Should we welcome hallucinated references?: citation integrity beyond fabricated sources

The phenomenon of hallucinated references, non-existent publications invented by generative AI tools that seem plausible, has prompted understandable concern in scholarly publishing. I argue that the prevailing response is too myopic. Hallucinated references present a serious research integrity problem, but they are also a particularly visible manifestation of a broader citation pathology that is the use of real references to support claims that the cited works do not in fact establish. The latter problem is harder to detect because it cannot be resolved by checking whether a bibliographic item exists, requiring instead semantic judgement about the relation between a citing claim and a cited source. I argue that the current scandal should therefore be used to widen, rather than narrow, editorial concern with reference integrity. The existing infrastructure and tools already address parts of this problem, from metadata checking to citation-context classification, but they do not remove the need for institutional responsibility. Thus, the response to the current crisis should aim at accountability for citation practices, rather than merely better detection of fabricated sources.

Keywords
research integrity, citation integrity, hallucinated references, generative AI, scholarly publishing

OGNJEN ARANDJELOVIĆ
University of St Andrews

Introduction

We are in the midst of an uproar concerning the state of academic publishing brought about by the phenomenon of hallucinated references: non-existent references 'invented' by generative AI tools such as ChatGPT, Claude and Gemini that seem plausible. There can be little doubt that there is cause for concern. Science is a social endeavour and an institution critically supported by trust. Hallucinated references violate that trust in multiple ways, undermining the credibility of researchers, claims and the broader process itself. This is not only an authorial problem but also a structural one, because the present scandal exposes what quality-checking work is done by journals and publishers themselves, what is offloaded to reviewers and what is de facto, if not nominally, treated as outside the realm of editorial responsibility.

The problem is particularly pernicious because generative AI systems can produce references that appear linguistically and bibliographically plausible while failing to correspond to real publications. For example, Walters and Wilder (2023) examined bibliographic citations generated by ChatGPT and found both fabricated citations and citations to real works but contaminated with errors. More recently, Topaz et al. (2026) reported a large-scale audit of fabricated citations in open-access biomedical articles. These and similar findings justify concern, highlighting that the problem is best understood not as a clerical flaw but rather a deep epistemic crack that damages the evidential scaffolding of an article and wastes the time of readers, reviewers, editors, librarians, and authors whose names may be attached to fictitious publications.

Yet, while natural, the current reaction to this epistemic injury has been overly specific. Fabricated references are a new and perhaps a particularly ugly manifestation of a longer-standing problem of references often being treated as markers of scholarly seriousness rather than as genuine and precise evidential commitments. The older and still persisting problem does not consist only in sloppiness manifested through references containing typographic errors, though that has been documented for decades, but, more seriously, includes the invocation of references as support for claims that the cited works do not in fact support (Arandjelović, 2026; Schuklenk, 2026). Hallucinated references deserve attention, but they also point to a broader problem and a poignant question: Why have publishers, journals, and reviewers been more ready to notice non-existent references than references that exist but fail in their evidential role?

'Fabricated references are a new and perhaps a particularly ugly manifestation of a longer-standing problem'

Citation problems prior to generative AI

Citation inaccuracy was a recognized publishing concern long before large language models entered the scene. For example, in a study of medical journals, de Lacey et al. (1985) found that the original author was misquoted in 15% of references examined and that citation errors occurred in 24% of cases, with some errors serious enough to prevent a straightforward identification of the actual source. Similarly, Evans et al. (1990) described quotation and reference inaccuracy in surgical journals as a continuing peer-review problem.

The evidential contribution of this earlier literature is central to the argument of the present article because it helps to separate two kinds of failure that are often conflated. One is bibliographic failure, concerning references with incorrect volume number, author initial, page range, title, DOI or journal name. These errors are irritating and sometimes consequential, for example, when they are large enough to prevent a reader from finding the source. The other kind of failure is more substantive and is manifest by citations that point to real sources but misrepresent what that source says; whether by overstating its evidential significance, or by using it to support a claim it does not establish. This is often referred to as a quotation error, though the term should not mislead us into thinking that the problem is limited to literal quotations. The bottom line is that the issue here is the relation between the citing claim and the cited source.

Mogull (2017) reviewed studies of quotation errors in medical research articles and recalculated the rate of incorrectly cited assertions after applying more consistent definitions. The resulting estimate was lower than that from some earlier summaries, though still troubling, with the estimated quotation error rate of 14.5%, and most content errors consisting of major errors in which the cited source failed to substantiate, was unrelated to, or contradicted the relevant assertion. Smith and Cumberledge (2020) extended this line of inquiry to high-impact general science journals and framed quotation errors in precisely the way that supports my present argument, as cases where a reference fails to substantiate the claim for which it is cited.

The same general concern is visible in work on citation copying and citation distortion. Simkin and Roychowdhury (2003)'s wonderfully embarrassing article, with its intentionally admonitory title 'Read before you cite!', used patterns of repeated citation misprints to estimate that only about one fifth of authors had read the article they cited. The exact figure may be contestable, but the broader point is clear: citation copying without serious acquaintance with the cited work precedes ChatGPT and the current panic and is probably old and established enough to qualify as a scholarly tradition. Greenberg (2009)'s *BMJ* analysis of citation distortion in biomedicine presents a troubling picture, showing how citation networks can help create unfounded authority for a claim through bias and amplification. Repeated citation, even of an unsupported or incorrect claim, can create the appearance of evidential convergence. This past record extends beyond mere historical interest and bears directly on the present debate. It is tempting to idealize

'citation copying without serious acquaintance with the cited work precedes ChatGPT'

the past and imagine that academic publishing had a generally functioning citation culture and practice that has been particularly starkly corrupted by generative AI. My point is that this rosy picture is fictional itself. The generative-AI scandal has made one specific kind of reference corruption easier to see because the fabricated reference is often, at least in principle, checkable as fabricated. But the deeper problem has been present for a long time, and it concerns the ordinary scholarly relationship between claims, sources, readers and institutions.

Diagnosis and where it goes wrong

The current diagnosis of the problem is too myopic and too focused on the specifics of the case, missing the more serious problem that has lain insufficiently noticed for too long. If understood properly, the proliferation of hallucinated references could turn out to be helpful in exposing the deeper and older citation pathology: the widespread use of real references cited as support for claims that they in fact do not provide.

'the proliferation of hallucinated references could turn out to be helpful in exposing the deeper and older citation pathology'

To be clear, I am not referring to the familiar small errors in bibliographic metadata, irritating though these are to all involved: editors, reviewers and readers alike. Rather, I am talking about the more consequential practice in which an author cites an existing article for a proposition which that article does not assert, establish or even address in the relevant sense (dishearteningly, as an author, I note that many references to my own work fit this mould).

One reason why existing but misused references pose a distinct and more challenging problem than hallucinated ones is clear: the latter are generally easier to detect and to verify as problematic. It is this legibility of hallucinated references that has led to the widespread awareness of the issue. The importance of legibility is exemplified by the 2026 *Lancet* analysis by Topaz et al. (2026), which used an automated verification system to audit nearly 2.5 million open-access biomedical articles in PubMed Central and identified 4,046 fabricated references across 2,810 articles. The larger scale of this audit is made possible by the relative tractability of checking whether a reference exists. A robust automated database search, a DOI check when available, or a suspicious reviewer will go a long way. The identification of a real article as evidentially misused is a much more demanding task: the citing claim must be read and understood, as must the cited work itself, and a judgement made about whether the source actually bears the burden placed on it. This is time-consuming and non-trivial. The broader problem is thus not new – what AI has done is create a new manifestation of it. It is true that the convenience and speed at which AI can generate content, coupled with often perverse institutional incentives, may mean that AI will fabricate references at a scale that human negligence could not match. But crude scale is a poor metric of epistemic danger. A failure that can usually be detected mechanically is different from one that demands expert judgement to expose.

The sweeping changes generative AI has already effected, the many more that are likely to follow and the uncertainty of what even the near future will look like make the focus on AI understandable. Yet, this focus is largely misplaced, and the crucial issue is not distinctively AI driven. The bottom line is that the epistemic problem is not exhausted by checking if a bibliographic entry exists – existence is a low bar.

There is a revealing asymmetry in our reaction to hallucinated references. The recent studies and commentaries show that the panic is not merely a function of visibility. We are scandalized when a machine invents a reference and an author fails to notice. We call it lazy and irresponsible. Yet, when a human author uses a real reference as credibility adornment or tribal signalling, apparently without having checked what it actually supports, we seem less bothered. The failure of AI is treated as alien and threatening in a way that familiar human negligence is not.

'The failure of AI is treated as alien and threatening in a way that familiar human negligence is not'

Existing remedies and the limits of their reach

Previous concerns over citation integrity have not been without a response. Different partial remedies have been proposed and implemented, and an examination of their effects is instructive. In short, academic publishing has become reasonably good at checking some aspects of references – such as their existence and metadata accuracy – while remaining poorly equipped to address the harder question of evidential correctness.

At the bibliographic level, scholarly publishing is much more robust than it once was. Persistent identifiers, metadata registries, reference-linking infrastructures and automated checks have made it easier to determine whether a cited item exists and whether its metadata are coherent. Crossref – described by Hendricks et al. (2020) as community-owned scholarly metadata infrastructure, collecting and curating metadata that underpin reference linking and the identification of scholarly objects – is a central and valuable part of this institutional landscape, supporting DOI resolution, reference linking, metadata enrichment and automated detection of some forms of bibliographic mismatch. It is at this level that hallucinated references are most naturally addressed. It would be hard to justify journals and publishers not making greater use of such tools, especially given that reviewers are already overburdened and are not well placed to perform clerical verification of every item in a bibliography (Schuklenk, 2026). But bibliographic verification has an obvious semantic limit: it cannot detect epistemic misuse of references, one that turns on the methodology, the data, the interpretation of results, or on whether a cited article reports original research, provides commentary, synthesizes prior work, or offers speculation.

There have been some attempts to move beyond traditional bibliographic checking. Nicholson et al. (2021), for example, describe scite as a citation index that displays the textual context around citations and classifies citation statements as supporting, contrasting or mentioning the cited work. In this manner, the tool helps make the epistemic status of references visible, thus making it a valuable asset for literature discovery, evaluation and research synthesis. Yet, its remedial reach remains limited. For example, a classification system may tell us that a reference is presented as supportive but it cannot check that this presentation is correct: the cited work may not support the claim, or it may support a similar but weaker claim, or it may be methodologically flawed and itself epistemically compromised. Recent work has begun to address the problem more directly. Sarol et al. (2024) focus on citation integrity in biomedical publications and ask whether natural language processing can help detect quotation errors at scale. Their work is important because it recognizes and attempts to address the deeper citation problem I highlight in the present article, as well as because it shows why the problem has been difficult to address institutionally: reliable checking requires access to full text, carefully and expertly labelled training data and judgement about what counts as adequate support.

These limitations explain why previous remedies have not succeeded in eliminating the problem. Some proposed remedies rely on authors reading what they cite; they should, but the obligation is already known. Some rely on reviewers noticing bad citations; they sometimes do, but reviewing is unpaid, little-acknowledged and time-constrained. Some rely on copy-editors; they can check formatting and metadata but cannot realistically inspect the semantic relation between every citation and every cited work – not only because of the labour involved, but also because of the expertize required. Automated tools can help at both the bibliographic and basic semantic levels, but not at the deep, epistemic level. The result is an institutional gap, with everyone agreeing that references should be accurate and responsible, and with everyone having some reason to think that someone else is doing the required work. In practice, the checking is partial, variable and concentrated on checks that are easiest to perform. And this is precisely why hallucinated references are useful, as a catalyst of a scandal that has exposed the first layer of the problem in a form that is embarrassing and difficult to ignore, and that should help reveal the less visible but more pernicious layers underneath.

‘The result is an institutional gap … with everyone having some reason to think that someone else is doing the required work’

Potential solutions: institutional, not personal change

The obvious response is to insist that authors should read what they cite, reviewers should check what is cited and editors should exercise greater oversight. All this may very well be true and still be perfectly useless for effecting change. Authors already know what they ought to do. As do reviewers. As do editors. The solution, if any exists, is to be found in institutional practices which make bad citation more visible and more costly.

Here is a concrete possible proposal. The first and crudest layer of veridicality checking would be straightforward in that it can be performed purely mechanically. Publication venues (journals, conferences) could perform checks that the cited works exist and that the associated metadata are coherent (e.g., consistency of DOIs, titles, authors, venues and dates). There is little reason not to do this and to instead leave such menial work to reviewers and, to a lesser extent, editors who are already pressured for time and overworked. This first step would not be able to ensure citation integrity, but it would remove an avoidable burden and catch many of the crudest citation failures. The second layer would be semantic. Here we find a good case where modern AI can actually be useful, though, as ever, it should not be treated as an oracle. A system could compare the citing sentence and its context with the content of, where available, the full text of the cited work, and flag cases in which the evidential link appears questionable. There would be errors, of course, but these need not be unduly epistemically costly if the system is used only to identify cases for further human scrutiny, the aim thus being not unattainable perfection but triage. Since subjecting every citation to full semantic analysis would be unnecessary and unworkable, journals could sample citations with the surrounding context for human scrutiny, prioritizing content that is central to the work.

The accountability aspect of my proposal is worth stressing because without appropriate accountability, the consequent outcome will not be meaningful change but a reshaped performance of compliance. Thus, authors could be asked, as they indeed sometimes already are, to declare that all references have been checked or to promise that no generative AI was used in specific ways. Such declarations may have some value, but they are only weakly able to address the broader epistemic problem I have highlighted.

Indeed, the history of citation inaccuracy suggests that the problem is not primarily in the ignorance of rules but in structure-mediated behavioural adaptation that balances factors such as speed and incentives against perceived risk, including the tacit belief that citation is a low-risk part of academic writing. The obvious response is to alter the balance of this structure by increasing the probability of detection and the severity of consequences. If a reference is found to be problematic after inspection by a human, the follow-up should fit the nature of the discrepancy. At the same time, we should not be overly hasty and overreact – a single questionable citation may be readily explainable by a simple slip, whereas the presence of many such references becomes informative about the author's relationship to evidence and should weigh on the strength of the institutional response. Thus, the author may be required to add clarification, change the work cited or otherwise revise the manuscript. In serious or repeated cases, the matter may warrant formal editorial sanctions or referral for academic misconduct.

Ultimately, the problem is institutional rather than one concerning individual research ethics alone. The question of citation integrity sits at the nexus of writing norms and practice, peer-review culture, editorial diligence and policy and technology, both publishing technology and technology in a broader sense. To rely purely on personal virtue or to treat the problem as a purely technical one would be a serious misunderstanding of the institutional ecology. Instead, the more practicable route is to demand layered responsibility, with authors being answerable for what they cite, reviewers asked to challenge suspicious references and editors and publishers expected to provide the infrastructure that makes the easiest failures easy to catch and the harder failures more likely to surface.

'The question of citation integrity sits at the nexus of writing norms and practice, peer-review culture, editorial diligence and policy and technology'

Conclusion

The ongoing hallucinated references scandal may be precisely the kind of shock that academic publishing needs – and that it has needed for a while. For this shock to yield positive change, our efforts should not be wasted on a narrowly AI-centred response. So, should we welcome hallucinated references? In that they are scholarly failures, certainly not. But we may have reason to welcome the fact that, by virtue of their crudeness, they have made visible the wider problem that was easier to ignore when it appeared in a more respectable form. The particularly pernicious references are those that exist and seem trustworthy but surreptitiously fail to support the claim for which they have been enlisted. If the present scandal merely gives us better bibliography checkers, we will have missed an opportunity for real change: to stop treating references as scholarly decoration and start treating them as epistemic commitments for which authors are answerable.

'stop treating references as scholarly decoration and start treating them as epistemic commitments for which authors are answerable'

Abbreviations and Acronyms

A list of the abbreviations and acronyms used in this and other *Insight*s articles can be accessed here – click on the following URL and then select the 'full list of industry A&As' link: **http://www.uksg.org/publications#aa**.

Competing interests

The author has declared no competing interests.

References

Arandjelović, O. (2026). Against moral panic and citation fiction: A critique of "panem, corticoids and circenses" and a proposal for editorial gatekeeping on reference integrity. *Bioethics*, *40*(4), 365–372. **https://onlinelibrary.wiley.com/doi/10.1111/bioe.70099** (accessed 29 September 2026).

de Lacey, G., Record, C., & Wade, J. (1985). How accurate are quotations and references in medical journals? *British Medical Journal (Clinical Research Ed.)*, *291*(6499), 884–886. **https://www.bmj.com/content/291/6499/884** (accessed 29 September 2026).

Evans, J. T., Nadjari, H. I., & Burchell, S. A. (1990). Quotational and reference accuracy in surgical journals: A continuing peer review problem. *JAMA*, *263*(10), 1353–1354. **https://jamanetwork.com/journals/jama/article-abstract/380949** (accessed 29 September 2026).

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, *339*:b2680. **https://www.bmj.com/content/339/bmj.b2680** (accessed 29 September 2026).

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), 414–427. **https://direct.mit.edu/qss/article/1/1/414/15577/Crossref-The-sustainable-source-of-community-owned** (accessed 29 September 2026).

Mogull, S. A. (2017). Accuracy of cited "facts" in medical research articles: A review of study methodology and recalculation of quotation error rate. *PLoS One*, *12*(9):e0184727. **https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184727** (accessed 29 September 2026).

Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P., & Rife, S. C. (2021). scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, *2*(3), 882–898. **https://direct.mit.edu/qss/article/2/3/882/102990/scite-A-smart-citation-index-that-displays-the** (accessed 29 September 2026).

Sarol, M. J., Ming, S., Radhakrishna, S., Schneider, J., & Kilicoglu, H. (2024). Assessing citation integrity in biomedical publications: Corpus annotation and NLP models. *Bioinformatics*, *40*(7), btae420. **https://academic.oup.com/bioinformatics/article/40/7/btae420/7699794** (accessed 29 September 2026).

Schuklenk, U. (2026). Academic integrity is coming increasingly under attack from academics using AI tools as a short-cut to publications. *Bioethics*, *40*(4), 363–364. **https://onlinelibrary.wiley.com/doi/10.1111/bioe.70100** (accessed 29 September 2026).

Simkin, M. V., & Roychowdhury, V. P. (2003). Read before you cite! *Complex Systems*, *14*(3), 269–274. **https://www.complex-systems.com/abstracts/v14_i03_a05/**

Smith, N., Jr., & Cumberledge, A. (2020). Quotation errors in general science journals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *476*(2242):20200538. **https://royalsocietypublishing.org/rspa/article/476/2242/20200538/80897/Quotation-errors-in-general-science** (accessed 29 September 2026).

Topaz, M., Roguin, N., Gupta, P., Zhang, Z., & Peltonen, L.-M. (2026). Fabricated citations: An audit across 2·5 million biomedical papers. *The Lancet*, *407*(10541), 1779–1781. **https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)00603-3/fulltext** (accessed 29 September 2026).

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, *13*(1):14045. **https://www.nature.com/articles/s41598-023-41032-5** (accessed 29 September 2026).

Article copyright: © 2026 Ognjen Arandjelović. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use and distribution provided the original author and source are credited.



Corresponding author:
Ognjen Arandjelović
University of St Andrews
United Kingdom
E-mail: ognjen.arandjelovic@gmail.com
ORCID ID: https://orcid.org/0000-0002-9314-194X

To cite this article:
Arandjelović O, "Should we welcome hallucinated references?: citation integrity beyond fabricated sources," *Insights*, 2026, 39: 22, 1–7; DOI: https://doi.org/10.1629/uksg.786

Submitted on 21 May 2026 Accepted on 30 July 2026 Published on 10 November 2026

Published by UKSG in association with Ubiquity Press.