Zoodiversity, 60(3): 301–308, 2026

DOI 10.15407/zoo2026.03.301
UDC 001.81:004.8:050

PHANTOM REFERENCES AND THE PEER-REVIEW CRISIS: HOW ARTIFICIAL INTELLIGENCE TESTS THE RESILIENCE OF SCIENTIFIC PERIODICALS

V. O. Kharchenko, V. O. Korneyev *, N. S. Filimonova

Zoodiversity Journal Editorial Board

* Corresponding author

E-mail: valery.korneyev@gmail.com

V. O. Kharchenko (https://orcid.org/0000-0002-3824-2078)
V. O. Korneyev (https://orcid.org/0000-0001-9631-1038)

Phantom references and the peer-review crisis: how artificial intelligence tests the resilience of scientific periodicals. Kharchenko, V. O., Korneyev, V. O., Filimonova, N. S. — The critical vulnerability of the traditional blind peer-review system to the challenges posed by the rapid development of generative tools is examined. Based on a real-world precedent involving the discovery of a completely fabricated bibliography in a submitted manuscript, we analyse the basic mechanisms behind the creation of false references, such as anachronisms, "Frankensteinisation", and professional biases. The study demonstrates the evolution of the threat: a transition from the obvious errors of early algorithms to the deep "semantic hallucinations" of modern RAG-based search engines, which are capable of generating perfectly formatted yet conceptually empty texts derived from real databases. To protect the publication process, an updated algorithm for editorial control is proposed, requiring the mandatory validation of digital object identifiers (DOIs) and a clear declaration of the algorithms utilised by the authors. The main conclusion emphasises the necessary and unalterable transition to the Open Science Framework paradigm, where textual material is viewed merely as an accompanying document to a verified array of primary datasets, open-source code, and deposited collection specimens.

Key words: academic integrity; LLM; generative hallucinations; open science; data deposition; research falsification.

Introduction

The rapid development of large language models (LLMs) has created an unprecedented challenge for academic integrity. While artificial intelligence initially dispelled fears by acting as a convenient tool for improving academic English, editorial

boards of scientific journals today face a new, far more dangerous trend — the complete fabrication of scientific data and bibliography. The propensity of LLMs to generate plausible-sounding but fictitious references was identified as a significant failure mode shortly after their widespread release. Early specialised investigations into LLM behavior estimated that between 30% and 69% of references generated in biomedical and scientific contexts were entirely fabricated (Athaluri et al., 2023; Walters & Wilder, 2023). Recent systematic audits demonstrate that this has since escalated into a rapidly escalating global crisis: an analysis of 2.5 million biomedical papers revealed a more than 12-fold increase in the rate of fabricated references between 2023 and early 2026 (Topaz et al., 2026).

It is important to emphasize that modern Large Language Models (LLMs) are fundamentally not information-processing tools, but language-processing systems. They mimic the linguistic patterns and structures of texts—rather than the substantive factual content—found in their training datasets (Walters & Wilder, 2023). Consequently, all such models are prone to "hallucinations" — generating factually incorrect but plausible-sounding responses. Current efforts to mitigate these inherent flaws involve the integration of external knowledge retrieval (RAG) and real-time web-search capabilities to ground outputs in verifiable data. However, as the core generative engine remains probabilistic, the challenge of ensuring factual accuracy remains a persistent concern.

This article analyses a real precedent encountered by our journal's editorial board.

A year ago, the editorial board received a review article manuscript from a group of authors, dedicated to ecological studies of one of insect groups in North Africa. The processing of this manuscript revealed all the chronic diseases of modern scientific periodicals.

A severe shortage of reviewers. We faced mass refusals from specialists to review due to excessive workload.

Superficial peer review. The manuscript eventually underwent double-blind peer review. Both reviewers provided positive feedback, having spent an inadequately long time analysing the text, but, as it turned out later, performed the check purely nominally.

Editorial routine and pressure. Having received positive reviews, the article was preliminarily accepted for publication. A lengthy and meticulous proofreading stage began: text editing, checking for compliance with British English, and bringing the manuscript in line with the journal's requirements. This process was accompanied by constant pressure from the authors, who regularly bombarded the editorial office with emails demanding to expedite the publication: "What is the status of the article now? When will it be published online?"

The turning point that collapsed this house of cards was the final technical check of the reference list. The production editor's attention was drawn to the lack of working DOIs in a significant portion of the modern links. A selective manual check of the sources in journal archives yielded a shocking result: a substantial part of the cited publications did not physically exist. They turned out to be the product of artificial intelligence hallucinations (most likely early versions of GPT, used without search plugins).

A detailed analysis of the fabricated reference list allowed us to identify three main mechanisms of "hallucinations" of early versions of generative language models, which the authors uncritically integrated into their text and which every editor should know.

Anachronisms of volumes and issues. AI compiles a real journal name with random volume numbers and years. For example, the list featured the 12th volume of the journal *Entomological Research* for 2020, whereas, in reality, the 50th volume of this publication was issued in 2020. The model is unable to verify the actual publishing history.

"Frankensteinisation" of sources. The most insidious type of forgery. The algorithm takes a real journal, correct year and volume, real page ranges, and even real specialist authors, but completely invents the article title so that it perfectly fits the narrative of the manuscript. In our case, on the pages of the journal *Agronomy for Sustainable Development*, which supposedly contained a review of Orthoptera ecology, an article by completely different authors about weed control in legume crops was actually published.

Professional and spatial displacements. The model does not distinguish the narrow specialisation of scientists. In a falsified reference to *Biodiversity Journal* about Moroccan insects, the co-author turned out to be a real and very famous Moroccan professor-agronomist whose specialisation is herbology (weed science), not entomology. The specified pages of the journal actually contained works on the fauna of Italy and India.

The scale of the problem is corroborated by independent audits. Research into biomedical literature over the last three years indicates very many of references generated by large language models (LLMs) in biomedical contexts are entirely fabricated (Topaz et al., 2026). This is not merely a technical glitch but a systemic phenomenon, as these references are often impeccably formatted and attributed to real researchers, making them virtually "invisible" to traditional peer review.

Conclusions and warnings for the scientific community. Effectively, we are entering a stage where one of the primary challenges in science is no longer accessing information, but verifying whether this information actually exists. Our position is categorical: if authors falsify sources so carelessly or deliberately, there is no reason to trust their own results and primary data. They are just as likely to be fabricated. This case requires wide publicity within ethical limits. The discovery of a completely fabricated reference list inevitably casts doubt on the reliability of any results presented in the article. The manuscript was immediately rejected already at the typesetting preparation phase, as publishing such material would deal an irreparable reputational blow to the journal, despite prior approval by reviewers. This case requires publicity within ethical boundaries (without revealing the names of the authors of the unpublished manuscript), as it exposes a global problem.

The traditional blind peer-review system is built on the presumption of the author's integrity. Today, this presumption is dead. Reviewers are not prepared to spend hours fact-checking every reference, which academic fraudsters actively exploit. We

are entering an era of new challenges, where the uncontrolled use of large language models can destroy the foundation of scientific trust.

The experience of large-scale audits (e.g., scanning 2.5 million papers in PubMed Central) proves that manual bibliography verification is no longer sufficient. A viable solution lies in implementing automated reference validation systems directly into the editorial workflow, as proposed in our updated editorial control algorithm.

Addressing our fellow editors of academic journals, we urge the editorial boards of all academic journals to implement new control protocols:

Selective checks: mandatory manual verification of 3–5 sources from different parts of the reference list (especially those lacking a DOI). Traditional peer review, based on the presumption of authors' academic integrity, has proven powerless against generative AI. Reviewers, focusing on the logic of the text, no longer check the physical existence of sources, trusting properly formatted reference lists.

Mandatory DOI validation prior to peer review. Technical editors must carry out mandatory automated or many validation ofseveral sources from different parts of the reference list (especially those lacking a DOI). Manuscripts without active digital object identifiers for modern sources must be returned to the authors even before peer review begins.

Declaration of AI. The author guidelines must enshrine a strict requirement: any use of language models (even for translation or formatting) must be detailed in the Acknowledgements or Methods section, specifying exactly how the generated information was verified. The absence of such a declaration upon the discovery of AI traces should be grounds for a lifetime ban of the authors from the publication.

Briefing reviewers. Update reviewer forms by adding a point on the mandatory selective verification of 3–5 sources from the reference list regarding their physical existence.

The scientific community must acknowledge: artificial intelligence tools are an irreversible reality of science, they are already here, and ignoring them is impossible. But without the introduction of strict verification barriers at the editorial level, scientific literature risks drowning in a flood of plausible but completely empty simulacra. It depends solely on our vigilance whether a scientific article will remain the gold standard of verified knowledge or turn into a generated illusion.

A few words about the "nuts and bolts". Artificial intelligence does not feel an "unwillingness"to work and is not "lazy" due to an influx of millions of users. Models do not have consciousness to choose the path of least resistance out of fatigue, at the core of this lies a fundamental mathematical problem of the transformer architecture — Maximum Likelihood Estimation. LLMs (large language models) are trained to predict the next word so that the text looks as natural as possible. It is mathematically "easier"(i. e., requires less computational effort from the algorithm) for them to generate a perfectly smooth, syntactically correct, but fabricated article title than to search within their weights for an exact, specific, and rare fact. Furthermore, the so-called "laziness" (when a model gives short, superficial answers) is often a consequence of alignment procedures (RLHF — Reinforcement Learning from Human Feedback), where developers artificially restrict models so they do not generate unsafe content or to save computational resources (tokens) on company servers.

However, over the past year, the landscape of AI tools for scientists has changed dramatically. Modern researchers rarely use "bare" ChatGPT for literature searches, as the aforementioned authors did a year ago. Here is the current arsenal now being used in the academic environment, which brings both benefits and new threats:

Basic universal LLMs (text and code generators):

ChatGPT (OpenAI, GPT-4o models): Has web access and plugins for data analysis. Used for drafting, structuring articles, and writing scripts (Python/R) for processing statistical data.

Gemini (Google, Pro/Advanced models): Integrated with the Google ecosystem (including Google Scholar via an extension). Capable of processing vast arrays of text. Used for analysing large PDF files. Google's separate NotebookLM tool is now massively used by scientists to create personal knowledge bases from uploaded articles.

Claude (Sonnet 4.6*): Has a massive context window, allowing authors to upload a dozen articles and request a compiled review.

Specialised scientific AI search engines (RAG systems):

Perplexity AI: The most popular search tool at the moment. It does not simply generate text but searches for real sources online and compiles an answer with direct links.

Elicit, Consensus, SciSpace: Highly specialised tools trained exclusively on databases of scientific articles (Semantic Scholar, etc.). They analyse PDFs, "extract" methodology, results, and create summary tables.

Evolution of illusions: from primitive errors to perfect falsifications. It should be understood that the manuscript, which became the subject of our analysis, was created about a year ago. At that time, the authors likely used basic versions of language models without direct internet access, which led to the appearance of obvious "hallucination" markers — non-existent volumes and pages. However, over the past year, tectonic shifts have occurred in the field of artificial intelligence.

Today, the academic world is massively armed with a new generation of AI tools. Universal models (ChatGPT-4o, Gemini Advanced, Claude 3.5) with huge context windows capable of analysing dozens of uploaded articles simultaneously have been joined by specialised scientific search engines and analysers (Perplexity, Elicit, Consensus, SciSpace). They use RAG (Retrieval-Augmented Generation) technology, relying on real databases of scientific publications.

Expectedly, the technological leap did not diminish the threat to academic integrity, but merely made it more insidious. A new systemic problem has emerged: a fundamental discrepancy between how algorithms work and what users expect from them. Large language models by their nature are neither knowledge bases nor search engines — they are probabilistic text generators. Their mathematical goal is to create a syntactically perfect and statistically plausible sequence of words. When a model encounters a complex, highly specialised query (for example, searching for empirical data on local fauna), searching for the exact fact within the neural network's weights requires considerable "effort". It is mathematically easier for the algorithm to take the path of least resistance: compile an averaged, maximally plausible text that looks like

science, sounds like science, but contains no actual facts. This effect is amplified by safety and optimisation settings from the developers themselves (RLHF), which force models to provide generalised, superficial answers instead of deep analysis.

As a result, we get a perfect trap for dishonest or lazy authors: AI generates flawlessly written English text with perfect structure and actually existing references (thanks to new search tools), but the conclusions, correlations, or synthesis of ideas themselves are completely empty, "averaged" hallucinations of a higher order. Detecting such a generated review or discussion is much more difficult: the bibliography will be genuine, but the link between the cited source and the author's thesis may be completely distorted by the algorithm, which simply tailored the text to the required narrative. Accordingly, the burden of verification again falls on the reviewer, who will now have to verify not only the existence of the article but also whether it actually states what the author claims. Modern AI can perfectly imitate form, so the protection of scientific knowledge must be based exclusively on strict control of content and primary data.

New algorithm for editorial verification: how to recognise hidden generation. Since modern RAG systems (AI-based search tools) have made it almost impossible to detect fake sources by external signs, editorial boards and reviewers need to shift their focus. Falsifications are moving from the level of bibliography to the level of meanings and primary data. We propose another algorithm for verifying manuscripts to detect the uncritical or fraudulent use of the latest AI models:

1. Search for "semantic hallucinations" and the effect of "perfect emptiness" Modern LLMs generate syntactically flawless but conceptually "sterile" texts. A reviewer should look out for the following markers:

Absence of scientific conflict: AI algorithmically avoids cognitive dissonance and "inconsistencies" in hypotheses. If a review article or discussion section perfectly reconciles all sources and contains no analysis of contradictory or anomalous results, this is a marker of generated averaging.

Gap between citation and conclusion: The source is real, but it does not support the thesis put forward by the author. AI often takes a real article but "invents" a conclusion needed for the smoothness of the current narrative. The reviewer must selectively verify not the fact of the article's existence, but the essence of the conclusion drawn from it.

Methodological sterility: Generated descriptions of materials and methods often look like idealised protocols from textbooks, devoid of the specific roughness of real field or laboratory work.

2. Verification of primary materials (Technical barrier). The most vulnerable point of artificial intelligence is its inability to generate a flawlessly consistent array of raw, empirical data with real physical parameters. Therefore, editorial boards must introduce a presumption of openness for primary materials. A manuscript should not be admitted to the peer-review stage without fulfilling the following conditions:

Deposition of raw data: All measurement tables, feature matrices, and statistical samples must be uploaded to open repositories (e. g., Zenodo or Dryad) by the time the article is submitted. Journals must abandon the practice of "data available upon request".

Digital and physical footprint: For biological research, the provision of

accurate geographical metadata is required. For instance, collection coordinates must be provided in a strictly standardised format, such as [0.717° S, 77.567°W], which complicates the automatic generation of random numbers.

Inventory of material**:** Every mentioned specimen or sample must contain a link to a real voucher number in a recognised depository (while paying attention to minor formal details that certify the author's manual work, for example, that the depository abbreviation ends with a full stop.). The absence of collection inventory numbers is grounds for immediate rejection.

3. Requirements for analytical tools if authors claim to have conducted complex statistical or bioinformatic analysis, the requirement for them must be categorical:

Open code: Provision of original scripts (R, Python) and program execution logs. AI perfectly writes code on demand, but often this code generates perfect graphs on fake datasets. Providing logs allows verifying whether the code was actually applied to the declared raw data.

Conclusion

It is time to admit: the era when it was enough to send a well-formatted text (PDF/Word) for publication has ended. The response to the challenge of generative models must be a transition to the concept of the Open Science Framework, where the text of the article is merely an accompanying note to a verified array of primary data. If an author cannot support their results with raw data, vouchers, and calculation algorithms, no perfectly written English text has the right to be called a scientific publication.

Declaration on the use of artificial intelligence and Acknowledgements. The authors officially declare that this editorial text (including the formulation of new requirements for authors and the conceptualisation of the problem) was generated, structured, and stylistically polished in direct dialogue with the large language model Gemini Pro. We make this statement not for the sake of an ironic metaphor, but as a demonstration of the profound and highly uncomfortable crisis of our time. We have just rejected an article by dishonest authors for delegating analytical work to an algorithm, and we ourselves — simultaneously delegated the creation of a text about the inadmissibility of such actions to an algorithm.

This fact should frighten the scientific community far more than a fabricated reference list. The fact that a machine is capable of flawlessly, with the required level of academic anger, structural logic, and imitation of human principledness, writing a manifesto against machine falsifications proves: the boundary between authorial text and machine generation is finally erased. We demand transparency from authors, so we start with ourselves. Responsibility for the ideas, implemented rules, and rejection of the aforementioned manuscript lies wholly and entirely with the human authors.

However, we acknowledge: the toolkit for expressing academic thought is no longer exclusively human. If we do not start strictly controlling primary data today, tomorrow it will be technically impossible to distinguish real science from generated imitation.

The authors express their sincere (and not without a touch of academic irony)

gratitude to the large language model Gemini Pro for the ruthless dissection of its own digital "relatives" algorithms, revealing the anatomy of machine hallucinations, and assisting in structuring this text. The fact that an article about the existential threat of uncontrolled artificial intelligence was analysed and written in dialogue with artificial intelligence is a classic embodiment of the mythological Ouroboros, the serpent eating its own tail.

However, this paradox brilliantly proves our main postulate: AI is neither an independent evil nor a panacea. It is merely a hyper-powerful "optical instrument". In the hands of a dishonest author, it generates convincing mirages, but under the strict control of a critical human mind — it becomes a microscope capable of debunking these mirages. The responsibility for every word, conclusion, and rule introduced in this article, as is appropriate in genuine science, is borne exclusively by the human authors.

REFERENCES

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V. L., Dave, T. & Duddumpudi, R. T. S. 2023. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15, e37432, 1–5.
https://doi.org/10.7759/cureus.37432

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, 1779–1780. https://doi.org/10.1016/S0140-6736(26)00603-3

Walters, W.H. & Wilder, E. I. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, 14045, 1–9.

Received 14 May 2026
Accepted 30 June 2026