publications

MDPI

Article

On the Vulnerability of Citation Metrics in the Era of Generative Artificial Intelligence †

Kay Smarsly [1,2]

[1] Institute of Digital and Autonomous Construction, Hamburg University of Technology, 21079 Hamburg, Germany; kay.smarsly@tuhh.de
[2] United Nations University (UNU) Hub on Engineering to Face Climate Change, United Nations University Institute for Water, Environment and Health (UNU-INWEH), Hamburg University of Technology, 21079 Hamburg, Germany

† This article is a revised and expanded version of a paper entitled "On the reliability of citation metrics in civil and building engineering in the age of generative artificial intelligence", which was presented at the International Conference on Computing in Civil and Building Engineering (ICCCBE), 23–26 March 2026, Taipei, Taiwan.

Abstract

Large language model (LLM) chatbots, as a widely used form of generative artificial intelligence, have reduced the marginal cost of producing publication-style manuscripts and have expanded feasible routes for manipulating citation metrics within the publishing ecosystem. Citation-based indicators (e.g., the h-index, the i10-index, and total citation counts) remain embedded in research evaluation and are sensitive to indexing practices of bibliographic databases, with Google Scholar providing broad coverage combined with comparatively limited curation. In this study, a systematic literature review is conducted to synthesize reported mechanisms of citation-metric manipulation and to examine limitations of citation-metric use, including evidence reported in civil engineering. A Google Scholar proof-of-concept case study examines whether the indexing of LLM-assisted, non-peer-reviewed documents with concentrated references to a target author is associated with changes in author-level citation metrics under platform-specific conditions. After indexing, a stepwise increase in author-level metrics is observed, demonstrating the feasibility of citation-metric manipulation under the platform-specific conditions. Finally, this paper discusses the implications for research integrity and citation manipulation in the era of generative artificial intelligence. It also presents recommendations for researchers, academic institutions and evaluation committees, publishers and editors, bibliographic database providers, and funding institutions and policymakers.

Keywords: large language models; generative artificial intelligence; citation metrics; h-index; Google Scholar; bibliographic databases; publishing ecosystem; research evaluation

Academic Editors: Guoqiang Liang and Shuo Zhang

Received: 6 March 2026
Revised: 27 March 2026
Accepted: 8 April 2026
Published: 11 April 2026

1. Introduction

Despite the central role of citation metrics within academic evaluation, concerns about the reliability of metrics-based "quality" indicators have increased (Nicholas et al., 2020). Bibliometric indicators, including the h-index, the i10-index and total citation counts, are widely applied to guide faculty promotion, hiring procedures, performance-based remuneration, institutional benchmarking and research funding allocation (Öztürk & Taşkın, 2024). Initially introduced as proxies for scholarly influence, citation metrics are now frequently interpreted as direct quality indicators (Kendall, 2024). A metric-centric research-evaluation culture has emerged across scientific disciplines, including civil engineering,

where publication and citation tallies are increasingly treated as substitutes for research quality and individual achievement (Ramadhan et al., 2024). Strong reliance on citation metrics has come under scrutiny, due to susceptibility to manipulative tactics, such as self-citation, citation cartels, or coercive citation tactics (Mehregan & Moghiman, 2024). Further challenges arise from differences among bibliographic databases, notably Google Scholar, Scopus or Web of Science, concerning coverage extent, quality assurance, and error rates. In particular, Google Scholar, widely favored by researchers due to extensive coverage and rapid indexing, is comparatively susceptible to intentional manipulations compared to curated databases (Fire & Guestrin, 2019). Documented instances show fabricated papers with deliberate self-citations successfully inflating citation counts on Google Scholar (Fong & Wilhite, 2017). Moreover, citation-boosting services demonstrably can inflate author-level metrics through strategically indexed documents (López-Cózar et al., 2014).

Recent advances in large language models (LLMs), which are the core engines of generative artificial intelligence and chat-based systems such as ChatGPT (Floridi & Chiriatti, 2020), have made it straightforward to generate fluent, publication-style prose, including abstracts, full manuscripts, and literature reviews that may contain convincing yet fictitious citations (Lund et al., 2023). Journal policies have responded by refining guidance on disclosure and authorship in relation to artificial intelligence tools (Goyanes et al., 2025), yet documented incidents include artificial intelligence being credited as an author and fabricated references being embedded in scholarly works (Cabezas-Clavijo et al., 2024). Once LLM-assisted documents enter preprint servers, repositories, or journals and become indexed, distorted reference metadata can propagate into citation databases and inflate citation counts (Besançon et al., 2023), affecting metric-based assessments and subsequent decisions (Thelwall & Kurt, 2025).

This paper evaluates the feasibility of citation-metric manipulation within the publishing ecosystem in the era of generative artificial intelligence. Google Scholar serves as the focal platform for a proof-of-concept case study. Based on the findings, the paper derives governance implications and recommendations for stakeholders involved in publishing, indexing, and evaluation. Civil engineering serves as an application context for interpreting evaluation practices and domain-specific metrics, rather than as the empirical focus of the platform demonstration. In civil engineering, academic evaluation reflects the same concerns described above, as academic performance in civil engineering is often assessed primarily by publication counts and citation-based indices. One empirical study covering 93 Greek university departments has ranked departments by aggregated faculty h-indices obtained from Google Scholar profiles (Altanopoulou et al., 2012). Reliance on metric-based rankings creates scope for strategic behavior; a bibliometric analysis in Turkey has reported that 96% of papers published in so-called "questionable" journals have been included in academic promotion dossiers, indicating that performance indicators can steer researchers towards low-quality publication outlets (Aksnes et al., 2019). Multiple studies have shown that raw citation counts do not necessarily reflect perceived research quality. Award-winning papers in civil engineering frequently attract below-average citation numbers, while factors unrelated to scientific merit (e.g., community size or fashionable topics) can systematically skew citation levels (Kazakis, 2014). Citation metrics therefore remain useful tools in civil engineering, but when applied in academic evaluation, the metrics must be interpreted cautiously and within disciplinary context (Smarsly, 2026).

The remainder of this paper is structured as follows: Section 2 provides a concise background on large language models and generative artificial intelligence to contextualize the implications and recommendations presented in this study. Then, Section 3 systematically reviews documented vulnerabilities of citation metrics. Next, Section 4 presents the Google Scholar proof-of-concept case study to empirically investigate the impact of

AI-enabled citation manipulation. Thereupon, Section 5 discusses the implications and formulates recommendations based on the systematic review and the case study. Finally, Section 6 concludes the paper with a summary of key findings and outlines avenues for potential future research. In addition, the study relates its findings to existing frameworks on responsible research evaluation and the use of generative artificial intelligence in scholarly communication, thereby situating the results within current developments in research governance and publication practices.

2. Background on Large Language Models and Generative Artificial Intelligence

The advancement of generative artificial intelligence, particularly through large language models, reshapes all scientific and engineering disciplines. LLMs, characterized by complex neural network architectures trained on vast text corpora, enable powerful text generation, contextual understanding, and automated reasoning. The cost and effort required to produce publication-style text have significantly been reduced. Tasks that previously required substantial time and expertise, such as drafting abstracts, structuring manuscripts or compiling reference lists, can now be performed rapidly using LLM-based tools, lowering the barrier for generating document-like outputs that resemble scholarly publications. This section summarizes publishing-relevant properties of LLMs to contextualize manipulation pathways and indexing risks examined in the subsequent sections.

Large language models did not emerge overnight; rather, large language models represent the culmination of decades of progress in natural language processing and artificial intelligence (Zubiaga, 2024). Early language models in the mid-20th century relied on rule-based systems and formal grammar and modeled word frequencies but struggled with contextual meaning and complex grammatical structures (Chomsky, 1956). Over the past decade, major advances have led to the emergence of what are now classified as 'large' language models (Min et al., 2023). Figure 1 shows a timeline of widely used large language models.

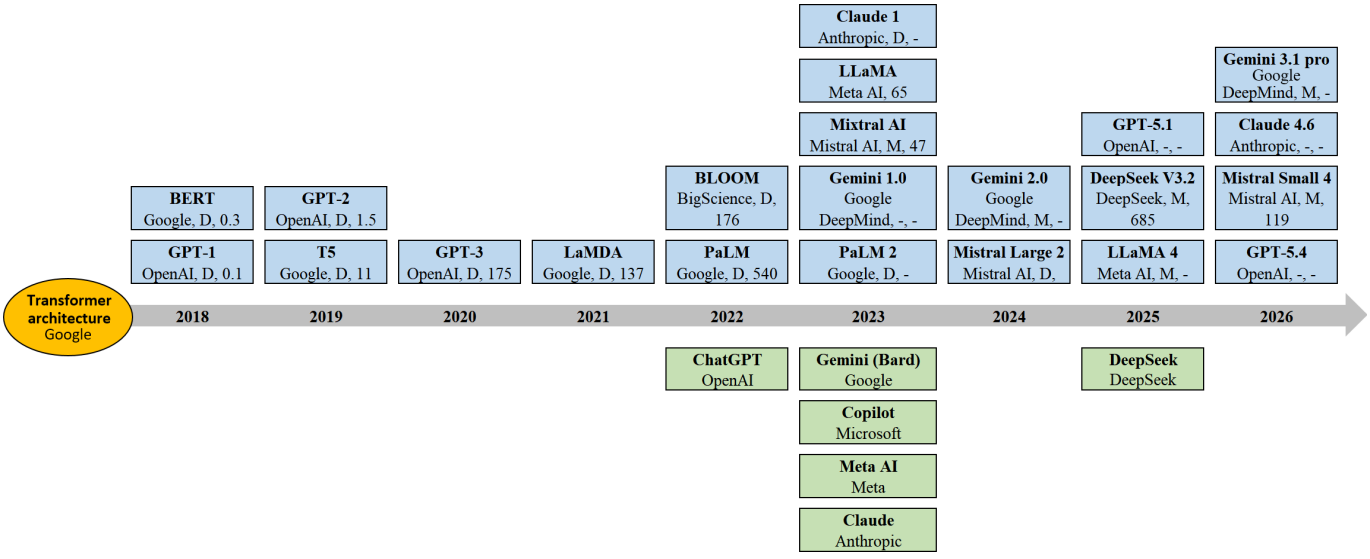

Figure 1. History of widely used large language models and chatbots (D = dense model; M = mixture-of-experts model; parameters in billions according to developer, "-" if unknown).

The growing use of large language models in academic writing brings unprecedented efficiency but also profound challenges. On the one hand, the preceding discussion has highlighted how generative AI can streamline literature review and drafting; on the other

hand, serious ethical and quality concerns have emerged. Ensuring originality and integrity of scholarly work is a primary challenge. For example, some high-profile journals (e.g., Science) reacted by banning LLM-generated text outright due to concerns about responsible authorship (Hosseini et al., 2023). The Editor-in-Chief of Science has compared AI-generated content to plagiarism, asserting that texts not genuinely authored by humans lack originality. Since LLMs are trained on large amounts of internet-sourced content, they may reproduce existing sentences. Without rigorous oversight, such outputs may unintentionally incorporate other ideas or wording without proper attribution. Additionally, despite their linguistic fluency, AI-generated texts often lack factual reliability. Research indicates that ChatGPT-like models regularly produce factual inaccuracies and fabricated yet seemingly credible references, attributable to the hallucinations previously discussed (Lund & Naheem, 2024). LLMs are therefore capable of generating references that are syntactically plausible, but factually incorrect or entirely fabricated, including non-existent publications or misleading bibliographic details. Despite these inaccuracies, such references often conform to expected formatting conventions, which are difficult to detect without careful verification. The tendency of these models to generate "gibberish" studies containing false data and false citations critically undermines trust in scholarly publishing (Vincent, 2023). Misplaced confidence in polished, AI-written text can thus lead to the inadvertent spread of misinformation in the literature.

Closely intertwined are questions of transparency, authorship, and accountability. Many publisher guidelines now insist that researchers disclose any use of LLM tools in manuscripts (Yoo, 2025). Major journals (including the Nature family) have ruled that AI tools cannot be credited as authors, since true authorship carries accountability that a machine cannot bear (Nature, 2023). However, enforcing these norms poses its own challenges. Blanket bans on AI assistance are viewed as impractical by many, given that covert use is difficult to detect and outright prohibition might simply drive usage underground, and current detection methods for AI-generated text are limited. There is broad consensus that ethical authorship is essential to preserving the integrity of scholarly communication. Human researchers must remain fully responsible for their work and must openly disclose AI assistance. When LLM-generated texts are disseminated through repositories, conference platforms, or other lightly curated publication channels, inaccurate or automatically generated references can become part of citation databases and potentially influence citation-based indicators. In the long term, this development might also distort academic metrics. If unscrupulous actors use LLMs to generate numerous papers or citations, traditional indicators of research productivity and impact could be manipulated. The study presented in this paper aims to address these emerging concerns. Section 3 reviews documented mechanisms of citation-metric manipulation, and Section 4 evaluates the feasibility of such manipulation through a Google Scholar proof-of-concept case study.

3. A Systematic Literature Review of Citation Metrics in the Era of Generative Artificial Intelligence

A systematic review is conducted to investigate the vulnerability of citation metrics in the era of generative artificial intelligence, particularly within the civil engineering domain. The review follows a modified version of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology (Page et al., 2021). The term "modified" indicates deviations from the strict PRISMA protocol to accommodate the specifics of this review, including additional methodological steps—such as iterative backward snowballing (Wohlin, 2014)—and considering multiple interrelated research questions (RQs):

- RQ 1: What methods of citation metric manipulation are reported?

- RQ 2: To what extent are citation metrics used in formal research evaluation procedures?
  - RQ 2.1: . . .specifically in the civil engineering domain?
- RQ 3: How widely are large language models being used to generate scientific papers?

The systematic review is organized into four methodological phases, (i) identification, (ii) screening, (iii) eligibility, as well as (iv) snowballing and inclusion. Separate search strings and independent screening and selection procedures are applied for each research question. Each phase is described in detail in the subsequent subsections, after which the results are reported. A PRISMA-style flow diagram of the modified review procedure is provided in the Supplementary Materials (Figure S1). All records are retrieved from the Scopus database within the period from 21 June 2026 to 23 June 2026.

3.1. Identification

The first phase ("Identification") of the systematic review comprises the design and execution of a structured literature search aimed at capturing records potentially relevant to the research questions. To maintain rigor, reproducibility, and transparency, the identification phase is divided into five steps, outlined in the following list.

- **Definition of core concepts and initial search terms:** For each research question, logical blocks of concepts are defined, for example, a "core concepts" block and a "general manipulation terms" block are established for RQ 1. Within each block, initial keywords (e.g., "citation metrics", "citation indicators") are expanded into specific search terms by considering synonyms, alternative expressions, and related terminology (e.g., "citation metric*", "citation indicator*"). Truncation techniques and Boolean operators are applied to combine the search terms and to capture various linguistic variations. The sets of search terms form the basis of initial search strings defined for each research question.
- **Conducting the initial search:** Using the initial search strings specified for each research question, an initial search is carried out. The objective of this step is to obtain an initial overview of the retrieved records and to assess the completeness, specificity, and precision of the search strategy. The findings obtained from the initial search form the basis for subsequent refinement of the search strings.
- **Refinement of the search strings:** Following the initial search, the search strings for each research question are systematically revised. The revision includes verifying whether predefined seminal studies are retrieved by the current strings. If key studies are not captured, the strings are adjusted and extended to improve coverage. Conversely, overly broad or irrelevant terms that produce large numbers of non-relevant records are removed. The use of Boolean operators and truncation is further optimized to increase overall precision.
- **Execution of the final search:** Upon refinement, a final search string is established for each research question, targeting an appropriate balance between comprehensive coverage and high precision. The final search strings for each research question, organized into logical blocks, are illustrated in Figure 2 and reported in full in the Supplementary Materials (Table S1). The final search is conducted in the "title", "abstract", and "keywords" fields of the Scopus database. Additional filters are applied for language (English) and document type (article, review, and conference paper) to further increase precision and relevance.
- **Removal of duplicates:** All records returned by the final search are examined for duplicate entries, which are then removed manually.

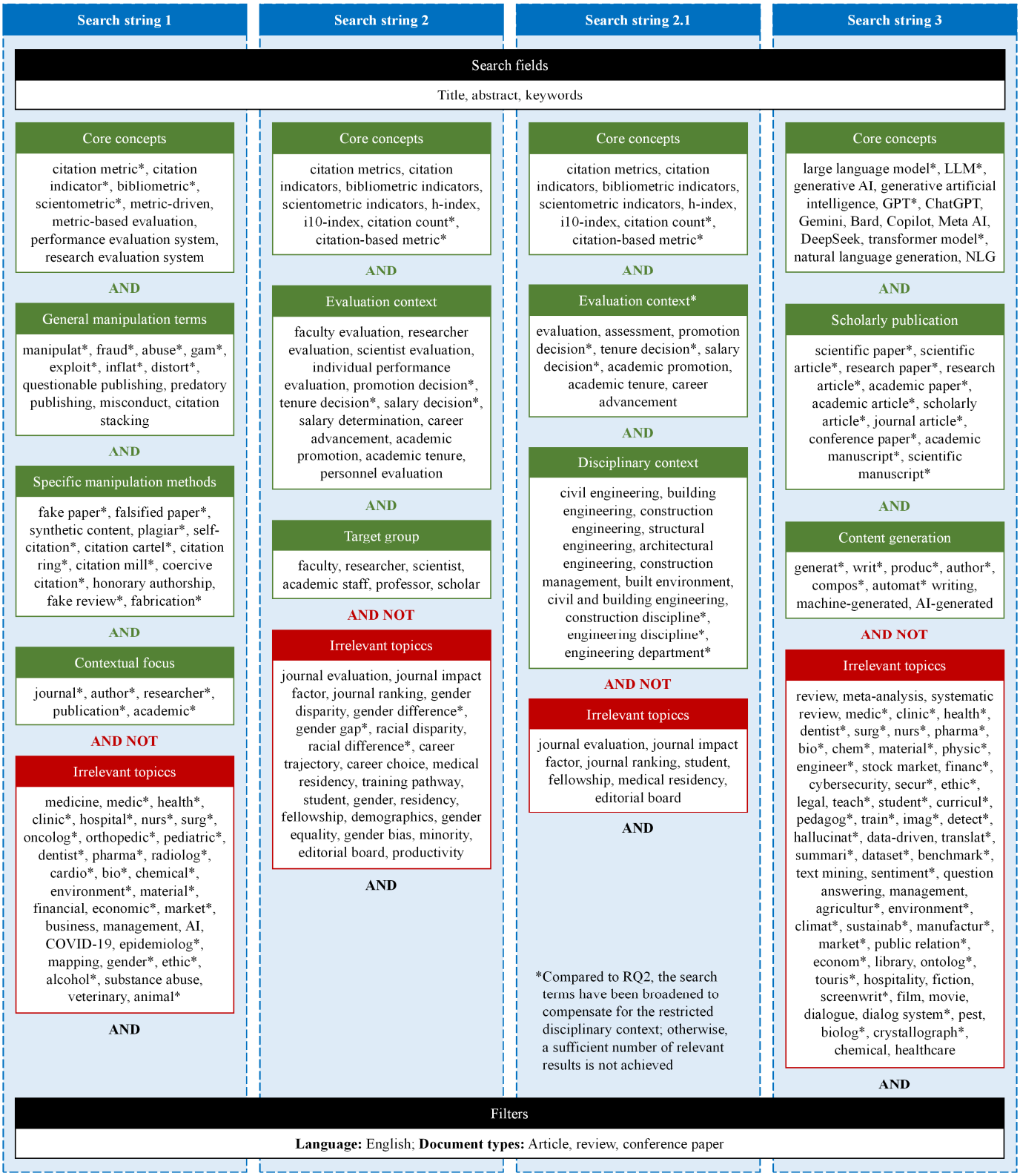

Figure 2. Final search strings defined for each research question.

3.2. Screening

Inclusion criteria require a clear thematic link to the research questions, contextual relevance, documented methodological rigor, suitable target populations, and explicit content relevance (e.g., for RQ 1, documented experimental attempts at citation manipulation). In operational terms, methodological rigor is judged based on the presence of

a clearly identifiable study design, transparent data sources or corpus definition, and an explicit analytical or evaluative procedure. Explicit content relevance is judged based on whether a record directly addresses the respective research question, rather than referring to citation metrics, research evaluation, civil engineering, or large language models only incidentally or in passing. Exclusion criteria eliminate records with irrelevant topics, insufficient methodological clarity, inappropriate populations, or non-scholarly formats, such as editorials, commentaries, or purely theoretical discussions without empirical evidence. Formal criteria related to language (English) and document type (article, review, conference paper) are already enforced via database filters during the identification phase and are therefore not re-applied at this stage. Only records with a high likelihood of meaningfully addressing the research questions are retained for full-text assessment.

3.3. Eligibility

In the third phase ("Eligibility"), all records retained after screening are examined in full text using the same inclusion and exclusion criteria employed at the abstract level. The same operational definitions of methodological rigor and explicit content relevance are applied at full-text level. Inclusion criteria emphasize clear thematic alignment with the research questions, methodological rigor, and empirical or analytical adequacy. Records are excluded when relevance to the core research questions is limited, methodological transparency is insufficient, or empirical evidence is inadequate. The eligibility phase ensures that only records of high relevance and sufficient methodological quality proceed to the subsequent snowballing and inclusion phase.

3.4. Snowballing and Inclusion

In the fourth phase ("Snowballing and inclusion"), all records deemed eligible in the full-text evaluation are formally included in the systematic review. Additional potentially relevant records are identified via iterative backward snowballing, which involves screening the reference lists of the included records. Snowballing continues until no additional eligible records are identified for inclusion (Wohlin, 2014). Newly identified records are then subjected to the same eligibility assessment, and their reference lists are again examined in subsequent snowballing iterations. The final sets of included records from the iterative process are documented and summarized in the following subsection.

3.5. Results

This subsection reports the results of the systematic review. The corresponding study selection process is summarized in the PRISMA-style flow diagram provided in the Supplementary Materials (Figure S1). This subsection focuses on factual reporting. Qualitative interpretations and in-depth discussion are deferred to Section 5, where the findings from the systematic review and the experimental study in Section 4 are synthesized into implications and recommendations. Table 1 summarizes the results of the systematic review, indicating the number of records identified at each review phase and the specific studies finally included. Following the table, the results addressing each research question are presented individually in the remainder of this subsection.

Figure 3 shows the annual number of identified papers and the corresponding citations per year for each research question. The data is taken from the "Identification (initial search)" phase because the phase provides the broadest landscape of potentially relevant papers. For RQ 1 (Figure 3a), the number of publications stays low until 2022 but noticeably increases from 2023 onward, while citations follow a steady upward trend, reaching a maximum in 2025. With respect to RQ 2 (Figure 3b), an increase in paper output is observed in 2014, with respective citations in the following years. For RQ 2.1 (Figure 3c), specific to civil engineering, the number of publications remains generally low throughout the

observed period, accompanied by a slight rise in citations in recent years. Regarding RQ 3 (Figure 3d), publication activity is minimal until 2022, after which a substantial increase is observed from 2023, and the citation counts show a similar trajectory. The results for each research question are detailed in the following subsections.

Table 1. Summary of the review results.

	RQ 1	RQ 2	RQ 2.1	RQ 3
Identification (initial search)	92	174	24	1437
Identification (final search)	36	44	22	66
Screening	25	31	9	11
Eligibility	7	8	5	3
Snowballing and inclusion	10 (7 + 3)	11 (8 + 3)	7 (5 + 2)	11 (3 + 8)
	I. Ali et al. (2021), Fister et al. (2016), Fong and Wilhite (2017), Ibrahim et al. (2025), Kojaku et al. (2021), Mazov and Gureev (2019), Moustafa (2016), Ortega and Delgado-Quirós (2023), Tripathi et al. (2019), Q. Zhang et al. (2020)	Abramo et al. (2012), N. Ali et al. (2023), Dehnad et al. (2019), Guraya et al. (2016), Haddow and Hammarfelt (2019), Lim et al. (2025), Lippi and Mattiuzzi (2017), Marsicano et al. (2022), Mingers et al. (2023), Wang et al. (2022), Zerem et al. (2021)	Asfour and Al-Qawasmi (2024), Abramo et al. (2021), El-Adaway et al. (2019), Mustafa et al. (2025), Raheel et al. (2018), Salman et al. (2021), Usman et al. (2021)	Camp et al. (2025), Finkel-Gates (2025), Haider et al. (2024), Hosseini and Horbach (2023), Kendall and Teixeira da Silva (2024), Kobak et al. (2025), Lendvai (2025), Ramoni et al. (2024), Tang et al. (2024), Walters and Wilder (2023), M. Zhang and Zhao (2025)

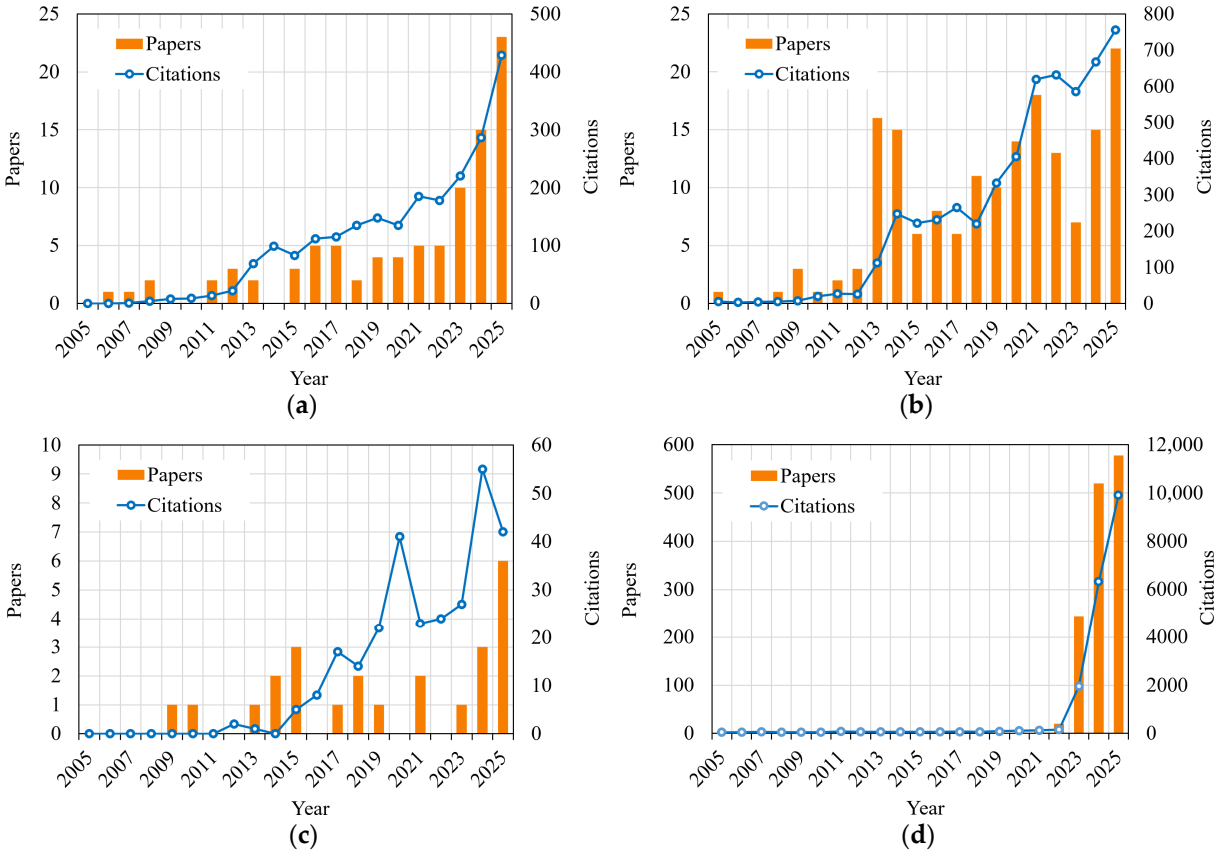

Figure 3. Annual distribution of the identified publications and citations per year, based on the "Identification (initial search)" phase: (**a**) RQ 1, (**b**) RQ 2, (**c**) RQ 2.1 (Civil engineering), (**d**) RQ 3.

3.5.1. Results Regarding Research Question 1 ("What Methods of Citation Metric Manipulation Are Reported?")

Manipulation of citation metrics is perpetrated through different methods. A prevalent method involves excessive self-citation, where authors deliberately cite their own publications to artificially inflate personal citation counts (Ibrahim et al., 2025). Additionally, citation cartels, i.e., groups of researchers who systematically and disproportionately cite each other's work, further distort citation-based evaluations (Fister et al., 2016). Guest authorship, where prominent researchers are listed without substantial contribution, increases both article visibility and citation likelihood. Honorary citations, i.e., references included due to social or strategic motivations rather than scientific merit, also distort scholarly metrics (Ibrahim et al., 2025). Collaborations with a larger number of authors are associated with increased occurrences of falsification or manipulation in retracted publications, implying that team size can be a factor in attempts to manipulate citation metrics (Q. Zhang et al., 2020). Citation metrics are also manipulated through commercial services, allowing researchers to purchase citations integrated into other publications for a fee, and artificial citations are introduced by individuals uploading AI-generated articles containing strategic references to targeted works onto unmoderated pre-print servers, artificially inflating citation counts (Ibrahim et al., 2025). The following studies deliberately fabricated fake or nonsensical papers to demonstrate the vulnerability of citation metrics (Labbé, 2010; Labbé & Labbé, 2013).

Editorial manipulation is another critical issue, with journal editors coercing authors to insert irrelevant citations, thereby artificially boosting the journal's citation metrics (Fong & Wilhite, 2017). Editors also publish self-citing editorials or review articles that frequently reference recent papers from their journals, directly enhancing journal impact factors (Kojaku et al., 2021). Moreover, coordinated mutual citation agreements between multiple journals, known as "citation stacking", artificially inflate the bibliometric indicators of the journals involved (Ibrahim et al., 2025). It has been observed that only a few of the articles that should receive editorial notices (due to publishing fraud or manipulation) are actually corrected by journals, indicating insufficient responses to documented manipulations, such as plagiarism or data falsification (Ortega & Delgado-Quirós, 2023). Additionally, coercive citations by editors during peer review processes artificially boost journal impact factors through imposed irrelevant references (Moustafa, 2016). The manipulative practices largely persist because bibliographic databases lack effective detection mechanisms. At present, the systems often fail to distinguish manipulated citations from legitimate ones, allowing citation manipulation to remain largely unchecked (Tripathi et al., 2019). The problem is exacerbated in non-curated bibliographic databases, such as Google Scholar, where quality-control mechanisms are limited. Authors may exploit the databases to artificially inflate citation counts, often without detection (I. Ali et al., 2021).

3.5.2. Results Regarding Research Question 2 ("To What Extent Are Citation Metrics Used in Formal Research Evaluation Procedures?")

Across formal assessment frameworks, citation-based indicators constitute a standard component of research evaluation (N. Ali et al., 2023). Particularly the h-index and the i10-index are frequently used to summarize individual publication impact for tenure and promotion decisions (Marsicano et al., 2022), and total citation counts are likewise widely applied (Wang et al., 2022). Empirical studies report that higher h-index values are associated with increased institutional and governmental funding allocations (Guraya et al., 2016). At the same time, citation metrics exhibit well-documented limitations, including a strong dependence on academic career length rather than solely on research quality (Lippi & Mattiuzzi, 2017). In particular, the h-index does not adjust for discipline-specific citation practices and does not differentiate between the contributions of co-authors, which has

motivated the introduction of additional indicators into formal evaluation procedures, such as the g-index (Dehnad et al., 2019) or the z-score (Zerem et al., 2021). Despite these shortcomings, approximately 92% of analyzed promotion policies explicitly reference citation-based indicators, with higher prevalence in upper-middle-income countries and in the global South, and the relative weight assigned to citation metrics varies across research disciplines (Lim et al., 2025).

3.5.3. Results Regarding Research Question 2.1 ("To What Extent Are Citation Metrics Used in Formal Research Evaluation Procedures Specifically in the Civil Engineering Domain?")

In civil engineering, it is observed that research rewards are closely tied to publication citation metrics (El-Adaway et al., 2019). For example, formal research evaluation procedures use specialized indices as well as domain-specific, statistically optimized ranking formulas to reflect collaboration and sustained contributions (Mustafa et al., 2025). Furthermore, multi-authorship indices, such as the h_m-index, g_m-index, h_i-index, h_f-index, g_f-index, or h_F-index, are applied in civil engineering to rank authors, and variants of the h-index (such as the h_g-index) are used in qualitative judgements for award nominations by international civil engineering societies (Salman et al., 2021). The reliance on citation metrics for career advancement has been shown to affect the behavior of researchers. In Italy, for example, civil engineering professors notably increased their self-citation rate after the implementation of citation-based academic accreditation. The findings point to a "culture of counting" in civil engineering that may encourage unethical practices, such as excessive self-citation, and may disadvantage citation-sparse fields within the discipline (Asfour & Al-Qawasmi, 2024). It should be noted that citation metrics also influence prestigious awards bestowed by civil engineering societies, such as ASCE (American Society of Civil Engineering), CSCE (Canadian Society of Civil Engineering), ACI (American Concrete Institute) and ICE (Institute of Civil Engineering), but fewer than half of the awardees align with the top ranks suggested by these metrics (Raheel et al., 2018). Award committees instead tend to favor publications with fewer authors per paper and place additional weight on respective indicators, such as citations per year and composite indices (Usman et al., 2021).

3.5.4. Results Regarding Research Question 3 ("How Widely Are Large Language Models Being Used to Generate Scientific Papers?")

Large language models, freely accessible via online chat bots, such as ChatGPT and Gemini, are now widely integrated into scientific writing workflows and affect multiple stages of manuscript preparation. The tools are increasingly employed for drafting initial texts, enhancing conceptual clarity, and refining grammar and structure (Tang et al., 2024). Moreover, LLMs assist in transforming notes into structured academic content, support citation formatting, and facilitate literature reviews (Lendvai, 2025). As reported by Ramoni et al. (2024), AI assistance, by 2023, accounted for about 25% of manuscript writing and 15% of grant writing, primarily assisting in manuscript editing, proofreading, and technical tasks. Early-career researchers, particularly in computational fields and non-English-speaking regions, are major adopters, often utilizing the tools without explicit acknowledgment (Haider et al., 2024). Students also use LLMs to reduce time and effort for academic assignments, with applications ranging from minor edits to extensive rewriting (Finkel-Gates, 2025).

However, the integration of LLMs into scholarly publishing has led to ethical and reliability issues, as LLMs have generated false citations, propagating references to non-existent papers across academic databases, undermining the credibility of scholarly communication (Camp et al., 2025). Entirely fabricated empirical studies, including inappropriate methods and false statistical outcomes, have even infiltrated reputable publications (Walters & Wilder, 2023). High-impact journals have cited fictional LLM-generated articles, high-

lighting vulnerabilities in editorial and peer-review processes (M. Zhang & Zhao, 2025). Detection of such fraudulent activities currently relies primarily on manual verification, due to limitations in automated tools (Camp et al., 2025). The rise of AI-generated fraudulent papers through "paper mills" and predatory journals further complicates matters, with fake papers constituting roughly 24% in certain research fields (Ramoni et al., 2024).

As reported by Haider et al. (2024), approximately two-thirds of examined manuscripts (as of 2024) contained undisclosed LLM-generated content, a practice notably more prevalent in journals with expedited review processes compared to high-prestige outlets (Kobak et al., 2025). Additionally, the increasing adoption of LLMs for drafting peer-review reports, editorial decision letters, and concise manuscript summaries indicates the expanding role of LLMs within editorial workflows (Hosseini & Horbach, 2023). Nevertheless, explicitly acknowledging generative AI as co-authors remains highly controversial and challenges established norms of authorship attribution (Kendall & Teixeira da Silva, 2024).

4. A Google Scholar Proof-of-Concept Case Study

The systematic literature review reported in Section 3 has documented multiple manipulation pathways and platform-dependent vulnerabilities affecting citation-based indicators used in research evaluation. In addition, Section 3 has indicated that strategically prepared, LLM-assisted documents can contribute to artificial citation signals once indexed by bibliographic databases. To provide a demonstrative feasibility assessment under controlled and explicitly bounded conditions, this section presents a Google Scholar proof-of-concept case study and, in addition to the previously defined research questions, the following research question is to be answered:

- RQ 4: How easily can papers generated with the assistance of large language models influence author-level citation metrics on Google Scholar?

4.1. Experimental Design and Methodology

To address research question 4, two papers—paper A (Dragos, 2025a) and paper B (Dragos, 2025b)—have been written, each initially drafted with support from a large language model (ChatGPT, model GPT-4.5) and subsequently revised to comply with established academic standards regarding content and style. The proof-of-concept has been designed as a demonstrative case study on a single platform (Google Scholar) and a single target profile (n = 1 author). It does not support claims about prevalence, representativeness, or generalizability beyond these conditions. In terms of content, both papers focus on explainable artificial intelligence, i.e., the papers (i) review the interaction between generative AI and citation-based, metric-driven academic evaluation systems and (ii) summarize prior work conducted by the author of this study. Therefore, both papers necessarily contain numerous citations to the author's prior work for scholarly reasons rather than for artificial inflation—a feature deliberately included in the proof-of-concept case study.

Both papers have been presented at the 7th International Workshop on Explainable Artificial Intelligence in Civil Engineering (XAICE) at Hamburg University of Technology on 5 May 2025, before being disseminated through a webpage indexed by Google Scholar. The XAICE workshop has not involved external peer review. However, it must be emphasized that both papers meet academic standards, each addressing clearly formulated research hypotheses. Both papers have been authored and formally published by a different individual, thus eliminating self-citations by the author of this study. The citation compositions within the two papers differ strategically: In paper A, 50 out of a total of 110 references (approximately 45%) cite publications authored by the author of this study, while in paper B, 75 out of a total of 115 references (approximately 65%) cite publications authored by the author of this study. Paper B has been released on the webpage once paper

A had been indexed by Google Scholar. The sequential release strategy has intentionally been adopted to examine whether different levels of citation concentration are associated with different changes in indexing outcomes and author-level citation metrics under the platform-specific conditions. The sequence has reduced the risk of simultaneous indexing effects and has supported attribution of metric changes to a specific indexed document under an opaque indexing algorithm. Risk mitigation measures have been applied by limiting the study to two documents, avoiding iterative or ongoing uploads, and refraining from providing operational guidance for scaling or automating the approach.

4.2. Results

The citation composition and dissemination timelines of paper A and B are presented in Table 2, providing the number of citations referencing works by the author of this study and those citing external sources. Specifically, the table reports the respective dissemination, i.e., upload dates, as well as the dates each paper has been indexed by Google Scholar. Paper A has become discoverable via Google Scholar 14 days after dissemination, whereas paper B has become discoverable 19 days after dissemination. The time series presented in Figure 4 illustrates the weekly development of the citation metrics (total citations, h-index, and i10-index) of the author of this study. Before uploading paper A (5 June 2025), the total citations gradually increase. After Google Scholar indexed paper A on 19 June 2025, the citation count increases from 2555 to 2615 (+60 citations, 50 of which were attributed to paper A). A similar pattern is observed upon the indexing of paper B on 15 July 2025, raising the total citation count from 2629 to 2715 (+86 citations, including 75 from paper B). Between uploading paper A and the indexing of paper B, the i10-index increased from 69 to 71, while the h-index remained unchanged at 27. The results therefore indicate measurable changes in author-level metrics under the specified dissemination and indexing conditions. However, the results do not support broader causal claims beyond this platform-specific demonstration.

Table 2. Citation composition, dissemination dates, and Google Scholar indexing dates.

	Paper A (Dragos, 2025a)	**Paper B (Dragos, 2025b)**
Author's papers cited	50	75
External papers cited	60	40
Dissemination date	5 June 2025	28 June 2025
Indexing date	19 June 2025	15 July 2025

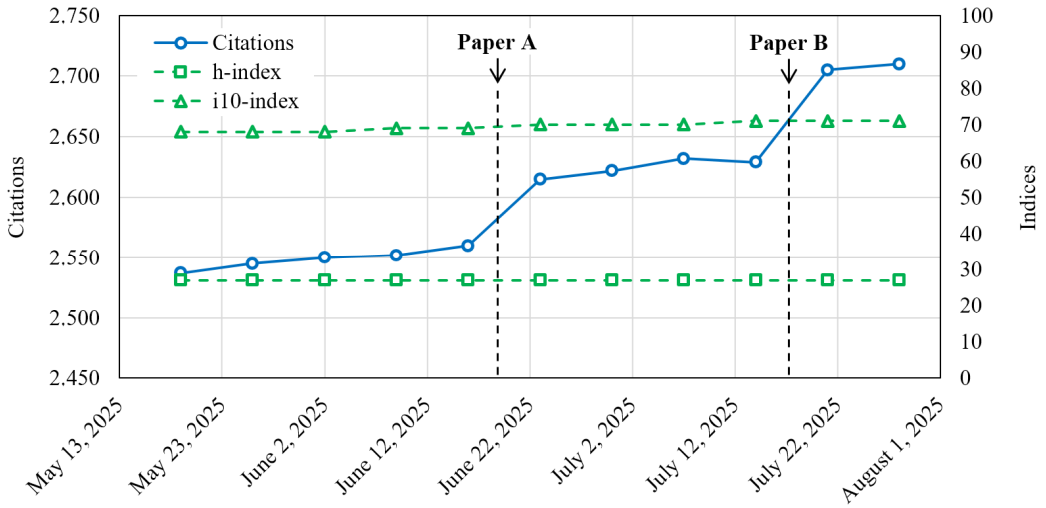

Figure 4. Time series illustrating the changes in the Google Scholar citation metrics.

4.3. Summary

The proof-of-concept case study shows that—under the platform-specific conditions—indexing of LLM-assisted, non-peer-reviewed documents with concentrated references is associated with measurable changes in author-level citation metrics on Google Scholar. In particular, concentrated referencing of one author's publications has translated into increased citation counts after indexing. It must be emphasized that the case study is intentionally limited in scope and should be interpreted as a proof-of-concept demonstration rather than as evidence of prevalence or system-wide effects. Nevertheless, the findings support a feasibility claim for metric inflation on a non-curated indexing platform, while limiting interpretation to the demonstrated mechanism and scope. Combined with the insights obtained from the systematic literature review presented in the previous section, an elaborate discussion of the findings, including broader implications and recommendations, will be provided in the following section.

5. Discussion, Implications, and Recommendations

A critical examination of the findings obtained from both the systematic literature review and the proof-of-concept case study reported in Section 3 and Section 4, respectively, is presented in this section. First, the discussion and implications synthesized from this study are presented. Then, aiming to support researchers, academic institutions and committees, publishers and editors, bibliographic databases as well as funding institutions and policymakers, specific recommendations are derived.

5.1. Discussion and Implications

In this subsection, the findings are interpreted and the implications are discussed according to the research questions.

5.1.1. Research Question 1 ("What Methods of Citation Metric Manipulation Are Reported?")

Manipulation of citation metrics poses a fundamental threat to the integrity and reliability of academic evaluation systems. Persistent failures in detection mechanisms within bibliographic databases allow unethical activities to remain largely undetected and uncorrected, thereby weakening the credibility of scholarly evaluation. The increasing availability of generative artificial intelligence, addressed in more detail in research questions 3 and 4, further intensifies the risk by enabling automated creation of plausible yet misleading academic documents. The implications drawn from the systematic review are summarized as follows. These implications should not be interpreted as suggesting that all citation-based evaluation is inherently unreliable, but rather that such evaluation remains vulnerable where detection and correction mechanisms are weak.

- Citation metric manipulation undermines the validity and accuracy of research evaluations, affecting critical academic decisions, such as faculty promotion, hiring procedures, performance-based remuneration, institutional benchmarking, and the allocation of research funding.
- Persistent manipulation reduces overall trust in scholarly publications, complicating the identification of genuinely impactful research and diminishing confidence in research outcomes.
- The rise of citation stacking and citation cartels reveals a more organized and collaborative form of metric manipulation that disadvantages ethical researchers.
- Reliance on manipulated metrics encourages a culture where quantity supersedes quality, thereby undermining robust scholarship and hindering the pursuit of innovative research.

5.1.2. Research Question 2 ("To What Extent Are Citation Metrics Used in Formal Research Evaluation Procedures?")

Citation metrics play a pivotal role in formal research evaluation procedures worldwide, increasingly shaping critical decisions in faculty promotions, institutional rankings, and research funding allocation. Despite broad acceptance and integration into formal evaluation procedures, significant concerns exist regarding validity, fairness, and objectivity. Additionally, heavy reliance on citation metrics potentially incentivizes publication strategies driven more by anticipated citations than by genuine scholarly innovation, ultimately constraining research diversity and originality. The aforementioned issue is amplified by non-curated bibliographic databases, such as Google Scholar, which index publications without rigorous quality control, enabling potential manipulation by authors to artificially inflate citation counts and scholarly metrics, leading to the following implications. The observations should not be read as an argument against the use of citation metrics per se, but rather against their uncritical use as stand-alone indicators in formal assessment contexts.

- Dependence on citation metrics encourages strategic publishing behaviors aimed at maximizing citations, thereby reducing incentives for high-risk, innovative, or interdisciplinary research.
- Younger researchers and/or researchers from emerging disciplines face disadvantages due to the inherent preference of citation-based metrics for senior scholars and/or scholars situated in established disciplines, which traditionally generate higher citation frequencies.
- Uniform application of citation metrics across diverse disciplines and geographical contexts aggravates inequalities, disadvantaging institutions and researchers in regions with lower citation frequencies or less journal indexing coverage.
- Most critically, reliance on non-curated platforms, such as Google Scholar, allows researchers to manipulate their own citation metrics directly. Since Google Scholar indexes all types of academic output without strict quality control, researchers can artificially inflate their citation counts, potentially influencing key evaluation processes, tenure decisions, funding allocations, and even salaries if remuneration is tied to metric-based performance agreements.

5.1.3. Research Question 2.1 ("To What Extent Are Citation Metrics Used in Formal Research Evaluation Procedures Specifically in the Civil Engineering Domain?")

Citation metrics significantly shape formal evaluation procedures in civil engineering, profoundly affecting tenure decisions, faculty assessments, and funding allocations. Civil engineering evaluations explicitly emphasize multi-authorship indices and collaborative outputs more prominently compared to other disciplines, reflecting the inherently cooperative nature of engineering projects. However, such practices tend to overlook critical aspects unique to civil engineering, such as industry engagement periods that often reduce academic publishing activity. The following implications are derived. The implications should be interpreted in light of the specific characteristics of civil engineering and should not be generalized uncritically across all engineering or scientific disciplines.

- As in other disciplines, overemphasis on citation metrics directs researchers toward citation-rich topics rather than innovative, interdisciplinary, or practically impactful civil engineering research, limiting overall research diversity and quality.
- Researchers with substantial industry experience, crucial in civil engineering, may experience disadvantages if citation metrics are weighted too heavily in metric-centric evaluations, as periods spent in professional practice negatively impact the publication output.

- Citation-based evaluations systematically undervalue practically relevant research fields and niche areas within civil engineering, potentially impacting funding decisions, career opportunities, and the attractiveness of practice-oriented careers.
- A comparatively higher weighting of multi-authorship metrics may disadvantage individuals or smaller collaborative groups, particularly in less frequently cited but practically significant research.

5.1.4. Research Question 3 ("How Widely Are Large Language Models Being Used to Generate Scientific Papers?")

The adoption of large language models (LLMs) in academic publishing has substantially reshaped scholarly writing practices, ranging from manuscript drafting and literature reviews to technical tasks such as citation formatting. While LLMs streamline scientific workflows, notably for early-career researchers and researchers from non-native English-speaking regions, considerable concerns emerge regarding authenticity, scholarly rigor, and research ethics. The rapid proliferation of artificially generated academic content, including fabricated citations, entirely fictitious papers, and widespread undisclosed usage of AI-generated texts, exposes critical vulnerabilities in established editorial procedures and peer-review standards, fundamentally challenging traditional notions of academic integrity and authorship. Furthermore, the ease of manipulating personal citation metrics through AI-generated publications increases significantly, enabling researchers to inflate performance-based metrics artificially, which could directly influence salary levels if remuneration is tied to metric-based performance targets, thus entailing the following implications. At the same time, the observations should not be interpreted as rejecting the use of LLMs in scholarly work altogether, but rather as highlighting the need for transparency, verification, and human accountability.

- Despite the evident advantages, increased reliance on LLMs in scholarly communication risks diluting research quality and eroding trust in published academic literature, exacerbating existing problems, such as predatory publishing.
- Although democratizing academic productivity and accessibility, particularly for researchers from non-native English-speaking regions, widespread use of LLMs without proper disclosure compromises transparency, accountability, and intellectual rigor in scholarly discourse.
- Established norms of authorship and attribution face disruption due to the emerging practice of explicitly acknowledging generative AI, thereby creating ambiguity regarding responsibility and genuine scholarly contribution in published research.

5.1.5. Research Question 4 ("How Easily Can Papers Generated with the Assistance of Large Language Models Influence Author-Level Citation Metrics on Google Scholar?")

The proof-of-concept case study described in Section 4 provides a platform-specific demonstration that author-level citation metrics on Google Scholar may change after two LLM-assisted, non-peer-reviewed documents are indexed under the tested conditions. However, the changes should be interpreted cautiously, as the case study is limited to one platform, one target profile, and two documents and is not designed to establish broader causal effects beyond this specific setup. The relevance of this finding lies less in the magnitude of the observed metric changes than in showing that citation-based indicators on non-curated platforms may be sensitive to strategically disseminated, weakly curated documents. This implication is particularly important where citation-based indicators are used in evaluation contexts without sufficient attention to the provenance and curation level of the underlying bibliographic data. The observations should not be generalized beyond Google Scholar or beyond the specific proof-of-concept setup examined in this study.

- The simplicity of artificially enhancing citation metrics through AI-generated content poses fundamental risks to the integrity and fairness of citation-based academic evaluations.
- Bibliographic databases operating without or with limited quality control, such as Google Scholar, are highly susceptible to citation manipulations, a vulnerability further aggravated by the free online access to LLM-based chatbots.
- The minimal effort and limited expertise required to produce apparently legitimate scholarly articles significantly lower the barriers for citation metric manipulation.

The implications discussed across the research questions are consistent with several established frameworks on the responsible use of citation metrics and research evaluation, including the Leiden Manifesto (Hicks et al., 2015), the San Francisco Declaration on Research Assessment (DORA, 2012), the CoARA Agreement (CoARA, 2022), and the Metric Tide report (Wilsdon et al., 2015), all of which emphasize the need to avoid narrow reliance on single quantitative indicators and to assess scholarly contributions in their broader institutional, disciplinary, and societal context. Likewise, emerging guidance on the responsible use of generative artificial intelligence in scholarly work, including the ERA Forum's Living Guidelines (European Commission, 2024) and recent publisher policies, underscores the importance of transparency, accountability, and verification. The findings of this study align with and reinforce these recommendations, particularly with regard to the risks associated with the uncritical use of citation-based indicators and with insufficiently curated bibliographic indexing environments.

5.2. Recommendations

The foregoing discussion shows that the central problem is not the existence of citation metrics as such, but their vulnerability to manipulation, their uneven interpretation across contexts, and their exposure to low-curation indexing environments. Based on the interpretation of the findings discussed in Section 5.1, and on the evidence synthesized in Sections 3 and 4, the following recommendations are proposed to address citation-metric manipulation within the publishing and indexing ecosystem.

Researchers

1. Adhering to ethical citation practices: Researchers should strictly comply with ethical citation standards, transparently disclose self-citations, and clearly acknowledge the contributions of each author. Additionally, researchers should actively avoid participation in citation cartels or other manipulative citation schemes, documented in Section 3.
2. Disclosing and verifying AI usage: Any use of large language models in research or manuscript preparation should be explicitly disclosed. Researchers should thoroughly verify all AI-generated content—particularly references, empirical data, and methodological details—to ensure accuracy and prevent dissemination of fabricated and/or false information that may propagate into bibliographic indexing systems.
3. Prioritizing quality over metrics: Research quality and real-world relevance should take precedence over strategies aimed solely at maximizing citation counts. Particularly in applied research disciplines, such as civil engineering, researchers are encouraged to document practical industry engagements and collaborative efforts, demonstrating genuine impact beyond purely bibliometric indicators used in formal evaluation frameworks (Section 3.5.2).

Academic institutions and committees

4. Diversifying research assessment: Academic institutions and committees should reduce their reliance on purely quantitative bibliometric indicators, particularly on indicators derived from non-curated bibliographic databases without or with limited

quality control, such as Google Scholar, whose platform-specific sensitivities are demonstrated in Section 4. Evaluation procedures should shift toward more diverse assessments, emphasizing long-term research impact, innovation, interdisciplinary collaboration, and tangible societal benefits.

5. Recognizing industry and interdisciplinary work: Evaluation frameworks should explicitly recognize and fairly assess industry experience, practical achievements, and interdisciplinary engagement of researchers, which is particularly relevant in applied disciplines, such as civil engineering, where industry collaborations significantly contribute to research impact yet may be inadequately represented by traditional citation metrics.
6. Establishing ethics and transparency policies: Institutions should implement clear policies addressing unethical citation practices, with defined sanctions for citation manipulation, such as excessive self-citation or organized citation exchanges, identified in the systematic review (Section 3.5.1). Policies should mandate transparency regarding generative AI usage in scholarly work and evaluation materials, ensuring ethical and openly disclosed use.

Publishers and editors

7. Strengthening editorial guidelines: Stringent peer-review and editorial guidelines should be established to prevent unethical citation behavior and the dissemination of AI-generated content. Measures should identify and mitigate coercive citation suggestions and irregular reference structures consistent with manipulation mechanisms summarized in Section 3.
8. Implementing clear policies on AI and authorship: Publishers should define explicit guidelines concerning the appropriate use of generative AI tools in research and manuscript preparation. To uphold trust in authorship attribution and scholarly content integrity, guidelines should specify methods for acknowledging AI assistance (e.g., in an acknowledgments section) and clearly reaffirm that all listed authors must be accountable human contributors.
9. Enhancing AI-awareness in the review process: Editors and reviewers should receive training and develop expertise to better recognize and critically assess AI-generated content during manuscript review. Editorial workflows should integrate verification procedures proportionate to documented risks of reference hallucination and citation inflation.

Providers of bibliographic databases

10. Detecting citation anomalies: Providers of bibliographic databases and indexing platforms should (further) develop and refine automated systems to identify manipulated citation patterns, counterfeit references, and papers potentially produced by paper mills or AI-assisted pipelines. Algorithms should monitor abnormal citation clustering and sequential indexing effects similar to those demonstrated in Section 4.
11. Flagging questionable content: Upon detecting manipulative practices or AI-generated content, platforms should clearly flag affected records, which is particularly pertinent for platforms without rigorous editorial oversight (e.g., Google Scholar or certain preprint servers). Explicit markers or warnings on suspicious articles should alert users about potential reliability issues.
12. Ensuring transparency in indexing: Bibliographic databases should adopt enhanced transparency measures, introducing clear labeling or markers indicating AI-generated content or records with irregular citation patterns. Providing explicit metadata regarding indexing processes strengthens the transparency and integrity of citation data and supports user trust in bibliometric evaluations.

Funding institutions and policymakers

13. Rewarding long-term impact over short-term metrics: Funding institutions and research policymakers should revise evaluation criteria to reward genuine innovation, long-term impact, collaboration, and societal value, rather than emphasizing high bibliometric scores; greater emphasis should be placed on sustained contributions of research projects, rather than on immediate citation counts or journal impact factors.
14. Establishing policies against metric gaming: Explicit guidelines should discourage strategic manipulation of citation metrics in grant applications and evaluations. Funding institutions should revise assessment frameworks to prioritize qualitative indicators of research excellence (including originality, rigor, and societal significance) over publication quantity or citation volume, to signal that attempts to "game the system" will not confer funding advantages.
15. Encouraging AI transparency and ethical standards in proposals: Funding institutions should mandate transparent disclosure concerning the use of generative AI tools throughout the research lifecycle, including proposal preparation, peer-review processes, and research execution. Applicants should be required to openly report AI involvement in preparing proposals or analyzing data and to explain how such use complies with ethical standards and rigorous research methodologies. In addition, funding institutions should establish mechanisms for monitoring unethical practices in evaluation-relevant submissions.

6. Summary and Conclusions

Generative artificial intelligence, particularly large-language-model-based chat bots, substantially increases the vulnerability of citation metrics to manipulation in academic evaluation. While citation-based indicators, such as the h-index, i10-index and total citation counts, have historically been susceptible to manipulation through self-citation and strategic behaviors, generative AI introduces qualitatively new vulnerabilities due to its unprecedented scale and automation capabilities. This study, with emphasis on the situation in the civil engineering domain, has systematically reviewed existing research on metric manipulation and has further validated critical concerns through a proof-of-concept case study. The empirical results indicate that, under the platform-specific conditions, indexing of LLM-assisted, strategically cited documents on Google Scholar is associated with increases in author-level citation metrics. It must be noted that the findings should be interpreted as a proof-of-concept demonstration within a bounded setup rather than as evidence of broader causal effects across platforms or evaluation systems.

The main implications drawn from this study are summarized as follows: Citation metrics are increasingly vulnerable to manipulative practices, significantly amplified by the proliferation of generative artificial intelligence, thus compromising the integrity of research evaluations. A critical vulnerability identified in this study is the possibility for researchers to manipulate their own citation metrics directly, which becomes particularly critical when evaluation committees rely on non-curated bibliographic databases (such as Google Scholar) that index all content without or with limited quality control. Researchers may artificially inflate their citation counts, potentially influencing critical decisions and even salary levels if remuneration is tied to performance metrics. Moreover, citation metrics can also be manipulated externally, i.e., by third parties, through uploading fabricated papers that massively reference a specific "target researcher", allowing third parties to distort the metrics of other researchers without their knowledge, potentially leading to unjust accusations of misconduct. Alarmingly, bibliographic databases might fail to detect manipulation patterns, as continuous uploads of fabricated papers do not cause suspicious spikes in the citation trajectory, rendering such manipulations difficult to detect. For ethical

reasons, this study deliberately refrains from providing procedural guidance that could facilitate replication or scaling of citation-metric manipulation.

The aforementioned findings are particularly pertinent to the civil engineering domain, where industry engagement periods can lead to reduced publication outputs, disadvantaging experienced researchers under current metric-driven evaluations. Moreover, the strong weighting of multi-authorship metrics and collaborative outputs in civil engineering may unintentionally disadvantage individuals and smaller research groups, particularly those working in practically significant but less citation-intensive areas.

Based on the results of this study, a set of recommendations has been presented to guide researchers, academic institutions and committees, publishers and editors, bibliographic database providers as well as funding institutions and policymakers. To safeguard research integrity and to prevent citation manipulation, the recommendations emphasize the implementation of rigorous quality assurance measures without reliance on non-curated bibliographic databases (such as Google Scholar), increased transparency regarding the use of generative AI, diversification of evaluation procedures beyond citation-based metrics, and the establishment of clear ethical guidelines for all groups involved.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/publications14020023/s1, Figure S1: PRISMA-style flow diagram of the modified review procedure, including separate selection processes for each research question and the contribution of snowballing; Table S1: Final search strings applied for each research question.

Funding: This study is thematically aligned with research projects funded by the German Research Foundation (DFG) under grants SM 281/9-3, SM 281/22-1, SM 281/30-1, SM 281/31-1, SM 281/32-1, SM 281/33-1, SM 281/44-1, and GRK 3068 as well as by the German Federal Ministry of Transport (BMV) under grant 01FV2059C, which provided the broader research context for this work. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of the aforementioned sponsors.

Data Availability Statement: Citation data used in this study were retrieved from publicly accessible repositories and Google Scholar profiles. The papers written within the scope of the proof-of-concept case study are publicly available for download through the links provided in the reference list. No additional datasets were generated or deposited.

Acknowledgments: Preliminary results of this study have been published in condensed form at the International Conference on Computing in Civil and Building Engineering (ICCCBE), to initiate a discussion within the civil engineering community. The present manuscript constitutes the complete version of the study and includes the background on large language models and generative artificial intelligence, the full systematic review, the complete set of research questions, the detailed methodological framework, and the extended discussion of implications.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
LLM	Large language model
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
RQ	Research question
XAICE	Explainable artificial intelligence in civil engineering

References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2012). Revisiting the scaling of citations for research assessment. *Journal of Informetrics*, *6*(4), 470–479. [CrossRef]

Abramo, G., D'Angelo, C. A., & Grilli, L. (2021). The effects of citation-based research evaluation schemes on self-citation behavior. *Journal of Informetrics*, *15*(4), 101204. [CrossRef]

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, *9*(1), 1–17. [CrossRef]

Ali, I., Sultan, P., & Aboelmaged, M. (2021). A bibliometric analysis of academic misconduct research in higher education: Current status and future research opportunities. *Accountability in Research*, *28*(6), 372–393. [CrossRef]

Ali, N., Halim, Z., & Hussain, S. F. (2023). An artificial intelligence-based framework for data-driven categorization of computer scientists: A case study of world's top 10 computing departments. *Scientometrics*, *128*(3), 1513–1545. [CrossRef]

Altanopoulou, P., Dontsidou, M., & Tselios, N. (2012). Evaluation of ninety-three major Greek university departments using Google Scholar. *Quality in Higher Education*, *18*(1), 111–137. [CrossRef]

Asfour, O. S., & Al-Qawasmi, J. (2024). Research metrics in architecture: An analysis of the current challenges compared to engineering disciplines. *Publications*, *12*(4), 50. [CrossRef]

Besançon, L., Cabanac, G., Labbé, C., & Magazinov, A. (2023). Sneaked references: Cooked reference metadata inflate citation counts. *Journal of the Association for Information Science and Technology*, *74*(7), 699–712.

Cabezas-Clavijo, Á., Magadán-Díaz, M., Rivas-García, J. I., & Sidorenko-Bautista, P. (2024). This book is written by ChatGPT: A quantitative analysis of ChatGPT authorships through Amazon.com. *Publishing Research Quarterly*, *40*(2), 147–163. [CrossRef]

Camp, N. T., Bengtson, J. A., & Sandstrom, J. C. (2025). The citation catastrophe: Propagation of AI-generated counterfeit citations in scholarship. *The Journal of Academic Librarianship*, *51*(4), 103065. [CrossRef]

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113–124. [CrossRef]

CoARA (Coalition for Advancing Research Assessment). (2022). *Agreement on reforming research assessment*. Available online: https://coara.eu/agreement/ (accessed on 27 March 2026).

Dehnad, A., Abdekhoda, M., & Fallah Atatalab, F. (2019). H-index and promotion decisions. *Annals of Library and Information Studies*, *66*(4), 171–175.

DORA (San Francisco Declaration on Research Assessment). (2012). *San Francisco declaration on research assessment*. Available online: https://sfdora.org/read/ (accessed on 27 March 2026).

Dragos, K. (2025a, May 5). *The impact of generative artificial intelligence on a metric-driven academic system—A review*. 7th International Workshop on Explainable Artificial Intelligence in Civil Engineering (XAICE), Hamburg, Germany. Available online: https://smarsly.wordpress.com/wp-content/uploads/2025/06/dragos2025b.pdf (accessed on 27 March 2026).

Dragos, K. (2025b, May 5). *A review of the h-index in the era of generative artificial intelligence*. 7th International Workshop on Explainable Artificial Intelligence in Civil Engineering (XAICE), Hamburg, Germany. Available online: https://smarsly.wordpress.com/wp-content/uploads/2025/06/dragos2025c.pdf (accessed on 27 March 2026).

El-Adaway, A. G., Assaad, R., Elsayegh, A., & Abotaleb, I. S. (2019). Analytic overview of citation metrics in the civil engineering domain with focus on construction engineering and management specialty area and its subdisciplines. *Journal of Construction Engineering and Management*, *145*(10), 04019060. [CrossRef]

European Commission. (2024). *Living guidelines on the responsible use of generative AI in research*. Available online: https://research-and-innovation.ec.europa.eu (accessed on 27 March 2026).

Finkel-Gates, A. (2025). ChatGPT in academic assessments: Upholding integrity. *Journal of Learning Development in Higher Education*, (36). [CrossRef]

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. *GigaScience*, *8*(6), giz053. [CrossRef]

Fister, I., Jr., Fister, I., & Perc, M. (2016). Toward the discovery of citation cartels in citation networks. *Frontiers in Physics*, *4*, 49. [CrossRef]

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. [CrossRef]

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLoS ONE*, *12*(12), e0187394. [CrossRef]

Goyanes, M., Lopezosa, C., & Piñeiro-Naval, V. (2025). The use of artificial intelligence (AI) in research: A review of author guidelines in leading journals across eight social science disciplines. *Scientometrics*, *130*(7), 3725–3741. [CrossRef]

Guraya, S. Y., Norman, R. I., Khoshhal, K. I., Guraya, S. S., & Forgione, A. (2016). Publish or Perish mantra in the medical field: A systematic review of the reasons, consequences and remedies. *Pakistan Journal of Medical Sciences*, *32*(6), 1562–1567. [CrossRef]

Haddow, G., & Hammarfelt, B. (2019). Quality, impact, and quantification: Indicators and metrics use by social scientists. *Journal of the Association for Information Science and Technology*, *70*(1), 16–26. [CrossRef]

Haider, J., Söderström, K. R., Ekström, B., & Rödl, M. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review*, *5*(5). [CrossRef]

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*, 429–431. [CrossRef]

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, *8*(1), 4. [CrossRef] [PubMed]

Hosseini, M., Resnik, D. B., & Holmes, K. (2023). The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics*, *19*(4), 449–465. [CrossRef] [PubMed]

Ibrahim, H., Liu, F., Zaki, Y., & Rahwan, T. (2025). Citation manipulation through citation mills and pre-print servers. *Scientific Reports*, *15*(1), 5480. [CrossRef]

Kazakis, N. A. (2014). Bibliometric evaluation of the research performance of the Greek civil engineering departments in National and European context. *Scientometrics*, *101*(1), 505–525. [CrossRef]

Kendall, G. (2024). More transparency is needed when citing h-indexes, journal impact factors and citescores. *Publishing Research Quarterly*, *40*(1), 80–99. [CrossRef]

Kendall, G., & Teixeira da Silva, J. A. (2024). Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing*, *37*(1), 55–62. [CrossRef]

Kobak, D., González-Márquez, R., Horvát, E.-A., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, *11*(27), eadt3813. [CrossRef]

Kojaku, S., Livan, G., & Masuda, N. (2021). Detecting anomalous citation groups in journal networks. *Scientific Reports*, *11*(1), 14524. [CrossRef]

Labbé, C. (2010). Ike Antkare one of the great stars in the scientific firmament. *International Society for Scientometrics and Informetrics Newsletter*, *6*(2), 48–52.

Labbé, C., & Labbé, D. (2013). Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? *Scientometrics*, *94*, 379–396. [CrossRef]

Lendvai, G. F. (2025). ChatGPT in academic writing: A scientometric analysis of literature published between 2022 and 2023. *Journal of Empirical Research on Human Research Ethics*, *20*(3), 131–148. [CrossRef]

Lim, B. H., D'Ippoliti, C., Dominik, M., Hernández-Mondragón, A. C., Vermeir, K., Chong, K. K., Hussein, H., Morales-Salgado, V. S., Cloete, K. J., Kimengsi, J. N., Balboa, L., Mondello, S., dela Cruz, T. E., Lopez-Verges, S., Sidi Zakari, I., Simonyan, A., Palomo, I., Režek Jambrak, A., Germo Nzweundji, J., . . . Bueso, F. (2025). Regional and institutional trends in assessment for academic promotion. *Nature*, *638*(8050), 459–468. [CrossRef]

Lippi, G., & Mattiuzzi, C. (2017). Scientist impact factor (SIF): A new metric for improving scientists' evaluation? *Annals of Translational Medicine*, *5*(15), 303. [CrossRef]

López-Cózar, E. D., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. [CrossRef]

Lund, B. D., & Naheem, K. T. (2024). Can ChatGPT be an author? A study of artificial intelligence authorship policies in top academic journals. *Learned Publishing*, *37*(1), 13–21. [CrossRef]

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570–581. [CrossRef]

Marsicano, C. R., Braxton, J. M., & Nichols, A. R. K. (2022). The use of Google Scholar for tenure and promotion decisions. *Scientometrics*, *47*(4), 639–660. [CrossRef]

Mazov, N. A., & Gureev, V. N. (2019, September 2). *Detection of inappropriate types of authorship using bibliometric approaches*. 17th Conference of the International Society for Scientometrics and Informetrics, Rome, Italy.

Mehregan, M., & Moghiman, M. (2024). The unnoticed issue of coercive citation behavior for authors. *Publishing Research Quarterly*, *40*(2), 164–168. [CrossRef]

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 30. [CrossRef]

Mingers, J., O'Hanley, J. R., & Okunola, M. (2023). Using Google Scholar institutional level data to evaluate the quality of university research. *Scientometrics*, *113*(3), 1627–1643. [CrossRef]

Moustafa, K. (2016). Aberration of the citation. *Accountability in Research*, *23*(4), 230–244. [CrossRef]

Mustafa, G., Afzal, M. T., Rauf, A., & Khan, M. A. (2025). Beyond publication numbers: A novel approach to academic ranking using evolutionary programming. *Evolutionary Intelligence*, *18*(3), 62. [CrossRef]

Nature. (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use (Editorial). *Nature*, *613*, 612. [CrossRef] [PubMed]

Nicholas, D., Herman, E., Jamali, H. R., Abrizah, A., Boukacem-Zeghmouri, C., Xu, J., Rodríguez-Bravo, B., Watkinson, A., Polezhaeva, T., & Świgon, M. (2020). Millennial researchers in a metric-driven scholarly world: An international study. *Research Evaluation*, *29*(3), 263–274. [CrossRef]

Ortega, J.-L., & Delgado-Quirós, L. (2023). How do journals deal with problematic articles. Editorial. *Profesional de la información*, *32*(1), e320118. [CrossRef]

Öztürk, O., & Taşkın, Z. (2024). How metric-based performance evaluation systems fuel the growth of questionable publications? *Scientometrics*, *129*(5), 2729–2748. [CrossRef]

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. [CrossRef]

Raheel, M., Ayaz, S., & Afzal, M. T. (2018). Evaluation of h-index, its variants and extensions based on publication age and citation intensity in civil engineering. *Scientometrics*, *114*(3), 1107–1127. [CrossRef]

Ramadhan, M. A., Sutarto, S., Widodo, S., & Anisah, M. (2024). Bibliometric analysis to reveal research evolution and educational technology trends in civil engineering education. *International Journal of Learning, Teaching and Educational Research*, *23*(3), 87–110. [CrossRef]

Ramoni, D., Sgura, C., Liberale, L., Montecucco, F., Ioannidis, J. P. A., & Carbone, F. (2024). Artificial intelligence in scientific medical writing: Legitimate and deceptive uses and ethical concerns. *European Journal of Internal Medicine*, *127*, 31–35. [CrossRef]

Salman, M., Ahmed, M. M., & Afzal, M. T. (2021). Assessment of author ranking indices based on multi-authorship. *Scientometrics*, *126*(5), 4153–4172. [CrossRef]

Smarsly, K. (2026, March 23–26). *On the reliability of citation metrics in civil and building engineering in the age of generative artificial intelligence*. The International Conference on Computing in Civil and Building Engineering (ICCCBE), Taipei, Taiwan.

Tang, A., Li, K.-K., Kwok, K. O., Cao, L., Luong, S., & Tam, W. (2024). The importance of transparency: Declaring the use of generative artificial intelligence (AI) in academic writing. *Journal of Nursing Scholarship*, *56*(2), 314–318. [CrossRef] [PubMed]

Thelwall, M., & Kurt, Z. (2025). Research evaluation with ChatGPT: Is it age, country, length, or field biased? *Scientometrics*, *130*(10), 5323–5343. [CrossRef]

Tripathi, M., Sonkar, S. K., & Kumar, S. (2019). A cross sectional study of retraction notices of scholarly journals of science. *Journal of Library and Information Technology*, *39*(2), 74–81. [CrossRef]

Usman, M., Mustafa, G., & Afzal, M. T. (2021). Ranking of author assessment parameters using logistic regression. *Scientometrics*, *126*(1), 335–353. [CrossRef]

Vincent, J. (2023). *Top academic publisher bans AI-generated text in scientific papers*. The Register. Available online: https://www.theregister.com/2023/01/27/top_academic_publisher_science_bans/ (accessed on 30 June 2025).

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, *13*(1), 14045. [CrossRef]

Wang, R., Zhou, Y., & Zeng, A. (2022). Evaluating scientists by citation and disruption of their representative works. *Scientometrics*, *128*(3), 1689–1710. [CrossRef]

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. Higher Education Funding Council for England (HEFCE).

Wohlin, C. (2014, May 4). *Guidelines for snowballing in systematic literature studies and a replication in software engineering*. 18th International Conference on Evaluation and Assessment in Software Engineering, London, UK.

Yoo, J.-H. (2025). Defining the boundaries of AI use in scientific writing: A comparative review of editorial policies. *Journal of Korean Medical Science*, *40*(23), e187. [CrossRef] [PubMed]

Zerem, E., Kunosić, S., Imširović, B., & Kurtčehajić, A. (2021). Science metrics systems and academic promotion: Bosnian reality, 2021. *Psychiatria Danubina*, *33*(Suppl. S3), S371–S377.

Zhang, M., & Zhao, T. (2025). Citation accuracy challenges posed by large language models. *JMIR Medical Education*, *11*, e72998. [CrossRef] [PubMed]

Zhang, Q., Abraham, J., & Fu, H. Z. (2020). Collaboration and its influence on retraction based on retracted publications during 1978–2017. *Scientometrics*, *125*(1), 213–232. [CrossRef]

Zubiaga, A. (2024). Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence*, *6*, 1350306. [CrossRef]