Dear Dr. Collins,

Thanks very much for providing these constructive reviews. We have taken this opportunity to make several substantial changes to the manuscript in-line with the reviewers’ suggestions and have provided in-line responses to these suggestions below.

In particular, we have augmented the detail in the methods, including cleaning up the github repository and releasing a Zenodo repository so that the methods are replicable and transparent. We have also performed a sensitivity analysis using different model set-ups and model choices. We have corrected a minor error in our analysis but show that our results still hold. We have also improved the discussion and introduction in-line with the reviewers comments.

We hope that this updated draft satisfies the reviewers reservations and is now suitable for acceptance at One Earth.

Thanks again,

Dr. Scott Spillias, et al.

Human-AI Collaboration to Identify Literature for Evidence Synthesis

ONE-EARTH-D-23-00604

Nov 22, 2023

Dear Dr. Spillias,

I hope this email finds you well. Thank you for your patience while we have managed the peer review of your manuscript "Human-AI Collaboration to Identify Literature for Evidence Synthesis", which has now been seen by our three reviewers, whose comments are enclosed here. As you will see from their comments, while all reviewers have commented on the potential significance of this work and Reviewers #1 and #2 are very complimentary and recommend only minor changes, Reviewer #3 (our AI expert) has identified a number of more significant concerns that require attention. Therefore, although we are unable to accept this version of the manuscript for publication, we would be interested in considering a revised version that sufficiently addressed the concerns raised.

I hope that you find the reviewers' comments constructive; if it is possible to satisfy the comments by obtaining further decisive data or conducting additional analysis, we would be happy to consider a substantially revised manuscript. However, please take into consideration that we would be reluctant to approach our referees again in the absence of major revisions. We are committed to providing a fair and constructive peer review process. If there are specific reviewer comments or requests that you feel are technically impossible, unlikely to yield a meaningful outcome, or extend the scope of the paper unnecessarily, please do contact us prior to resubmission.

We generally suggest a revision time frame of up to four months. Should the revision process take significantly longer than this timeframe, we will be happy to consider your manuscript at a later date, so long as nothing similar has been accepted for publication at One Earth or published anywhere else in the interim. We understand that changes to your circumstances can lead to challenges in completing revisions, especially as various global crises may be causing disruption for you and your colleagues. If that is the case for you and it has an impact on your ability to revise your manuscript, please let us know, and we will be happy to work with you on a plan that works for you to keep your paper moving forward.

Once you have had a chance to consider the reviewers' comments, please let us know if you plan on submitting a revision. When resubmitting your paper, please highlight all changes in the manuscript text file and provide a point-by-point response to the reviewer’s comments (this should be provided in a separate document to your cover letter). We recommend using line numbers to link your revise manuscript and point-by-point response.

Best wishes,

Lewis Collins, Ph.D.

Editor-in-Chief, One Earth

Reviewers' Comments:

Reviewer #1:

This article examined through a structured process the potential benefits of using AI technology to support or improve evidence synthesis tasks, in this case a scoping review process. The paper examined potential for AI to replace and/or "collaborate" with humans in the work of developing search algorithms and the labor intensive task of screening. I am writing this review as a fellow evidence synthesist, not an informatics/AI expert, but as such equally interested in understanding where the value of AI-assisted work may be. **I have designated that I find no issues with the submission and would (indeed, did) find it a useful read.** I especially appreciated the attention in the discussion to the potential trade-offs, and encouragement to be mindful of the full value of literature reviews to the development of new (or indeed even simply improving) researchers.

**If there are revisions, my one suggestion would be to note whether the human team included a reference librarian.** In my world of evidence-based health care delivery, a medical librarian is a highly respected and necessary team member. That's not to say that other researchers can't develop similar skills, perhaps especially in fields where indexed terminology is less critical (than it is in medical or healthcare research). But such team members, plus peer review by other skilled reference librarians, would be a touch point that would be looked for.

**Thank you for the positive comments. We note that while we haven’t employed a reference librarian, we highlight in the text that the ‘human team’ has expertise in evidence synthesis in the field of environmental sciences and that whilst a reference librarian would aid in developing a search string, our team has experience doing this for numerous projects.**

**96-96:** *To assess the usefulness of AI tools for evidence synthesis we assembled a ‘human team’ and an ‘AI team’. The human team consisted of five humans (PT, JD, DK, RS, RAJ) with expertise in undertaking various forms of evidence synthesis (e.g.* [*17,18*](https://www.zotero.org/google-docs/?hoTFC8)*).*

Reviewer #2:nesses technological advancements in the field of AI to enhance review capacity across potential large bodies of diverse literature. This of course is a timely inquiry given the emerging technologies increasingly playing roles in how research is undertaken. The manuscript makes a compelling and demonstrated argument for the use of such technology, as well as a proposed approach. It is moreover well written. For those reasons I commend the authors. **My comments are minor and refer to more general reflections on the paper.** I would invite the authors to consider the following comments:

AI clearly offers great value in enhancing an expanding quality of reviews of literature, particular to overcome human deficiencies. The case study demonstrates where the technology can overcome time and capacity constraints of the researcher. **However there are other human deficiency that may be worth examining too - for example language (any research published in other languages), other disciplinary focus, and literature types (beyond academic data bases, but to include the grey literature). These example dimensions may be beyond the scope of what the case study could offer**

The manuscript presents an innovative approach that har**and rather may be areas of further research testing, but it would be good to reflect more explicitly on further applications in the discussion (and what pitfalls exist there) as the tech evolves.**

**We agree that these results point to many promising avenues. We have added the following to the discussion to address these points:**

***331-333: These results also suggest other applications where human-AI collaborations could overcome barriers in research, such as in better including non-English research, non-peer-reviewed data sources, and important findings from beyond one’s domain expertise.***

**Did the authors consider involving CBFM experts in the review to help make assessments.** There are of course advantages and disadvantages to this in terms of methodological design. I don't suggest changing methodological design at this stage, but it may be worth adding a few lines of reflection in the methods about whether addition of CBFM experts would be beneficial or disruptive, as a justification to the current design.

**Thank you for the opportunity to reflect on this point. We have added the following text to the discussion:**

***349-360: In the searches and formulation of research questions we drew from the author group’s experience with marine social science, studies of fisheries governance and management, perceptions of local communities towards marine spaces and management and from previous collaborations in the Pacific region. We also emphasise that this is a pilot to explore the efficacy of AI methods. For a full review (as opposed to a pilot) it could be beneficial ensure that the team includes a high level of specific disciplinary expertise i.e., in this case, expertise in CBFM Given the high context-specificity of CBFM concepts (e.g., Aswani et al. 2017) for a real full systematic review we would include and be informed by local experts/organizations to ensure relevance and legitimacy of findings, and counter Wester-based terminology and post-colonial parachute science. However, given that similar reviews are often conducted by inexperienced researchers, such as PhD students as part of their initial exploration of a discipline/field, in the absence of domain experts, our results suggest that an AI assistant could be an invaluable source of guidance.***

**It seems the literature review was focused specifically on CBFM in the Pacific, rather than a global study of CBFM. This is not consistently communicated in the framing of the paper, in most cases this is simply a matter of adjusting wording.**

**Thank you for highlighting this inconsistency, we have gone back and re-worded where this may have been confusing.**

I encourage the authors to consider the above and with that support proceeding this manuscript to publication.

Reviewer #3:

The paper presents a study on the automation capability of the initial two tasks in the process of creating literature reviews (evidence syntheses) by generative AI agents (ChatGPT). These two tasks are the generation of literature search strings and candidate screening. To this end, they analyze three alternatives: results from humans, results from AIs (individually and in committee), and results from human-AI collaboration. The analysis focuses on a very specific case study of community-based fisheries management.

**Regardless of the case study, the use of general AIs (LLMs) for this problem seems to me a very interesting solution worthy of exploration**. The authors themselves (lines 69-75) note that AI use has expanded to support the different stages of systematic evidence syntheses. However, they point out that these other proposals are limited to specific literature review stages. **Curiously, after this assertion, the authors use this generative AI proposal based on ChatGPT to support the conducting phase (paper selection) in the systematic literature review process, which is possibly one of the most addressed with AI proposals, although mostly through supervised techniques and active learning.** I recognize the interest of the work, very timely, and I appreciate its reading. **I certainly enjoyed it, and it is easy to understand.**

**Nevertheless, the motivation for the work is scarce. Why do the authors focus on these tasks?** The authors seem to justify it with a short "Specifically, we evaluate the usefulness of an LLM to…". **Therefore, despite praising the virtues of LLMs as tools to support the evidence synthesis process, I still have doubts about whether an integral method for the entire review process or most of it is possible.** At least, it would be interesting to know why this phase of the process or if the generalist AI could cover the gaps in the process that other AI methods have not been able to cover so far.

**We are pleased that the reviewer found this work to be interesting, enjoyable and easy to read, and appreciate this guidance on how to better articulate the motivation of the specific approach that we adopted. We have added the following to the introduction to better articulate this:**

**83-84:***We chose these stages as they are the initial steps in a systematic review, and in particular the screening stage has been identified as being particularly suited to AI collaboration (Chappell et al., 2023; Hamel et al., 2021).*

**Overall, I liked the article and believe this idea has high potential.** While the paper presents several compelling aspects, there are key areas that would benefit from further clarification and detail. Addressing these concerns would greatly enhance the paper's current state.

1. Poorly defined methodology.

**The description of the methodology is very limited. Unfortunately, Figure 1 is far from sufficient. The entire experimental procedure that guides the interaction with ChatGPT should be detailed.** Explain how the team of AI agents is constituted, how many prompts compose each interaction, the patterns used (see the work of J. White that the authors themselves cite), how GPT's outputs were treated, the date of the interactions/model, etc.

Some of this information appears but is superficial and very scattered when it should really be a well-defined experimental framework.

**Thank you for this feedback. We note that for the screening section of the study, the code is provided in the repository and we detail how these data are obtained and processed in the methods section. We have updated the relevant parts of our methods to ensure that it is clear that full methodological detail is available in the repository. For the search string section of the study, we have provided links to the specific interactions with ChatGPT so that readers can verify the prompt and outputs.**

An example is the treatment of randomness. Even if GPT behaves deterministically, both the preprocessing (e.g., word sampling) and the context used in the interaction include non-deterministic components that could influence the response. For example, in Section B, when the text is indicated, what context is executed in the different repetitions?

**These data were collected from ChatGPT using the web interface before ‘Custom Instructions’ were introduced. Therefore there was no additional context beyond the AI’s training and whatever safety features OpenAI incorporated at this time (which are unknown to us and the general public). We have included links to the specific chats in the Readme on the Github repository and again updated the relevant text in the manuscript to clearly point toward this.**

The methodology should clarify and motivate the use of 3 agents (l. 287), the interaction of keywords with 5 separate instances of ChatGPT online interface (l. 392), the screening "repeatedly asking the same prompt three times" (l.427), or the use of API (Python code in the repository) for other interactions. It is confusing to know where what comes from. It is important to establish a clear method to mitigate potential biases and doubts in this regard.

**We have included links to the specific ChatGPT transcripts from late April 2023 in the Github Readme.**

*399 - 406: In parallel, the AI team used ChatGPT to develop an appropriate list of keywords (i.e., a search string). Specifically, we developed a prompt inspired by the strategy of (Wang et al., 2023)(Table 1), and submitted it to five separate instances of the ChatGPT online interface,* ***without any prior prompting or context. These conversations occurred in late April 2023 (prior to the introduction of ‘custom instructions’) and; we provide links to these in the Github repository: https://github.com/s-spillias/GPT-Screening.*** *We then provided the five resulting search strings and asked one instance of ChatGPT to decide which was best for the research topic (for the specific prompt see Supplement G).*

***430-431: Further, we used the API in Python, rather than the online web interface, so that the entire process could be automated and the OpenAI parameters controlled.***

***437 - 438: Based on preliminary trials where the AI outputs were prone to confabulations and incorrect interpretations compared to the human pilot screen data, we heavily modified the screening criteria for the AI (compare Supplement D & E) and also explored two techniques to mitigate against erroneous AI output.***

Regarding diversity (Annex F), the use of random strings also influences, which seemed a creative strategy considering the stateless nature of the interaction with ChatGPT. I understand that it could work for very low temperature cases. But here I want to go back to the comment above: in the absence of a clearly indicated procedure, it is difficult to assess whether there would be other alternatives, such as the use of certain patterns or parameters. On the one hand, has a more tailored strategy for the use of the language model been tested, such as the use of prompts that require follow-up, or few-shot prompting? The basic "Person" pattern has been used, but it would be interesting to know whether there were other attempts. For example, have the authors tried the "Question refinement pattern"? Similarly, I speculate whether the flipped interaction pattern would also help improve question writing. I cannot see a focus on how prompts were redacted, but it is an important matter too.

**We agree that there may indeed be better strategies for refining the process we have followed. We have highlighted these ideas as avenues of further study in the discussion in the hopes that other researchers pick up this thread:**

**328-329: *Further refinement of the process described here may also be possible, including the use of few-shot prompting, or other strategies that refine the questions posed to the AI.***

On the other hand, if the "temperature" and other parameters (e.g., "top-p") are adjusted, it could be interesting to perform a sensitivity analysis of the problem to these parameters and, even, to study whether the random string would still be necessary. The methodology should also clarify these points and justify the method used properly.

https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api-a-few-tips-and-tricks-on-controlling-the-creativity-deterministic-output-of-prompt-responses/172683/10

**Thank you for these suggestions. We did indeed experiment with these parameters, but found the best results with the configuration presented here. We have added a sensitivity analysis to the supplemental information where we demonstrate the influence of altering these parameters, rather than appending the random string for the best performing AI strategy (AnyYes). In summary, for a temperature of 1.0, default top-p (1.0), and no random string, the kappa-value was 0.41. This increased to 0.60 for a temperature of 0.0, which is better, but not as good as when the random string was used (k = 0.69). When we modified top-p to be 0.2, with a default temperature of 1.0 and no random string, we found a kappa-value of 0.40. In all three cases, these are compared to the final ‘Collaborative list’. Compared to the human team’s initial screen these kappa values are lower. So in both cases, merely adjusting these parameters without appending the random string does not yield good results. We have added this information to the supplemental section:**

**789-797: *We compared the influence of appending the random string compared to other conventional settings of temperature and top-p by performing a sensitivity analysis where we modified the top-p and temperature parameter using the best performing AI strategy (AnyYes). In summary, for a temperature of 1.0, a default top-p (1.0), and no random string, the kappa-value compared to the ‘Collaborative list’ was 0.41. This increased to 0.60 for a temperature of 0.0, which is better, but not as good as when the random string was used (k = 0.69). When we modified top-p to be 0.2, with a default temperature of 1.0 and no random string, we found a kappa-value of 0.40. These results suggest that appending the random string does indeed improve the quality of the screen beyond what can be achieved by modifying OpenAI’s default parameters.***

**And added the following to the results.**

**284-293: *Sensitivity and Reliability***

***We repeated the screening process in December 2023, using a variety of different OpenAI models and parameters. Interestingly, using the same settings, parameters, and prompts we were unable to obtain the same level of agreement with the collaborative list using any of the gpt-3.5-turbo models, including the legacy models. Instead, we found that the gpt-4-turbo model performed comparably to our results from May 2023, although its false negative rate was still higher at 3.3%. Closer examination of the model outputs suggested that many articles kept by the human team were being excluded by gpt-4-turbo because the abstracts mentioned ‘co-management’. We therefore performed the screen again with a modified second criterion without the ‘co-management’ stipulation, which resulted in a kappa value of 0.78, and a false negative rate of 1.3% (Figure 4).***

In the introduction, the authors mention that other AI techniques have already been employed for the different phases of the construction of evidence syntheses. The use of generative AI seems a natural evolution because of the advantages it offers. **However, an analysis of the accuracy and suitability of this technique against previous results obtained by other approaches would be convenient.** In the case of the conducting phase (search strings and paper selection), it is usual to mostly find supervised techniques or active learning. At this point, a general AI that is already trained but efficient can be an advantage to the end user. While generative AI's pre-trained nature may offer user advantages, the paper should address its generalizability (see next comment). Therefore, it would be beneficial to study more deeply how this proposal can overcome the limitations of other techniques, as well as its limitations compared to them. This might lead to a more convincing contribution.

**We highlight below the considerations around the generalizability of this method and its limitations.**

2. Results not generalizable.

The work deals with a very specific and surely well-defined example. **However, how does the domain affect the use of this tool?** Since the authors are working with a single case study, I wonder what exactly the contribution of the article is beyond the example of how it could be done (l. 366). I also wonder how generalizable the conclusions are. The title of the work, as well as the abstract, refer to the general use of ChatGPT "to undertake two initial stages of evidence syntheses". However, the work is limited to a single topic (case). Considering that ChatGPT has been trained with a knowledge base (I understand that unknown to the authors), I find it biased to base the conclusions on specific responses, since the depth and accuracy of these responses might be influenced by the research topic. This is more evident in highly technical and specialized fields. How does this factor affect your work? Could it affect the generation of search strings or the selection of articles? I speculate that the team of humans has a specialization in the topic of the review, while we cannot guarantee it for the AI or, at least, the precision of its knowledge. In this case, the limitation to a single case study to draw seemingly general conclusions is not convincing.

**We concur that this is an important limitation of the work. However, it is important to note that this work is still novel and worthy of publication for three reasons:**

1. **Although the training data set is unknown, it is unlikely that there would be more documents on CBFM in the Pacific, than any other topic. Thus, we would not expect ChatGPT to perform better on the topic of CBFM in the Pacific than other related, but equally niche, topics of environmental/sustainability importance relevant to readers of *One Earth.***
2. **At least for the screening step, the research topic is probably less important, as the task of classifying a chunk of text (abstract) based on specific screening criteria is often assigned to research assistants or PhD students who often have limited experience in the field that they are reviewing.**
3. **The publication of this singular case study will likely encourage others to attempt similar projects within their own domains and share their experience.**

**We have added the following text to reflect these points.**

**335-342: *Nevertheless, although the training data set is unknown, there is little reason to believe that this particular topic would be better represented than any other topic of socio-ecological importance. Thus, its performance in this case study holds promise that it might perform similarly when faced with other topics in this field. Because ChatGPT's responses cannot be anticipated ahead of time due to randomness, uncertainties in the training data set, and uncertainties in the underlying infrastructure its usefulness can only be explored empirically. Future case studies will be necessary before we can be confident that its performance is consistent across domains and research questions.***

Furthermore, please consider the following comments for improvement:

- Line 60. The authors state that "the terms evidence synthesis and reviews are used interchangeably". On one hand, I suggest using only one term to improve the readability of the work. On the other hand, although certainly they are very close terms, it is my understanding that "evidence synthesis" and "review" apparently would have a slightly different scope. In fact, I thought that an "evidence synthesis" implies a broader work in terms of integrating research findings with techniques such as, for example, data analysis; however, a "review" deals with collecting and interpreting current research systematically. Is that the case? (I may be wrong) If so, it would be interesting to clarify it properly in the manuscript.

**Thanks for highlighting this potential confusion. We have altered the text to be clearer that a review is a subset of evidence synthesis:**

**59-61: *There are many different methods of synthesis (e.g., systematic reviews, systematic maps, scoping reviews, rapid evidence assessments), with the method adopted for a specific context depending on the purpose, scope of the topic and research questions, or the availability of resources***[***3,4***](https://www.zotero.org/google-docs/?5FyR1h)***. Systematic methods can exceed the quality and strength of traditional literature reviews (e.g., narrative reviews) through the application of three key principles: rigour, transparency and replicability***[***5,6***](https://www.zotero.org/google-docs/?L8T9bu)***.***

- Line 88. I do not consider the use of a blog post ("The conversation") to serve here as part of the motivation for the work, in this case, justifying the use of human-AI collaboration for its better performance compared to a team of humans or to AI separately. Please use a trustable peer-reviewed reference instead.

**We have removed the blog-post citation.**

- Line 94. "human with modeling expertise". "Modeling" is a broad term. Could you please be more specific?

**We have edited the text as follows:**

***94-95: The AI team consisted of four humans (SS, MA, FB, RT) with modelling expertise in quantitative and qualitative socio-ecological modelling, [...]***

- Line 98. What are these AI Agents? Although explained in Methods, it would be interesting for the discourse not to require so much cross-referencing and to explain the concept lightly to understand it. Or use another term at this point in the discourse.

**We have replaced the word ‘agents’ with ‘replicates’ throughout. E.g.:**

***99: …and asking a ‘committee’ of AI replicates (separate API calls) the same prompt…***

- Line 125. The constant leaps between sections break the discourse at many moments, which worsens readability. For example, a case that could be resolved would be Table 1, whose reference is 10 pages later its first citation.

**We note that this is perhaps a product of the methods section occurring at the end of the article. We have changed the citation to the Table to better signpost where the Table can be found:**

**124: *B****ased on the prompts given by the independent team of reviewers 'AI team'* ***(SS, MA, FB, RT) (see Methods, Table 1),*** *the AI generated the following search string and screening criteria:*

- In the bibliography, there are incomplete references. Likewise, although there are articles on arXiv that are relevant (e.g., the mentioned one by J. White), their use should be limited. Avoid the use of blogs and other non-peer-reviewed references.

**We have removed the blogpost.**

Comments on the replicability package:

I appreciate the transparency of the authors and that they have made available in a GitHub repository the data from this work, as well as the source code. However, let me make the following recommendations:

1) Explain in the "readme.md" file the content of the package and of each directory/file properly. For example, we find a "GPT4" directory when apparently GPT3.5 was used. There are also duplicate files named with different indices or named as "- Copy". It is not intuitive to understand what is being viewed.

**We have cleaned up the repository.**

2) Once a consistent state for the current version of the repository has been reached, you could generate a release and use Zenodo (or another similar system) to obtain a snapshot of the repository's state. This ensures that the data will not be lost/vary with respect to what has been validated during the peer-review process.

**Thank you for pointing this out and for these suggestions. We have updated the repository to reflect only what is found here in the paper. We have also uploaded all of the data to Zenodo, as suggested:**

[**DOI: 10.5281/zenodo.10595355**](https://doi.org/10.5281/zenodo.10595355)

Summary:

I acknowledge that this is a work with significant potential and timely contribution. However, it has certain weaknesses that prevent me from accepting it in its current form. Moreover, the extent of some of these issues makes me doubt its acceptability unless the authors make considerable efforts to improve it. Even after revision, I am unsure if the article would meet the acceptance criteria. The most relevant points mentioned above are:

- The authors are encouraged to **discuss the limitations of their case study** and suggest exploring the applicability of their findings to other domains.

**As above, we have expanded the discussion section to further explore the limitations and possible extensions of this work.**

**335-342: *Nevertheless, although the training data set is unknown, there is little reason to believe that this particular topic would be better represented than any other topic of socio-ecological importance. Thus, its performance in this case study holds promise that it might perform similarly when faced with other topics in this field. Because ChatGPT's responses cannot be anticipated ahead of time due to randomness, uncertainties in the training data set, and uncertainties in the underlying infrastructure its usefulness can only be explored empirically. Future case studies will be necessary before we can be confident that its performance is consistent across domains and research questions.***

**370-372:  *We also found inconsistencies in performance over time and within iterations of a particular model and parameter configuration, which highlights the necessity for human oversight and underscores the need for open access models that are transparent to researchers.***

- **The authors should provide more detailed descriptions of their experimental setup and the rationale behind their methodological choices.**

**As described above, we have augmented the methods section with additional detail and have provided an improved repository in which all of the interactions with ChatGPT are accessible and the code used to contact that API are present.**

- The introduction rightly positions generative AI in evidence synthesis; **a detailed comparative analysis with existing methods to highlight its strengths and limitations is recommended.**

**We have compare our results to a very recent review of similar methodologies:**

**330-331:** *The false positive and false negative rates we report here (<5%) fall in the better performing end of the range of similar machine learning techniques (see (Chappell et al., 2023) for a full list).*

- Prioritize peer-reviewed literature in their citations to strengthen the paper's credibility.

**Done**

- Provide a more organized and comprehensive replicability package, possibly by following a recognized standard for data and code documentation.

**Done**