. Assoc. Inf. Sci. Technol.* **2024**, *75*, 1368–1379. [CrossRef]
60. Basil, D.Z.; Burton, S.; Soboleva, A.; Nesbit, P. COERCIVE CITATION: UNDERSTANDING THE PROBLEM AND WORKING TOWARD A SOLUTION. *Acad. Manag. Perspect.* **2023**, *37*, 205–219. [CrossRef]
61. Avros, R.; Keshet, S.; Toledano-Kitai, D.; Vexler, E.; Volkovich, Z. Detecting Manipulated Citations through Disturbed Node2Vec Embedding. In *2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*; IEEE: New York, NY, USA, 2023; pp. 274–278. [CrossRef]
62. Kojaku, S.; Livan, G.; Masuda, N. Detecting Anomalous Citation Groups in Journal Networks. *Sci. Rep.* **2021**, *11*, 14524. [CrossRef] [PubMed]
63. Haghighat, M.; Hayatdavoudi, J. How Hot Are Hot Papers? The Issue of Prolificacy and Self-Citation Stacking. *Scientometrics* **2021**, *126*, 565–578. [CrossRef]
64. Abramo, G.; D'Angelo, C.A.; Grilli, L. The Effects of Citation-Based Research Evaluation Schemes on Self-Citation Behavior. *J. Informetr.* **2021**, *15*, 101204. [CrossRef]
65. Baccini, A.; De Nicolao, G.; Petrovich, E. Citation Gaming Induced by Bibliometric Evaluation: A Country-Level Comparative Analysis. *PLoS ONE* **2019**, *14*, e0221212. [CrossRef] [PubMed]
66. Heneberg, P. From Excessive Journal Self-Cites to Citation Stacking: Analysis of Journal Self-Citation Kinetics in Search for Journals, Which Boost Their Scientometric Indicators. *PLoS ONE* **2016**, *11*, e0153730. [CrossRef] [PubMed]
67. Wilhite, A.W.; Fong, E.A. Coercive Citation in Academic Publishing. *Science* **2012**, *335*, 542–543. [CrossRef] [PubMed]
68. Mehregan, M.; Moghiman, M. The Unnoticed Issue of Coercive Citation Behavior for Authors. *Publ. Res. Q.* **2024**, *40*, 164–168. [CrossRef]
69. Májovský, M.; Černý, M.; Netuka, D.; Mikolov, T. Perfect Detection of Computer-Generated Text Faces Fundamental Challenges. *Cell Rep. Phys. Sci.* **2024**, *5*, 101769. [CrossRef]
70. Liang, W.; Izzo, Z.; Zhang, Y.; Lepp, H.; Cao, H.; Zhao, X.; Chen, L.; Ye, H.; Liu, S.; Huang, Z.; et al. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 21–27 July 2024*; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 29575–29620. Available online: https://proceedings.mlr.press/v235/liang24b.html (accessed on 15 October 2026).
71. Spennemann, D.H.; Spennemann, R.J.; Singh, C.L. Examining and Detecting Academic Misconduct in Written Documents Using Revision Save Identifier Numbers in MS Word as Exemplified by Multiple Scenarios. *Forensic Sci. Int. Digit. Investig.* **2024**, *51*, 301821. [CrossRef]
72. Bloch, L.; Rückert, J.; Friedrich, C.M. PreprintResolver: Improving Citation Quality by Resolving Published Versions of ArXiv Preprints Using Literature Databases. In *International Conference on Theory and Practice of Digital Libraries*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 14241, pp. 47–61. [CrossRef]
73. Wu, J.; Yang, S.; Zhan, R.; Yuan, Y.; Chao, L.S.; Wong, D.F. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *Comput. Linguist.* **2025**, *51*, 275–338. [CrossRef]
74. Kehkashan, T.; Riaz, R.A.; Al-Shamayleh, A.S.; Akhunzada, A.; Ali, N.; Hamza, M.; Akbar, F. AI-Generated Text Detection: A Comprehensive Review of Methods, Datasets, and Applications. *Comput. Sci. Rev.* **2025**, *58*, 100793. [CrossRef]
75. Sun, Y.; Sheng, D.; Zhou, Z.; Wu, Y. AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1278. [CrossRef]
76. Committee on Publication Ethics (COPE) Authorship and AI Tools: COPE Position Statement. Available online: https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools (accessed on 11 October 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.