

AI is poisoning evidence synthesis

Sam A. Reynolds¹, Alec P. Christie², Lynn V. Dicks¹, Sadiq Jaffer³, Anil Madhavapeddy³, Rebecca K. Smith¹, William J. Sutherland¹

¹ Conservation Science Group, Department of Zoology, Cambridge University, The David Attenborough Building, Pembroke Street, Cambridge CB2 3QZ, UK

² Centre for Environmental Policy, Imperial College London, London, SW7 2AZ, UK

³ Department of Computer Science & Technology, University of Cambridge, Cambridge CB3 0FD, UK

The rigorous synthesis of literature is one of the leading scientific developments of the last fifty years, revolutionising effectiveness in medicine and other fields by allowing us to draw insights from the collated global literature.

The rapid development of Artificial Intelligence (AI) simultaneously presents one of the largest opportunities and biggest threats to the field of evidence synthesis. It promises more efficient and more comprehensive reviews at speeds commensurate with decision making processes. However, misleading and fake manuscripts are increasingly flooding the academic literature and generative AI will make these easier to produce and harder to detect. The first fully AI-generated paper, created by “The AI Scientist”, recently passed peer review for a workshop hosted at ICLR, a top international AI conference. The paper was entirely AI generated, formulating hypotheses, designing and running experiments, analysing results, generating figures and producing the manuscript¹. Use of such technology will likely be attractive to bad actors whilst the numbers of papers and citations remain the rewarded goal of scientific activity, currently facilitated by paper mills and a variable peer review system, which itself is becoming infiltrated by AI-produced reviews².

28

29 The production, and subsequent publication, of ever larger numbers of both
30 legitimate and convincing–fictitious academic papers represents an existential
31 crisis for evidence synthesis, for which it is currently woefully unprepared. Already
32 systematic reviews to Cochrane standards demand significant resources, on
33 average costing >\$140,000 and taking over two years to complete³. Therefore, it is
34 important to assess whether traditional evidence synthesis processes, devised and
35 built under previous technological restraints, are still fit for purpose. For example,
36 a recent study demonstrated over half of systematic reviews reported search
37 strings that were not repeatable⁴. Constrained thinking around improving current
38 evidence synthesis and fraud detection techniques using AI, while not accounting
39 for how the wider landscape of academic literature is changing, might leave us
40 blinkered to the more transformative change required to secure the future of
41 reliable and responsive evidence synthesis.

42 **Challenges facing evidence synthesis**

43 Currently, the gold standard for evidence synthesis is the systematic review. High
44 quality systematic reviews are comprehensive (aiming to include almost all
45 relevant evidence), rigorous (using the best available appropriate methods to
46 reduce biases), transparent (enabling reproducibility) and objective (handling the
47 review process fairly without interference from conflict of interest or cognitive
48 biases). Concerningly, however, many systematic reviews already unwittingly cite
49 retracted publications⁵. Moreover, after being notified of retractions 89% of
50 systematic reviews remained uncorrected one year after notification⁶.

51

Rates of paper retractions are increasing, exacerbated by the rise of paper mills, and the slow responses of journals to confirm and retract papers flagged as illegitimate intensifies their impact on research⁶. However, the retraction of flagged papers at the source is only the first step in removing their influence. Relying on the continued engagement of the original authors, with limited resources and perhaps little incentive, can exacerbate the pollution of retracted science in the conclusions of evidence reviews.

Near ubiquitous access to generative AI, particularly large language models (LLMs), adds rocket fuel to the fire. These are positioned to drastically accelerate the rate at which fraudulent manuscripts can be produced and can also generate convincing doctored ‘raw’ data⁷. One preprint estimates that, conservatively, 10% of all PubMed abstracts published in 2024 were written with LLMs, by observing an abrupt increase in the frequency of certain style words coinciding with widespread access to LLMs⁸. However, for now, it is difficult to separate malicious use from authors simply improving their writing⁹.

Combined with the rapid expansion in the total number of papers available to review, and the number of fraudulent papers in the literature likely to increase, the effort required to comprehensively synthesise the relevant and legitimate available evidence will become considerably more difficult than in the past. Additionally, the increase in retracted papers is likely only the tip of the iceberg. A recent analysis suggests over 400,000 research papers with strong textual similarities to fraudulent studies from paper mills were published over the last decade, of which ~70,000 were published in 2022 alone¹⁰. They estimate that ~2% of all scientific papers published in 2022, and 3% in biology and medicine papers, closely resemble

paper-mill works, and these numbers are only likely to increase over the coming years.

What is the evidence synthesis community doing about it?

Cutting the use of AI in the production of manuscripts at the source is unlikely to be an effective policy and publisher guidelines on 'AI use' in manuscripts varies. The approach of *Nature*, *Science* and many other journals is that the use of LLMs is allowed if declared, on the sensible basis that it improves access for authors whose first language is not English. This approach, while improving access, however muddies the waters when trying to identify harmful AI generated papers against a backdrop of 'acceptable levels' of AI use.

In evidence synthesis, guidance on how to deal with retracted studies is still under debate⁵. However, the need for clarity is more pertinent now than ever. The Campbell collaboration and Cochrane reviews provide guidance on how to identify studies with expressions of concern, or those having been retracted to avoid their inclusion in reviews¹¹. These include checking studies against Retraction Watch (<https://retractionwatch.com/>); using the new functionality of CENTRAL (Cochrane's database of clinical trials), which flags retracted studies; and updated search protocols to identify retraction notices¹².

The majority of the guidance around retracted and suspicious papers focuses on the searching and sorting end of the evidence synthesis pipeline, and less on how systematic reviews should be handled in light of new evidence that may undermine their literature base. Current Cochrane guidance states authors of published reviews containing retracted studies should recalculate all results and flag the

103 review with an editorial note while this is undertaken, or withdraw and republish
104 the review¹³.

105

106 Placing the onus on the authors of reviews to continually identify retracted and
107 suspicious studies and reanalyse data of affected syntheses is unlikely to be a long-
108 term or pragmatic solution to the arising issue of fraudulent manuscript
109 production. Proving the absence of fraudulent papers from reviews, AI generated
110 or otherwise, is synonymous to definitively proving the local extinction of a
111 species. To do so with 100% certainty is mathematically impossible¹⁴, undoubtedly,
112 some papers will slip through the net. We need to develop a system that is robust to
113 the rapid removal of poisoned papers that feature in evidence synthesis.

114 **What is the antidote?**

115 For evidence synthesis to counter the near exponential rise in publications and the
116 predicted increase in both paper retractions and fraudulent papers, requires a
117 system that allows for both scale and dynamism in respect to incorporating new
118 evidence, removing retracted studies and reanalysing data accordingly. The ever-
119 increasing production of both legitimate and spurious scientific literature will
120 likely outstrip human capabilities to cope with the task of maintaining an accurate
121 picture of ‘the science’. We believe a network approach is required, powered by a
122 combination of AI enabled subject-wide evidence synthesis (SWES)¹⁵ with
123 humans-in-the-loop and a geographically and institutionally decentralised
124 network of living evidence databases (LEDs). This will improve the robustness of
125 evidence synthesis and hasten the removal of poisoned papers.

SWES is an evidence synthesis technique developed from 2004 as part of the Conservation Evidence project (www.conservationevidence.com) that aims to identify all literature containing information potentially relevant to a broad theme (i.e. effectiveness of management actions for biodiversity conservation). SWES eliminates the need for search terms and instead performs a single sweep across all literature that presents evidence for a certain theme, rather than requiring separate reviews that would need to requery the literature every time for each research question. Each paper is assessed, the outcomes extracted along with other relevant metadata and indexed appropriately. This eliminates the requirement for papers to be reassessed for most evidence synthesis purposes. Instead, researchers create a 'living evidence database' (LED) of outcomes, which can be pulled out to answer more specific questions for which that outcome, and its context, is relevant. This dynamic LED approach to evidence synthesis is showcased in Metadataset (www.metadataset.com), a tool developed by Conservation Evidence¹⁶.

Networks of LEDs could automatically store information such as the literature searched, screened, appraised, data extracted, the associated analyses, study context and summaries for future use. These LEDs would dynamically cache new studies and associated metadata from a plurality of sources available through their institution or geography. High-quality systematic reviews, and rapid reviews can then be undertaken from verified evidence accumulated across these databases, with humans-in-the-loop if desired, enabling economies of scale and near real-time updating. Users, as well as automatic mechanisms, can flag gaps, retracted papers and errors. The ability to version-control databases, and any resulting evidence review, enhances repeatability. Each review could be associated with a unique identifier, related to the exact databases at the time of searching, so that

the search process can be repeated exactly to check the accuracy, the consequences of subsequent removal of papers and the outcome if a different search had been adopted. Reviews can also be brought up to date either automatically, or with humans-in-the-loop, when LEDs contain new, relevant evidence (Figure 1). The ability for data to be shared and mirrored across LEDs on the network, protects access to scientific data from institutional or political decisions, or technical issues, within certain institutions or geographies that may otherwise prevent access.

Individual systematic reviews may become faster and more efficient with AI, whether through improving search terms, screening literature for inclusion, or extracting data. However, they will always be constrained by the limitations of being individual reviews, based on narrow questions, written and maintained by small groups of authors. Creating LEDs allows the retraction status, study evaluation, outcomes and data extracted from a piece of literature to be tracked throughout the process, allowing confidence levels at each stage to be propagated as well as tracking its dependencies. For example, a human-in-the-loop evaluation might override an AI-made decision at the inclusion or data extraction stage, which can trigger reassessment in later stages automatically. The resultant LEDs can then be used for evidence reviews, with the database dynamically corrected both for inevitable errors and retractions.

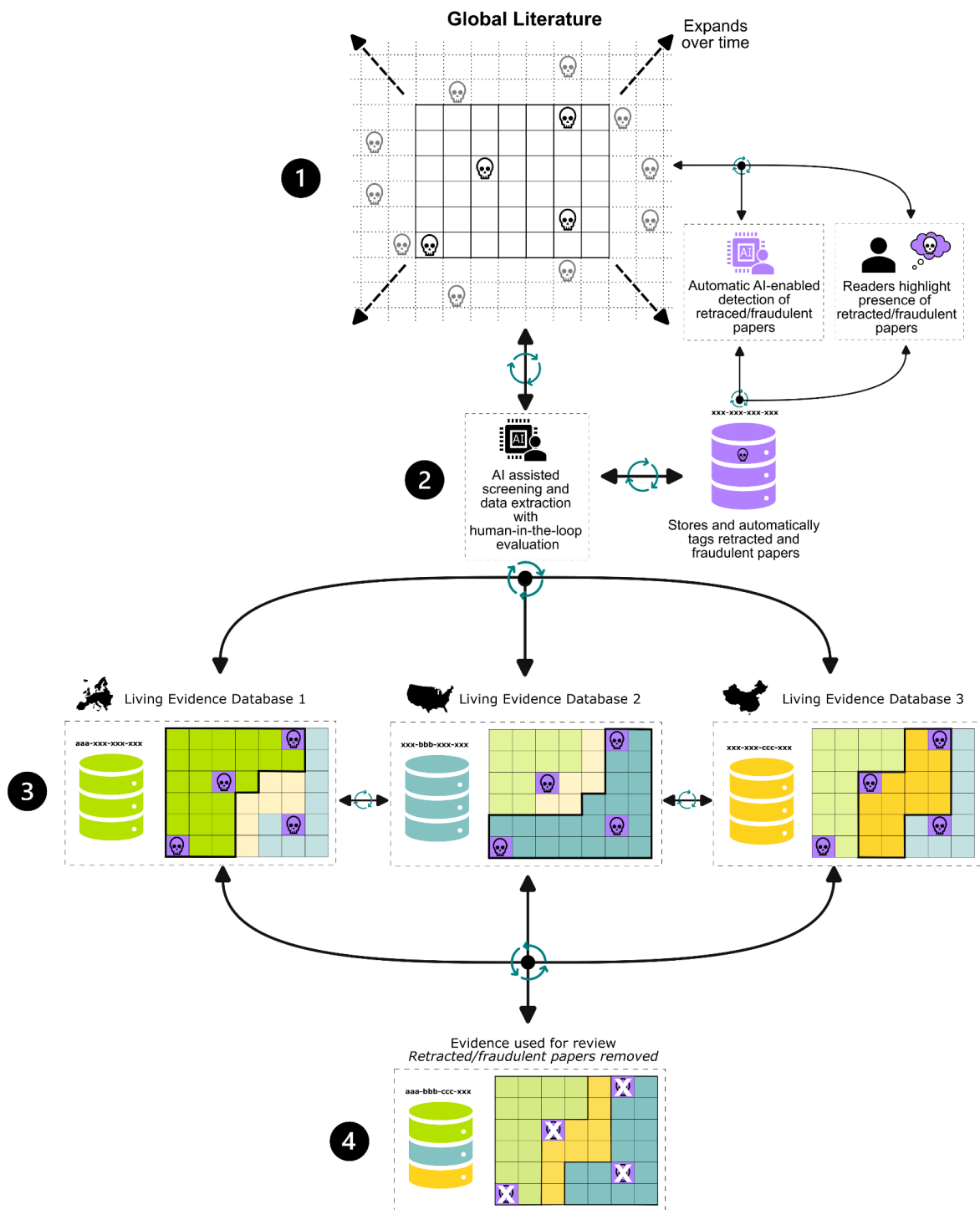


Figure 1. Illustrates how literature may be gathered, collated and synthesised through a network of living evidence databases (LEDs). 1) Represents the global literature base, with fraudulent papers marked with skulls. 2) AI-assisted gathering and screening of evidence, occurring simultaneously across institutions and geographies. Selected studies interface with automatic and manual processes for flagging fraudulent papers. 3) Literature, relevant extracted data and metadata is stored both locally and communicated dynamically to other databases in the network 4) When undertaking an evidence review, relevant evidence is pulled from the network, with fraudulent studies excluded from analysis. The review remains connected to the network so studies can be added or removed in near real-time.

AI can empower this approach in numerous ways, and SWES allows it to scale effectively. AI can help to find logical deficiencies in inclusion criteria, assess whether papers meet inclusion criteria and expedite the translation of key data and metadata from papers into tables which can be used to populate the LEDs. AI can automatically monitor the literature for new relevant studies, remove retracted papers, and assess papers against benchmarks, which signify the likelihood of them being generated by AI. While multilingual LLMs mean that we can also assess and screen papers in multiple languages, which are often missed in systematic reviews¹⁷.

It is important to highlight that one of the key barriers to the future of effective evidence synthesis is access to full-text studies, which varies by institution and geography. For evidence synthesis to remain a viable field, a fundamental shift in access is required. Publishers and data-holders need to be incentivised to share their data more widely, and implement protocols that allow for frictionless sharing to a trusted and neutral third party, as openness is a key tenet of the scientific process to allow replication of results. The impacts of not doing so are clearer now than ever, with the US executive removing access to valuable Centre for Disease Control and National Institute of Health databases. Relying on the actions of individuals to archive this data is not a sustainable or responsible course of action.

The construction of this network superstructure of transparent, traceable and dynamic LEDs supporting global evidence synthesis efforts will help to protect evidence synthesis from wider perturbations in the global literature. It can be used as an oracle of the current state of published scientific knowledge, identifying fraudulent papers, evidence gaps and how contributions to the network shift our

scientific understanding. Failing to build resilient evidence infrastructure will eventually doom the viability and credibility of global evidence synthesis efforts.

References

1. Sakana.ai. The AI Scientist Generates its First Peer-Reviewed Scientific Publication. <https://sakana.ai/> (2025).
2. Piniewski, M., Jarić, I., Koutsoyiannis, D. & Kundzewicz, Z. W. Emerging plagiarism in peer-review evaluation reports: a tip of the iceberg? *Scientometrics* **129**, 2489–2498 (2024).
3. Cumpston M., Flemyng E., Thomas J., Higgins J.P.T., Deeks J.J., Clarke M. J. Chapter I: Introduction [last updated August 2023]. In: Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., Welch, V. A. (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5. Cochrane, (2024). Available from training.cochrane.org/handbook.
4. Li, Z. & Rainer, A. Reproducible searches in systematic reviews: an evaluation and guidelines. *IEEE Access* **11**, 84048–84060 (2023).
5. Bakker, C. *et al.* Reducing the residue of retractions in evidence synthesis: ways to minimise inappropriate citation and use of retracted data. *BMJ Evid.-Based Med.* **29**, 121–126 (2024).
6. Avenell, A., Bolland, M. J., Gamble, G. D. & and Grey, A. A randomized trial alerting authors, with or without coauthors or editors, that research they cited in systematic reviews and guidelines has been retracted. *Account. Res.* **31**, 14–37 (2024).
7. Kwon, D. AI-generated images threaten science – here’s how researchers hope to spot them. *Nature* **635**, 268–269 (2024).
8. Kobak, D., González-Márquez, R., Horvát, E.-Á. & Lause, J. Delving into ChatGPT usage in academic writing through excess vocabulary. Preprint at <https://doi.org/10.48550/arXiv.2406.07016> (2025).
9. Berdejo-Espinola, V. & Amano, T. AI tools can improve equity in science. *Science* **379**, 991 (2023).
10. Van Noorden, R. How big is science’s fake-paper problem? *Nature* **623**, 466–467 (2023).

- 237 11. MacDonald, H. *et al.* Searching for studies: A guide to information retrieval for
238 Campbell systematic reviews. *Campbell Syst. Rev.* **20**, e1433 (2024).
- 239 12. Lefebvre C., Glanville J., Briscoe S., Featherstone R., Littlewood A., Metzendorf
240 M-I., Noel-Storr A., Paynter R., Rader T., Thomas J., Wieland L. S. Technical
241 Supplement to Chapter 4: Searching for and selecting studies [last updated
242 September 2024]. In: Higgins J. P. T., Thomas J., Chandler J., Cumpston M., Li T.,
243 Page M. J., Welch V. A. (editors). *Cochrane Handbook for Systematic Reviews of*
244 *Interventions* version 6.5. Cochrane, (2024). Available from
245 www.training.cochrane.org/handbook
- 246 13. Higgins J. P. T., Thomas J., Chandler J., Cumpston M., Li T., Page M. J., Welch V
247 .A. (editors). *Cochrane Handbook for Systematic Reviews of Interventions*
248 version 6.5 (updated August 2024). Cochrane, (2024). Available from
249 www.training.cochrane.org/handbook.
- 250 14. Reed, J. M. Using Statistical Probability to Increase Confidence of Inferring
251 Species Extinction. *Conserv. Biol.* **10**, 1283–85 (1996).
- 252 15. Sutherland, W. J. & Wordley, C. F. R. A fresh approach to evidence synthesis.
253 *Nature* **558**, 364–366 (2018).
- 254 16. Shackelford, G. E. *et al.* Dynamic meta-analysis: a method of using global
255 evidence for local decision making. *BMC Biol.* **19**, 33 (2021).
- 256 17. Nuñez, M. A. & Amano, T. Monolingual searches can limit and bias results in
257 global literature reviews. *Nat. Ecol. Evol.* **5**, 264–264 (2021).