ese / European Science Editing

Viewpoint

Guarding against artificial intelligence – hallucinated citations: The case for full-text reference deposit

Alex Glynn✉

Kornhauser Health Sciences Library, University of Louisville, Louisville, KY, United States of America

alex.glynn@louisville.edu

orcid.org/0000-0002-3027-7276

Received: 26 Mar 2025
Accepted: 2 Jul 2025
Published: 31 Jul 2025

Funding

The author received no specific funding for this work.

Prepublication

This manuscript has been posted as a preprint on arXiv: https://doi.org/10.48550/arXiv.2503.19848.

Declaration of interests

The author is the founder and maintainer of Academ-AI (reference 3), a non-revenue generating project. The author declares no other conflict of interest, financial or otherwise.

CC BY

This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

Citation

Glynn A. Guarding against artificial intelligence – hallucinated citations: The case for full-text reference deposit. *Eur Sci Ed.* 2025;51:e153973.

https://doi.org/10.3897/ese.2025.e153973

ese / European Science Editing

Abstract

The tendency of generative artificial intelligence (AI) to 'hallucinate' false information is well known; AI-generated citations to non-existent sources have penetrated the bibliographies of peer-reviewed publications. Drawing from the Transparency and Openness Promotion guidelines, American judicial contention with generative AI, and the submission of prior art to the US Patent and Trademark Office, the author proposes that journals require authors to submit the full text of each cited source along with their manuscripts, thereby preventing authors from citing material whose full text they cannot produce. This solution requires limited additional work by authors or editors while effectively immunizing journals against hallucinated references.

Keywords:

Artificial intelligence, bibliography, citation, large language models, references

On March 4, 2024, *PLoS One* published an article by Shoukat et al comparing blended learning with traditional instruction for undergraduates in Faisalabad, Pakistan.[1] Within the same month, commenters on PubPeer raised concerns regarding the article's reference list.[2] The first clear irregularity was the extraneous phrase 'Regenerate response', which appeared in reference #62. This phrase was formerly the label of a button on the ChatGPT interface. In at least 50 instances, authors have apparently copied the label along with the chatbot's output and pasted it into their subsequently published manuscripts.[3] On further investigation into the *PLoS One* article, PubPeer commenters were unable to find sources matching several of the references, including #62, concluding that they likely did not exist.[2] Generative artificial intelligence (AI) tools based on large language models (LLMs) have a known tendency to 'hallucinate' false information,[4] particularly bibliographic references.[5]

On April 18, 45 days after publication, the article was retracted. The editors cited the phrase 'Regenerate response' as evidence of 'potential undisclosed use of an AI tool' and reported that 18 of the article's 76 references could not be identified.[1] Had their authenticity been checked during peer review or other editorial processes, the study may never have been published. However, editorial verification is an inefficient solution to this problem. Proving definitively that a hallucinated reference does not exist is, like many negative assertions, time-consuming at best and impossible at worst. Proving that a real reference does exist is typically a trivial endeavour.

An instructive parallel comes from the legal world. In 2023, two attorneys were sanctioned in the Southern District of New York for 'submitt[ing] non-existent judicial opinions with fake quotes and citations created by the artificial intelligence tool ChatGPT'; *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443, 448 (S.D.N.Y. 2023). The non-existent opinions were cited in the Plaintiff's Affirmation in Opposition. When first the Defendant's counsel and then the Court itself were unable to locate the cited decisions, the Plaintiff's counsel were instructed to provide copies of them in full text. The attorneys were able to provide only 'excerpts', also hallucinated by ChatGPT, which themselves contained further citations to non-existent cases, among various other legal and logical flaws.

Two primary factors contributed to the discovery of the hallucination in this case. Firstly, the adversarial legal system, unlike academic peer review, strongly incentivizes lawyers to find and read the full text of cases cited by the opposing counsel. Secondly, when neither the Defense, nor the Court was able to do so, the Plaintiff's counsel were required to provide the full text themselves. Had the cases actually existed, this would have been an extremely simple task.

Among the '[m]any harms' caused by the submission of fake opinions, the Court cited the following:

> The opposing party wastes time and money in exposing the deception. The Court's time is taken from other important endeavors.

Searching fruitlessly for non-existent journal articles is no better use of time for editors and peer reviewers than searching fruitlessly for non-existent case law is for courts and law firms. In academic publishing, rather than attempting to require or incentivize peer reviewers to verify bibliographic references, editors could require authors, to the extent permissible by applicable copyright law to submit the full text of each reference as part of the submission process.

Further precedent comes from the increasingly common practice of data deposit. The Transparency and Openness Promotion (TOP) guidelines provide a set of data deposit standards for journals.[6] To achieve level II or III data transparency, journals must require that original research data associated with submitted articles be deposited in a trusted repository, such as Dataverse or Dryad, in the interest of reproducibility. As of March 12, 2025, 320 journals have achieved this standard.[7]

Depositing full-text references would not be as resource-intensive as depositing research data. The full texts of cited articles need not be *publicly* available; they would only need to be shown to the editors and peer reviewers to verify their authenticity and, as an added benefit, be available for the reviewers' use while appraising the submitted article.

It should not put undue strain on authors to submit copies of their sources in full text. Since responsible researchers do not cite articles based on their abstracts alone, the citing author should already have access to the full text. For longer works, such as books and monographs, only the relevant chapters or section would need to be provided. Items available only in physical formats could be scanned or photographed. Thanks to the successful implementation of Inter-Library Loan services and other resource sharing initiatives, accessing and sharing full-text sources is largely a solved logistical problem. An example workflow using a privately shared Zotero library is described in the Supplementary Material.

One foreseeable impediment to implementing the author's recommendation is that sending the full text of closed access sources to journal editors may infringe on the copyright of those sources. In the United States, exceptions to copyright exist on the grounds of fair use (17 U.S.C. § 107), which may protect the submission of full-text references to journals; American courts have found in favour of fair use defences in similar scenarios. For example, multiple courts[1] held that iParadigms, LLC, did not violate students' copyright by archiving full-text papers submitted to the TurnItIn plagiarism detector. Specifically, the use of 'the entirety of the original works' was deemed necessary 'to be successful in [iParadigms'] plagiarism detection services'; *A.V. v. iParadigms, LLC*, 544 F. Supp. 2d 473, 483 (E.D. Va. 2008). The Court further held:

> Clearly, iParadigms' use does not amount to a 'market substitute' for Plaintiffs' works. iParadigms stores its archived papers digitally and they are not publicly accessible or disseminated in any way. (*Id.* at 473–483)

Similarly, it may be argued that the author's proposal does not involve public access to or dissemination of full-text articles and therefore does not supplant them in the market. Sharing full texts with peer reviewers as well as editors may present a market substitution issue since, as subject experts, they are likely to overlap with the articles' target markets. This measure is not essential to the proposal; however, if necessary, journals could reserve the full texts for the eyes of internal editorial staff only.

A 2013 ruling upheld the fair use defence for a law firm making copies of scientific articles (non-patent literature, or NPL) for use as evidence in patent filings; *Am. Inst. of Physics v.*

[1] The District Court's finding of fair use was upheld by the Appeals Court for the Fourth Circuit; see *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F. 3d 630 (4th Cir. 2009).

Winstead PC, Civil Action No. 3:12-CV-1230-M (N.D. Tex. 2013). The Court found that 'the Defendants' copying of NPL serves an evidentiary function' and 'facilitates an efficient patent process' and that 'it is not merely the content of the NPL article that matters, but its existence in the available literature.' Likewise, the submission of full-text references to journals would serve an evidentiary function – verifying the references' existence in the available literature – and facilitate an efficient peer-review process.

However, an international publishing industry has to function within many different legal systems, including those lacking the exceptions to copyright protections found in American law; in at least some cases, exceptions to the full-text deposit requirement will likely be necessary. Following TOP data transparency level II, journals could require that exceptions to the full-text deposit requirement be identified and justified on submission.[6] Only these exceptions would then need to be investigated by editors or peer reviewers. Even if the full text cannot be provided, journals could require that all citations include persistent identifiers or links to records in bibliographic databases or library catalogues. Such links still require validation since they can be generated easily by LLMs, but testing a link is much less time-consuming than validating a plain bibliographic reference. To improve efficiency, highly selective journals could require full-text reference deposit only for articles that are provisionally accepted for publication.

It has always been possible for unscrupulous researchers to cite non-existent references, gambling on the assumption that editors cannot or will not verify their authenticity. However, the emergence of popular and affordable generative AI tools has enabled researchers to cite non-existent references not only with unprecedented ease but even without intent. So long as AI tools capable of hallucination are in use, this problem is likely to continue. The solution proposed here would prevent authors from citing non-existent sources; to continue doing so, authors would need to fabricate entire articles or create counterfeit bibliographic databases, measures which, while certainly not impossible, lack the expediency that attracts unscrupulous authors to generative AI and certainly rule out any accidental fabrication.

References

1. PLoS One Editors. Retraction: A comparative analysis of blended learning and traditional instruction: Effects on academic motivation and learning outcomes. *PLoS One*. 2024;19(4):e0302484. Retraction: Retraction: A comparative analysis of blended learning and traditional instruction: effects on academic motivation and learning outcomes. *PLoS One*. 2024;19(4):e0302484. [CrossRef]

2. Cabanac G, Abalkina A. Comments on "A comparative analysis of blended learning and traditional instruction: effects on academic motivation and learning outcomes". *PubPeer*; 2024. https://www.pubpeer.com/publications/7F01171F32B0421DF37EE8D1B49C04 .Accessed 25 Mar 2025.

3. Academ AI. *Suspected AI*. https://www.academ-ai.info/ .Accessed 25 Mar 2025.

4. Raimondo GM, Locascio LE. Artificial intelligence risk management framework generative artificial intelligence profile [Internet]; 2024. Gaithersburg (MD): National Institute of Standards and Technology Accessed 2025 Mar 25. [CrossRef]

5. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *J Med Internet Res*. 2024;26:e52935. [CrossRef]

6. Nosek BA, Alter G, Banks GC, et al. SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*. 2015;348(6242):1422-1425. [CrossRef]
7. Stodden V, Boycan E, Mellor DT, et al. TOP resources - evidence and practices. 2016. *Open Science Framework* [Internet]; 2025. In: Charlottesville (VA): Center for Open Science:c2011-c2025; Accessed 2025 Mar 16. [CrossRef]

ease / publications

ese / European Science Editing

European Science Editing is an official publication of EASE. It is an open access peer-reviewed journal that publishes original research, review and commentary on all aspects of scientific, scholarly editing and publishing.

https://ese.arphahub.com/
https://www.ease.org.uk
https://www.linkedin.com/company/easeeditors/
https://bsky.app/profile/easeeditors.bsky.social
https://www.facebook.com/EASEeditors/
https://mstdn.science/@EASE
https://www.instagram.com/easeeditors/
https://www.youtube.com/easeeditors

© 2025 the authors. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ISSN 2518-3354

AVES

arpha