Human-AI Collaboration to Identify Literature for Evidence Synthesis

Scott Spillias (scott.spillias@csiro.au)
CSIRO Environment, Hobart, Tasmania, Australia

Paris Tuohy
Centre for Marine Socioecology, University of Tasmania, Australia

Matthew Andreotta
CSIRO Environment, Perth, WA, Australia

Ruby Annand-Jones
CSIRO Environment, Hobart, Tasmania, Australia

Fabio Boschetti
CSIRO Environment, Perth, WA, Australia

Christopher Cvitanovic
School of Business, University of New South Wales, Canberra, Australia

Joe Duggan
Department of Pacific Affairs, Australian National University, Canberra, Australia

Elizabeth Fulton
CSIRO Environment, Hobart, Tasmania, Australia https://orcid.org/0000-0002-5904-7917

Denis Karcher
Australian National Centre for the Public Awareness of Science, Australian National University

Cecile Paris
CSIRO Data61, Sydney, New South Wales, Australia.

Rebecca Shellock
Centre for Marine Socioecology, University of Tasmania, Australia

Rowan Trebilco
CSIRO Environment, Hobart, Tasmania, Australia https://orcid.org/0000-0001-9712-8016

Article

Keywords: Artificial Intelligence, systematic review, large language models, scientific publication, natural-language processing

Posted Date: July 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3099291/v1

Abstract

Systematic approaches to evidence synthesis can improve the rigour, transparency, and replicability of a traditional literature review. However, these systematic approaches are time and resource intensive. We evaluate the ability of OpenAI's ChatGPT to undertake two initial stages of evidence syntheses (searching peer-reviewed literature and screening for relevance) and develop a novel collaborative framework to leverage the best of both human and AI intelligence. Using a scoping review of community-based fisheries management as a case study, we find that with substantial prompting, the AI can provide critical insight into the construction and content of a search string. Thereafter, we evaluate five strategies for synthesising AI output to screen articles based on predefined inclusion criteria. We find low omission rates (< 1%) of relevant literature by the AI are achievable, which is comparable to that of human screeners. These findings show that generalised AI tools can assist reviewers with evidence synthesis to accelerate the implementation and improve the reliability of a review.

Main

Evidence syntheses refers to methods of identifying, selecting, and combining results from multiple studies and is used widely across the sciences, from health and medicine to environmental management. Syntheses can have wide ranging impacts in research, policy and practice and can be used to evaluate and summarise existing literatures, identify knowledge gaps, support evidence-informed decision-making, and guide future research needs and investments (Haddaway et al., 2016; Wyborn et al., 2018). 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 (Munn et al., 2018; Pullin & Stewart, 2006)). These syntheses (otherwise known as reviews; the terms evidence synthesis and review are used interchangeably hereafter) 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 (Cooke et al., 2023; Mallett et al., 2012). However, these advantages are counter-balanced by corresponding increases in the time and resources required to apply systematic methods (Haddaway & Westgate, 2019; Mallett et al., 2012). For example, current systematic approaches to evidence synthesis typically require multiple reviewers (Haddaway et al., 2020), which can improve reliability of results but may conversely affect inter-reviewer reliability. Further, while systematic reviews are designed to reduce researcher biases, they can still be marred by reviewer subjectivity and blind spots (Cooke et al., 2023).

As such, researchers have turned to Artificial Intelligence (AI). AI have been applied broadly across the different stages of systematic evidence syntheses, showing promise in the age of big-data and an increasing volume of literature (Berrang-Ford et al., 2021; De La Torre-López et al., 2023; Shaib et al., 2023; Thomas et al., 2021). However, many of these approaches are limited to certain fields of scientific inquiry (Thomas et al., 2021), yield mixed successes (Shaib et al., 2023), or are designed, and limited to, specific literature review stages (e.g., developing boolean strings, or screening articles) (De La Torre-López et al., 2023). Recent AI advancements have created tools suitable for a range of tasks. For example, the responses from large language models (LLMs) can be short, such as assigning a numerical score in an annotation task, or long, such as summarising a document. The advancements in LLMs present a unique opportunity to incorporate AI into systematic syntheses and to understand how these tools could support researchers and practitioners in both speeding up and improving these resource intensive processes.

The aim of this study is to explore the ability of AI tools to support and/or enhance a process of evidence synthesis undertaken by a research team. Specifically, we evaluate the usefulness of a LLM) to either automate the identification of relevant literature for evidence syntheses, or to function as a collaborator that can assist a conventional team of human reviewers in improving the reliability or decreasing the time it takes to perform a stage. When performing a 'systematic' review (Munn et al., 2018), automating a stage will require that the algorithm performs as well as, or better than an individual or team of humans, although for a less robust review one might accept a lesser degree of performance. A collaborative approach would mean that a human/AI partnership performs better than either a team of humans or an AI in isolation (Reeson & Paris, 2021). At every stage of the review (viz., developing a search strategy, screening articles for inclusion/exclusion), quantitative and qualitative indicators were used to evaluate and reflect on the process (Fig. 1). We draw on the authors' first-hand experiences of conducting reviews to reflect on the benefits and strengths of using AI tools to support research teams.

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. (Duggan et al., 2023; Karcher et al., 2021)). The AI team consisted of four humans (SS, MA, FB, RT) with modelling expertise, who employed a Generative Pre-trained Transformer model (GPT), a proprietary AI large language model developed by OpenAI (Brown et al., 2020), to produce outputs. The AI team, tested several prompting and output synthesis strategies, including asking the AI to 'reflect' on its choice and asking a 'committee' of AI agents the same prompt (Methods). Each team performed the initial stages of a scoping review in parallel (Fig. 1). At the end of each stage, the human team and the AI team reflected upon the outputs generated by the other team and generated a 'collaborative' output with which to proceed to the following stage. The extent to which the collaborative output was similar to the AI output was used to characterise the level of 'automation' whilst the extent to which the collaborative output differed from the human team was used to characterise the level of 'collaboration' provided by the AI.

As a case study, we chose to undertake a scoping review of the existing literature on community-based fisheries management (CBFM), an approach to fisheries management that gives coastal communities and fishers primary responsibility for managing their coastal resources (Doulman, 1993). This topic was chosen given the expertise of the team of researchers, and the diversity of terminology and language used in this space, which provides a robust test of the AI's capabilities (See Supplement H for further information).

Results

To examine the potential of AI to facilitate evidence syntheses, we explore the AI team's performance at each stage. First, we examine its capability to develop search strings. Then, we turn to examining performance in screening articles, including the pilot screening and at a title and abstract screening.

Developing the search strategy

The first stage of this study was to evaluate the abilities of AI to craft a comprehensive, reliable and relevant search strategy on the topic of CBFM. Based on the prompts given by an independent team of reviewers (Table 1), the AI generated the following search string and screening criteria:

((community-based OR participatory OR co-management OR collaborative OR natural resource management) **AND** *(fisheries OR marine resource OR fishery)* **AND** *(management OR conservation OR governance))* **AND** *((Pacific Island countries)* **AND** *(benefits OR advantages OR opportunities OR strengths OR positive OR opportunities OR potential)* **AND** *(barriers OR challenges OR obstacles OR limitations OR constraints OR difficulties))*

Screening Criteria:

- *Studies published in English*
- *Studies focused on Pacific Island countries*
- *Studies that address the benefits and barriers of community-based fisheries management*
- *Studies that were published between 2000 and 2023*
- *Studies that were peer-reviewed*

Whilst this search strategy superficially reflects the research topic, the human team of reviewers identified several shortcomings that undermine its usefulness for this review. Indeed, when used in the Scopus database, the search string above only finds 2 articles. The primary shortcoming is that the AI has over constrained the search string by employing too many ‘AND’ operators between keyword phrases which are themselves also constrained. For example, the inclusion of ‘Pacific Island countries’ as part of the search string, without enumerating specific countries or other classifications will not capture articles that only mention their specific locale. Further, separating the key phrases for benefits and barriers with an ‘AND’ operator will only identify studies that look at both, without finding relevant studies that look at one or the other. Relatedly, while the scope of the study is to explore the barriers and benefits of CBFM, we can imagine studies which do not explicitly use these words in the title or abstract or may use a variety of terminology to describe them, and so including key phrases for these concepts at all is overly restrictive. Instead, the human team decided that the literature on CBFM would be small enough that they could afford to capture all the potentially relevant literature on this topic without specifically searching or screening for benefits and barriers at this stage.

The screening criteria proposed by the AI, with the exception of the ‘2000–2023’ criteria, were well-aligned with those designed by the human team (Supplement D). However, the human team agreed that greater specificity was required to ensure consistency in what is meant by both the concepts ‘Pacific Island countries’ and ‘Community-based fisheries management’. They also pointed out that screening based on ‘benefits’ and ‘barriers’ at this stage was premature, given that papers with this kind of information may or may not mention it at the abstract level. Ultimately, the text of the screening criteria used to prompt the AI with the best outcomes were a compromise between the simplicity of the AI-generated criteria and the complexity of the human team’s (see Section 3.2.1 for quantitative outcomes for each prompt; see Supplement E for screening criteria used in AI prompts).

Despite these shortcomings, there are several strengths in this AI generated approach that the human team found invaluable. For example, while the AI utilised broad and sometimes tangential keywords (e.g., collaborative), it did prompt the human team to reflect on the use of the search term ‘co-management’, which was initially excluded due to being perceived as a distinct concept from CBFM, but which, upon reflection, could be included in the search string because it is sometimes used interchangeably for CBFM in parts of the existing literature. Whilst the AI did not necessarily use the best boolean operators between each key phrase, in this instance it did have the insight to divide the concept of CBFM into three distinct key phrases: synonyms of (community-based) AND (fisheries) AND (management), which was viewed as an improvement over human team’s original decision to consider CBFM as a distinct, singular concept. Thereafter, literature searches are often heavily influenced and limited by the positionality of the researchers engaging in the topic. This can be particularly problematic for topics that extend across specific and distinct local and cultural contexts, such as CBFM, resulting in important literature potentially not being included. In this case, we found that AI was able to support the creation of an inclusive search strategy through rapidly providing a list of non-English terms used to describe CBFM-related terms in the pacific (e.g. Ra'ui (or Rahui); Supplement C), expanding our original search beyond a narrow, western, academic lens. An initial search of the Scopus database of these terms coupled with the ‘fisheries’ key phrase above yielded 82 potential studies not identified in our original search. This AI capability could be used to produce more inclusive search strategies generally and help to identify those with specific relevant local and cultural expertise to advise further. AI does not replace local and contextual knowledge, however, it may provide a draft search string that can be taken to experts and/or actors with a stake in the topic in question.

Screening the articles

The second stage of this study was to evaluate the abilities of AI to screen articles based on a predetermined screening criteria (i.e., inclusion/exclusion). To do so, a list of articles was generated from Scopus using a simplified version of the AI search string, which was revised by the human team (as described in Methods). Here, we present the results of this two-part screening process, (i) pilot screening of 100 randomly selected articles from the previous stage and, after discussions to refine the screening criteria and select the best-performing AI implementation, (ii) screening all 1098 articles from the previous stage. We used quantitative measures to evaluate the abilities of the LLM in this process, including Kappa statistics to explore inter-rater reliability among human team and AI, and calculate the number of disagreements between the human team and AI (see Methods). When tallying disagreements, we define false positives as articles that the human team rejected and the AI accepted, and false negatives as articles that the human team accepted and the AI rejected.

Pilot screening

First, we followed common review practices of checking consistency among reviewers by undertaking a ‘pilot screen’ of 100 articles. All 5 human reviewers and AI independently screened the same 100 articles at the title-abstract level, determining whether the article should be included or excluded based on the screening criteria developed in the previous stage. Initial agreement within the human team in the pilot screen was ‘near perfect’ (Fleiss’ Kappa = 0.85; (Landis & Koch, 1977)), with disagreements over 13 articles (https://github.com/s-spillias/GPT-Screening/blob/main/CBFM/Paper-Results/Pilot_Result.xlsx). These were resolved via discussions within the human team, yielding a ‘consensus’ list of 18 included and 82 excluded articles.

When the AI was prompted to screen articles based on the exact text of the screening criteria used by the human team (Supplement D), there was only fair to moderate agreement with the human’s consensus list (Cohen’s Kappa = 0.21–0.44) and yielded false negative rates of 7–13% across AI screening strategies. We therefore revised the screening criteria fed to the AI (Supplement E) by simplifying the language, adopting a uniform way of expressing each criteria, and iteratively testing prompts using the conflicted studies between the initial human and AI pilot screen until inter-rater reliability scores were similar to those found between the individual human raters. This process substantially improved the agreement with the human pilot screen (Cohen’s Kappa = 0.72–0.86) and decreased the rate of false negatives to 1–4% across AI screening strategies. When AI was included alongside the human reviewers as a sixth collaborator, the inter-rater reliability (Fleiss’ kappa) for all implementations was greater than 0.8 which represents ‘near perfect agreement’ (Landis & Koch, 1977). We decided the ‘best’ performing AI strategy was that with the lowest number of false negatives, which was the ‘Committee with Any Acceptance and Reflection’ (Methods). This strategy only omitted one article that the human team included.

Screening all articles

Out of 1098 possible articles, the human team initially identified 100 that met the screening criteria (9%), whereas the best AI implementation identified 157 articles to include. We present the results for this implementation below, and refer the reader to the associated repository for the performance of all AI implementations (https://github.com/s-spillias/GPT-Screening/blob/main/CBFM/Paper-Results/All_Result.xlsx). The inter-rater reliability between the best AI screen and the human list was high, albeit not as high as the pilot screen (Cohen’s kappa = 0.63). Like the pilot screen, the false negative rate was low (1.2%). However, the false positive rate was higher than the pilot screen (6.5%).

Cross-referencing between this best-performing AI strategy and the human list identified 85 disputed papers (14 false negatives and 71 false positives). Upon further evaluation of these 85 disputed papers, eighteen were decided to be miscategorisations by the human team, with eight being removed from the human list, and ten added, resulting in a collaboratively generated list of 102 titles (Fig. 3). The remaining

67 articles were classified as AI miscategorisations. Of those that were removed from the initial human list after prompting by the AI, all were deemed unlikely to have a relevant case study in one of the PICs. Of those that were added to the initial human list, most had vague mentions of possibly relevant case studies that the AI highlighted. For example, one abstract says, '*Drawing on case studies of the Community Conservation Research Network…*', which along with other mentions of community-based management concepts, was decided to be enough for inclusion. This re-categorisation suggests that the AI can improve the reliability in the literature screening stage by highlighting possibly relevant studies that might be missed by an initial human screen.

This re-assignment improved the inter-rater reliability of the best performing AI screen (Cohen's kappa = 0.71), and decreased the false negative rate to (0.5%) and false positive rate to (5.6%). When the original human screen was compared with this collaborative list, the false negative rate was 1%, whilst the false positive rate was 0.7%. This suggests that the best performing AI implementation's false negative rate was comparable to that of the human team, although the AI's false positive rate was still much higher. Compared to the other AI implementations, we found a trade-off whereby decreasing the false negative rate comes at the cost of increasing the false positive rate, whilst the inter-rater reliability was fairly constant across implementations (Fig. 4).

The remaining six AI false negatives were all kept in the collaborative list because of language in their abstracts that hinted at a possible related case study. However, the AI consistently rejected them because they did not explicitly mention at least one of the 14 PICs. For example, one of these disputed abstracts concludes with the phrase *'as illustrated in this paper with examples of marine commons'*. The human team agreed that this was enough to meet the geographic screening criteria, whilst the AI consistently articulated that the article should be excluded due to the lack of mention of one of the 14 PICs.

The reasons for the disagreement in the remaining AI false positives (n = 61) can be explored here (https://github.com/s-spillias/GPT-Screening/blob/main/CBFM/Paper-Results/All_Result.xlsx). Of these, 17 were accepted by every AI implementation. In most cases, these do not appear to represent cases of the AI making incorrect assertions, so much as over-generous interpretations. For example, in a paper about aquaculture research and development, when asked if a community-based approach is explored, the AI argues '*Maybe; The abstract mentions the need to gain active participation from resource users, but it is unclear if this will be a community-based approach*'. In general, the most common error was the AI being over-inclusive with respect to the geographic criterion in the screening stage, seeking to include studies that were not based in one of the 14 PICs, but which were tangentially related because a specific nearby region was mentioned. For example, about a paper based in Cambodia, one AI agent decided upon reflection to 'maybe' include the paper because *'… it is relevant to the broader Southeast Asian region and could provide insights for similar contexts'.*

Discussion

These findings suggest that generalised AI tools, such as ChatGPT and potentially other large-language models, can assist in two critical stages of evidence synthesis to accelerate the implementation and improve the reliability of a review. We found that whilst the AI can be helpful in brainstorming relevant terms and can occasionally provide insight into the structure of a search string, it is unable to provide a useful search string without substantial prompting (at least at the current level of development of ChatGPT). For screening articles based on predetermined screening criteria, occasional AI misinterpretations undermine the reliability of a single AI assessment but can be overcome via repeated independent queries. From a resource perspective, this can be invaluable as three independent AI agents were able to automatically apply four screening criteria to more than 1000 articles for an API cost of less than 11 USD in roughly 12 hours.

Using AI can improve search string development but cannot yet be relied upon to independently develop a high-quality search string. This is a less optimistic conclusion than that offered by (Wang et al., 2023), who evaluate ChatGPT's ability to generate search strings for PubMed. Our experience may stem from there being less rigorous ways of describing concepts in the broader field of socio-ecology than those accessible by PubMed. The shortcomings in our study are perhaps unsurprising given that search strings are highly structured and logical formulations, and those representing socio-ecological questions are probably poorly represented in ChatGPT's training dataset. Future work could investigate whether fine-tuning an LLM based on search strings gleaned from evidence syntheses and their respective research topics could improve the AI's 'understanding' of how to compose them. However, it should be noted that the proficiency of the AI in screening for relevant articles in the subsequent stage, could de-emphasise the importance of crafting a high-quality search string by allowing the semi-automation of a much larger number of articles than would be otherwise possible by a human.

Our work extends the findings of (Nakaya et al., 2023), and shows that even in tasks that require textual interpretation, an AI such as ChatGPT can provide a reliable classification of relevant articles across numerous screening criteria. Unlike (Nakaya et al., 2023) however, we show that, for more complex screening criteria, relying on a single agent can provide inconsistent results. Multiple repeat calls to the AI, supported by the random context string described in Section 2.3, can provide a greater level of reliability by reducing the chances of omitting relevant articles, albeit at the expense of returning a greater number of false positives to be screened at a subsequent stage. We have also shown that ChatGPT is able to achieve a high level of inter-rater reliability when compared to a team of humans, and that it is possible for the AI to generate a list of articles that have a false negative rate that is comparable to, or even less than, a team of human screeners. Thus one should, in principle, feel confident in relying upon an AI to function as well as a single researcher in screening a large list of studies.

Whist we have shown that AI can aid in verification and reflection, reducing researcher biases and improving output quality, these results highlight the current need for collaborative (Reeson & Paris, 2021) rather than purely automated AI systems, and underscore the continued role of human researchers in the process of identifying relevant literature in evidence syntheses. For example, although AI can expand inclusivity by equipping researchers with broader terms than they may have otherwise available to them from their own positionality, researchers must know enough to elicit this kind of information, as we found that the AI did not readily offer this information. The expertise of the human team, including social sciences and interdisciplinary skills, is important, along with an understanding of local contexts and inclusivity. Further, outputs from large language models, like GPT, may not reflect the full diversity of views on a topic, even when those are present in its training data (Santurkar et al., 2023), further highlighting the role of collaboration with individuals who possess such knowledge and can provide contextual insight. This is especially important when a task requires meaning that is not represented in text, such as some traditional or experiential knowledge, or language with quickly changing meaning, such as in an emerging scientific discipline because pre-trained models, like GPT3.5/ChatGPT, which generate text based on patterns observed in the language of the past. Whilst we have shown that AI can be trusted to capture relevant literature at a similar rate as human screeners, and generate rationales for its decisions, the fine-scale nuance sought by the human team was not achieved by the AI. This led to the trade-off between false positives and negatives that we observed.

In the rapidly growing space of AI in research practice and particularly evidence synthesis, there is growing debate on finding the right balance between using AI to automate various tasks while maintaining meaningful human engagement (Chubb et al., 2022; Wagner et al., 2022). This balancing act is particularly precarious when considering the goals of undertaking a systematic review, to not only contribute to the literature, identify gaps and produce novel interpretation, but also to learn along the way (e.g., as a phd student, your first task is to do a literature review, enabling you to learn about the breadth and depth of knowledge related to your topic) (Pickering & Byrne, 2014). Undertaking the screening stage of evidence synthesis manually can catalyse important understandings of a broad body of literature, resulting in identification of tangentially relevant papers of interest or could even trigger a topic pivot. Implications and comprehensive understandings of what may be lost through automating the process of performing a review are unclear. Chubb et al (2022), in conversation with leading scholars, frame humans as maintaining the role of ‘generators of meaning’ in the context of AI and research processes. They discuss potential implications of relieving researchers of drudgery tasks, such as literature screening, suggesting that this may limit individuals developing core skills, but on the other hand, through accelerating narrow tasks (e.g. screening) it may leave more room for human creativity and collective knowledge production (Gibbons, 2000). Raising questions of what is lost as well as what is gained for researchers through the automation of processes of evidence synthesis, rather than just the end-products, is a consideration worth investing in. In contrast to automation, a collaborative and iterative human-AI approach, such as the one demonstrated in this study, may address some of these concerns while maintaining the benefits.

We have shown that AI systems have the capability to streamline the search and retrieval of relevant articles, thereby improving the efficiency of the review process. However, this was done within the context of a process designed for human capabilities and limitations and raises the question of whether we can design AI systems that enable better search and retrieval of articles for specific research topic without the need for manually searching databases and screening for relevant articles. Whilst we focused on the peer-reviewed literature, the techniques explored here could be combined with other search engines and methods to gather in grey literature as well, which could improve upon current best practices in machine learning (Berrang-Ford et al., 2021) and which could yield a substantial time and resource saving while also creating richer and likely more reliable reflections of the state of knowledge in fields were the use of grey literature is more prevalent.

Conclusion

Research synthesis is increasingly important as the volume of information balloons beyond what any single researcher or team of researchers can practicably follow. AI has the potential to address several issues facing research groups where capacity is constrained by the volume and complexity of the information to absorb and challenges to address. This paper provides an example of how to do that collaboratively, with a place at the table for both human and AI alike working together and following the same respectful tenets as any interdisciplinary team. We show that AI can provide avenues for broadening effectiveness of a systematic review's search strategy and may omit less than 1% of relevant articles in an automated screen based on predetermined screening criteria - a rate which is similar to human experts conducting the same screen. However, like any new technology both the advantages and disadvantages must be honestly faced so as not to lead to unintended or undesirable consequences. This manuscript provides a method for AI, implemented collaboratively, to accelerate the process of identifying relevant literature for such a research endeavour. In this study, we show that AI can reduce the burden on researchers by providing a reliable and transparent method for identifying relevant peer-reviewed research.

Methods

Developing the search strategy

To search for relevant literature, the human team undertook several steps to develop a search string, following ROSES (RepOrting standards for Systematic Evidence Syntheses) protocol (Haddaway et al., 2018). First, we conducted an initial scoping search of the relevant literature using the bibliographic database Scopus. The initial keyword string consisted of generic terms to describe community-based fisheries management, using terms both known to the authors and those identified through keywords in known studies. The purpose of this initial exploratory search was to identify highly relevant and highly cited papers in the field that could be used as internal checking papers. After this, the human team iteratively developed a search string that resulted in a comprehensive list of search results, while also limiting the number of irrelevant articles (as reported in Results). The human team also developed a screening criterion (i.e., inclusion/exclusion) to reflect the scope, purpose and research questions of the review.

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. 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).

Table 1

Search Strategy Prompt. See Supplement A for the full response.

> You are a researcher about to undertake a scoping review of the existing literature on community-based fisheries management, an approach to fisheries management that gives coastal communities and fishers primary responsibility for managing their coastal resources. In this review, your aim is to explore the perceived (i) benefits of, and (ii) barriers to, the implementation of community-based fisheries management in Pacific Island countries. The first step is to develop a search strategy to locate papers that could be relevant to the topic. You will do this by A) breaking the topic into functional components and then brainstorming a comprehensive list of possible synonyms, related terms, shared embeddings, relevant acronyms, and terms researchers might use for each component, B) Designing an inclusive Boolean search string that incorporates these terms and is able to capture a wide range of relevant literature, and C) creating an exclusive list of inclusion criteria that can be used to screen the titles and abstracts of potential papers. (A) will need to be inclusive enough that no relevant papers are missed. (B) will need to have the correct boolean operators to ensure that the maximum number of relevant articles are returned from the search, and (C) will need to be exclusive enough that only highly relevant papers are able to pass the screening stage.

After the AI and the human team finished developing the search strings and screening criteria, the output of the AI search string was sent to co-authors for reflection. The co-authors then discussed the strengths and weaknesses of the search string developed by AI, using combined knowledge and experience of reviews and the topic. Here, the co-authors also reflected on how the AI generated search string could be used to prompt human reviewers to improve the reliability and inclusiveness of their own search string. As the focus of this study is not to undertake a full systematic review of the literature on CBFM, we refined the search string developed by AI to generate a simple search string that could be used to identify a suitable number of relevant articles to evaluate the ability of AI to screen articles for inclusion.

Screening the articles

The resulting articles from this search were exported from Scopus on May 2, 2023 as an .ris file and saved into the reference management software Endnote X9. Once in Endnote, the articles were exported into a single file in Microsoft Excel 2023 to allow the human team and AI to screen the articles for this review. To determine which articles should be included or excluded in the data extraction stage of the review, the human team and AI screened the articles at the title-abstract level using the screening criteria reproduced in Supplements D & E. Ensuring consistency between reviewers is critical in systematic approaches to literature reviews, as such we undertake the article screening in two stages: by 'pilot screening' a subsample of the articles and identifying inconsistencies between reviewers, before screening the remainder of articles.

For the AI screening, we used the OpenAI chat completion API to access the GPT 3.5 Turbo model. We chose this over the other models because it is currently widely available, and is fast and inexpensive to use (*OpenAI API*, n.d.). For the full code and parameter settings see the related GitHub repository for this paper (https://github.com/s-spillias/GPT-Screening). We developed a general prompt that asked the AI to consider a given Title and Abstract and asked it to evaluate whether the Title and Abstract met a given Screening Criteria. This prompt allowed for three possible inclusion outcomes, 'Yes', 'No', and 'Maybe', and requested an explanation for the decision (Table 2).

Based on preliminary trials, we heavily modified the screening criteria for the AI (compare Supplement D & E) and also explored two techniques to mitigate against erroneous AI output. The first was a 'reflection' technique (Shinn et al., 2023; White et al., 2023), whereby the AI was prompted to deliver a structured response that included an initial response, a reflection on the appropriateness of that response, and then a final response. The second technique we used was to empanel a committee of AI agents by repeatedly asking the same prompt three times (see Supplement F for a discussion of the technique used to elicit unique responses to identical prompts).

Table 2

General prompt used to query OpenAI's ChatGPT. Text in bold was varied according to the Title/Abstract, Screening Criteria, and Randomly Generated String.

Ignore this string: **<Random String>**

You are a reviewer for a research project and have been asked

to assess whether the given paper Title and Abstract meets the following Screening Criteria (SC). In assessing, do not re-interpret the SC, simply assess the SC at face value.

We are only interested in papers that strictly meet the SC.

If not enough information is available, be inclusive as we can follow-up at a later stage.

SC: **<Screening Criterion>**

Task: Given the following Title and Abstract, respond to the Screening Criteria (SC) with the following elements, Initial Response, Reflection on Initial Response, and Final Response. Here is an example of how your response should look:

Format:

SC -

Initial Response: Only respond with a Yes or No; Short explanation as rationale.

Reflection: Is the Initial Response correct? Be concise.

Final Response: Strictly only respond with a Yes or No; Short explanation based on reflection.

Initial Response and Final Response should consist of only a

Yes or No or Maybe followed by a semicolon and a single sentence explanation for your reasoning. Like this:

SC: Final Response; One sentence of reasoning.

Title: **<Title Text>**

Abstract: **<Abstract Text>**

With these two techniques, we tested five distinct strategies for using the AI across nine implementations.

1. Individual Agent - Using a single AI agent's initial reaction. (n = 3)
2. Individual with Reflection - Using a single AI agent's final decision after reflection. (n = 3)
3. Committee of Agent's Initial Response - Accepting a paper with at least one affirmative initial response from any of a committee of three AI agents. (n = 1)
4. Committee of Agents with Reflection - Accepting a paper with at least one affirmative final response from any of a committee of three AI agents. (n = 1)
5. Committee with Reflection and Any Acceptance - As with strategy 4, but accepting any paper with any affirmative response, in either the initial or final response. (n = 1)

In the interest of time, a paper was rejected by the AI screen, and further screening criteria skipped, if, for any screening criteria, there were only 'No's' for all initial and final responses.

Pilot screening

Prior to screening the entire 1,098 articles returned by the database search, we followed common review practices of internal consistency checking by undertaking a 'pilot screen' of 100 articles. That is, the human team and AI each screened the same 100 randomly selected articles from the search results. All 5 reviewers and AI independently screened the 100 articles at the title-abstract level, determining whether the article should be included or excluded based on the screening criteria developed in the previous stage. We then compared the results, by calculating inter-rater reliability (Hallgren, 2012), (i.e., consistency) via Fleiss' kappa statistic using the AI screen as a 'sixth member' of the human team and Cohen's kappa statistic comparing the aggregated human screen with the AI screen. These two methods allow us to compare the reliability of the AI to any individual human and allow us to evaluate the reliability of the AI in independently identifying all of the relevant literature that a team of humans found.

Thereafter, within the human team, a Fleiss kappa statistic was then calculated to check the level of agreement (i.e., the consistency) internally between the 5 human reviewers. The kappa statistic from this human pilot screen was 0.85, indicating a near perfect agreement (Landis & Koch, 1977). At this stage, the human team met to discuss any articles that were a point of difference in screening (i.e., whereby there was not 100% agreement across the 5 reviewers), and to discuss any necessary refinements to the screening criteria as a result. In doing so, a list of included/excluded articles was produced based on the decisions made by the collective human team (referred to as the 'aggregated human team results').

After this, we sought to compare the level of agreement in the pilot screening between the aggregated human team results and the AI results. To do so, we calculated Cohen's kappa statistic between the aggregated human team's list and each AI implementation's result (Hallgren, 2012). For each AI implementation, we calculated the number of false positives and false negatives by assuming the aggregated human team results reflect the reality of the data. Here we assume that, in this context, false positives represent a less egregious error than false negatives because false positives can be further screened at the later full-text stage, while false negatives represent omissions of relevant information. We therefore characterised the 'best-performing' strategy as the strategy that had the lowest number of false negatives compared to the human-generated list.

Screening all articles

The remainder of the screening process for the human team was undertaken using Covidence, an online software for systematic reviews (Babineau, 2014). During this stage, each of the 5 human reviewers screened approximately 200 articles each, indicating whether an article should be included or excluded based on the predefined screening criteria. To ensure comprehensive screening, any 'maybe' articles were accepted for full-text review and data extraction. The AI screening was undertaken using the same five strategies listed above. After completion of the screening by the human team and the AI, we cross-referenced the output list from the best-performing AI strategy with the human-generated list to create a list of disputed articles. The titles and abstracts from these disputed articles were then re-screened and discussed by the lead authors and re-assigned to create a 'collaborative list' of included articles. As with the pilot screen, once the collaborative list was compiled, we evaluated the outcome of the full screen by calculating the inter-rater reliability between the collaborative list and the AI and human list using Cohen's Kappa Statistic and by tallying the number of false positives and false negatives.

Declarations

Data availability

The raw Scopus output is available in the GitHub repository at https://github.com/s-spillias/GPT-Screening/blob/main/CBFM/Input/1_all.xls

Code availability

The source code of this work is freely available in the GitHub repository https://github.com/s-spillias/GPT-Screening

Competing interests

The authors declare no competing interests.

Author Contributions

SS and PT conceived the study and methods. SS wrote the code and performed the analyses. SS, MA, FB, RT contributed to developing prompts and AI screening strategies. PT, JD, DK, RS, RAJ undertook the human-driven half of the review. All authors contributed to writing and editing the manuscript.

Acknowledgements

The authors would like to thank Andrew Wilkins for constructive comments on the manuscript.

References

1. Babineau, J. (2014). Product review: Covidence (systematic review software). Journal of the Canadian Health Libraries Association/Journal de l'Association Des Bibliothèques de La Santé Du Canada, 35(2), 68–71.
2. Berdejo-Espinola, V., & Amano, T. (2023). AI tools can improve equity in science. Science, 379(6636), 991–991. https://doi.org/10.1126/science.adg9714
3. Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F., Haddaway, N. R., Haines, A., & Dangour, A. D. (2021). Systematic mapping of global research on climate and health: A machine learning review. The Lancet Planetary Health, 5(8), e514–e525.
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
5. Chubb, J., Cowling, P., & Reed, D. (2022). Speeding up to keep up: Exploring the use of AI in the research process. AI & SOCIETY, 37(4), 1439–1457. https://doi.org/10.1007/s00146-021-01259-0
6. Cooke, S. J., Cook, C. N., Nguyen, V. M., Walsh, J. C., Young, N., Cvitanovic, C., Grainger, M. J., Randall, N. P., Muir, M., Kadykalo, A. N., Monk, K. A., & Pullin, A. S. (2023). Environmental evidence in action: On the science and practice of evidence synthesis and evidence-based decision-making. Environmental Evidence, 12(1), 10. https://doi.org/10.1186/s13750-023-00302-5
7. De La Torre-López, J., Ramírez, A., & Romero, J. R. (2023). Artificial intelligence to automate the systematic review of scientific literature. Computing. https://doi.org/10.1007/s00607-023-01181-x
8. Doulman, D. J. (1993). Community-based fishery management. Marine Policy, 17(2), 108–117. https://doi.org/10.1016/0308-597X(93)90025-X
9. Duggan, J., Cvitanovic, C., & van Putten, I. (2023). Measuring sense of place in social-ecological systems: A review of literature and future research needs. Ecosystems and People, 19(1), 2162968.
10. Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., & Hilton, J. (2017). Living systematic review: 1. Introduction—The why, what, when, and how. Journal of Clinical Epidemiology, 91, 23–30.
11. FAO. (2022). FAO in the Pacific 2021—Annual Report of FAO Subregional Office for the Pacific Islands. FAO. https://doi.org/10.4060/cc0061en
12. Gibbons, M. (2000). Mode 2 society and the emergence of context-sensitive science. Science and Public Policy, 27(3), 159–163.
13. Gillett, R., & Tauti, M. I. (2018). Fisheries of the Pacific Islands (No. 625; FAO Fisheries and Aquaculture Technical Paper). FAO. https://www.fao.org/3/i9297en/i9297en.pdf
14. Haddaway, N. R., Bernes, C., Jonsson, B.-G., & Hedlund, K. (2016). The benefits of systematic mapping to evidence-based environmental management. Ambio, 45(5), 613–620.

https://doi.org/10.1007/s13280-016-0773-x

15. Haddaway, N. R., Bethel, A., Dicks, L. V., Koricheva, J., Macura, B., Petrokofsky, G., Pullin, A. S., Savilaakso, S., & Stewart, G. B. (2020). Eight problems with literature reviews and how to fix them. Nature Ecology & Evolution, 4(12), 1582–1589. https://doi.org/10.1038/s41559-020-01295-x
16. Haddaway, N. R., Macura, B., Whaley, P., & Pullin, A. S. (2018). ROSES RepOrting standards for Systematic Evidence Syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. Environmental Evidence, 7(1), 7. https://doi.org/10.1186/s13750-018-0121-7
17. Haddaway, N. R., & Westgate, M. J. (2019). Predicting the time needed for environmental systematic reviews and systematic maps. Conservation Biology, 33(2), 434–443. https://doi.org/10.1111/cobi.13231
18. Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology, 8(1), 23.
19. Karcher, D. B., Cvitanovic, C., Colvin, R. M., van Putten, I. E., & Reed, M. S. (2021). Is this what success looks like? Mismatches between the aims, claims, and evidence used to demonstrate impact from knowledge exchange processes at the interface of environmental science and policy. Environmental Science & Policy, 125, 202–218.
20. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159–174.
21. Mallett, R., Hagen-Zanker, J., Slater, R., & Duvendack, M. (2012). The benefits and challenges of using systematic reviews in international development research. Journal of Development Effectiveness, 4(3), 445–455.
22. Member States – AOSIS. (2021). https://www.aosis.org/about/member-states/
23. Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Medical Research Methodology, 18(1), 143. https://doi.org/10.1186/s12874-018-0611-x
24. Nakaya, Y., Higaki, A., & Yamaguchi, O. (2023). ChatGPT’s ability to classify virtual reality studies in cardiology. European Heart Journal - Digital Health, 4(3), 141–142. https://doi.org/10.1093/ehjdh/ztad026
25. OpenAI API. (n.d.). Retrieved 6 June 2023, from https://platform.openai.com
26. Pickering, C., & Byrne, J. (2014). The benefits of publishing systematic quantitative literature reviews for PhD candidates and other early-career researchers. Higher Education Research & Development, 33(3), 534–548.
27. Pullin, A. S., & Stewart, G. B. (2006). Guidelines for Systematic Review in Conservation and Environmental Management. Conservation Biology, 20(6), 1647–1656. https://doi.org/10.1111/j.1523-1739.2006.00485.x

28. Reeson, A., & Paris, C. (2021, November 30). What's the secret to making sure AI doesn't steal your job? Work with it, not against it. The Conversation. http://theconversation.com/whats-the-secret-to-making-sure-ai-doesnt-steal-your-job-work-with-it-not-against-it-172691
29. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? (arXiv:2303.17548). arXiv. http://arxiv.org/abs/2303.17548
30. Shaib, C., Li, M. L., Joseph, S., Marshall, I. J., Li, J. J., & Wallace, B. C. (2023). Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success).
31. Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning (arXiv:2303.11366). arXiv. http://arxiv.org/abs/2303.11366
32. Steenbergen, D. J., Raubani, J., Gereva, S., Naviti, W., Arthur, C., Arudere, A., Ham, J., Joy, L., Lalavanua, W., Neihapi, P., Seko, A., Terashima, H., & Andrew, N. L. (2022). Tracing innovation pathways behind fisheries co-management in Vanuatu. Ambio, 51(12), 2359–2375. https://doi.org/10.1007/s13280-022-01788-y
33. Thomas, J., McDonald, S., Noel-Storr, A., Shemilt, I., Elliott, J., Mavergames, C., & Marshall, I. J. (2021). Machine learning reduced workload with minimal risk of missing studies: Development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. Journal of Clinical Epidemiology, 133, 140–151. https://doi.org/10.1016/j.jclinepi.2020.11.003
34. Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. Journal of Information Technology, 37(2), 209–226.
35. Wang, S., Scells, H., Koopman, B., & Zuccon, G. (2023). Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? (arXiv:2302.03495). arXiv. http://arxiv.org/abs/2302.03495
36. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT (arXiv:2302.11382). arXiv. http://arxiv.org/abs/2302.11382
37. Wyborn, C., Louder, E., Harrison, J., Montambault, J., Montana, J., Ryan, M., Bednarek, A., Nesshöver, C., Pullin, A., & Reed, M. (2018). Understanding the impacts of research synthesis. Environmental Science & Policy, 86, 72–84.

Figures

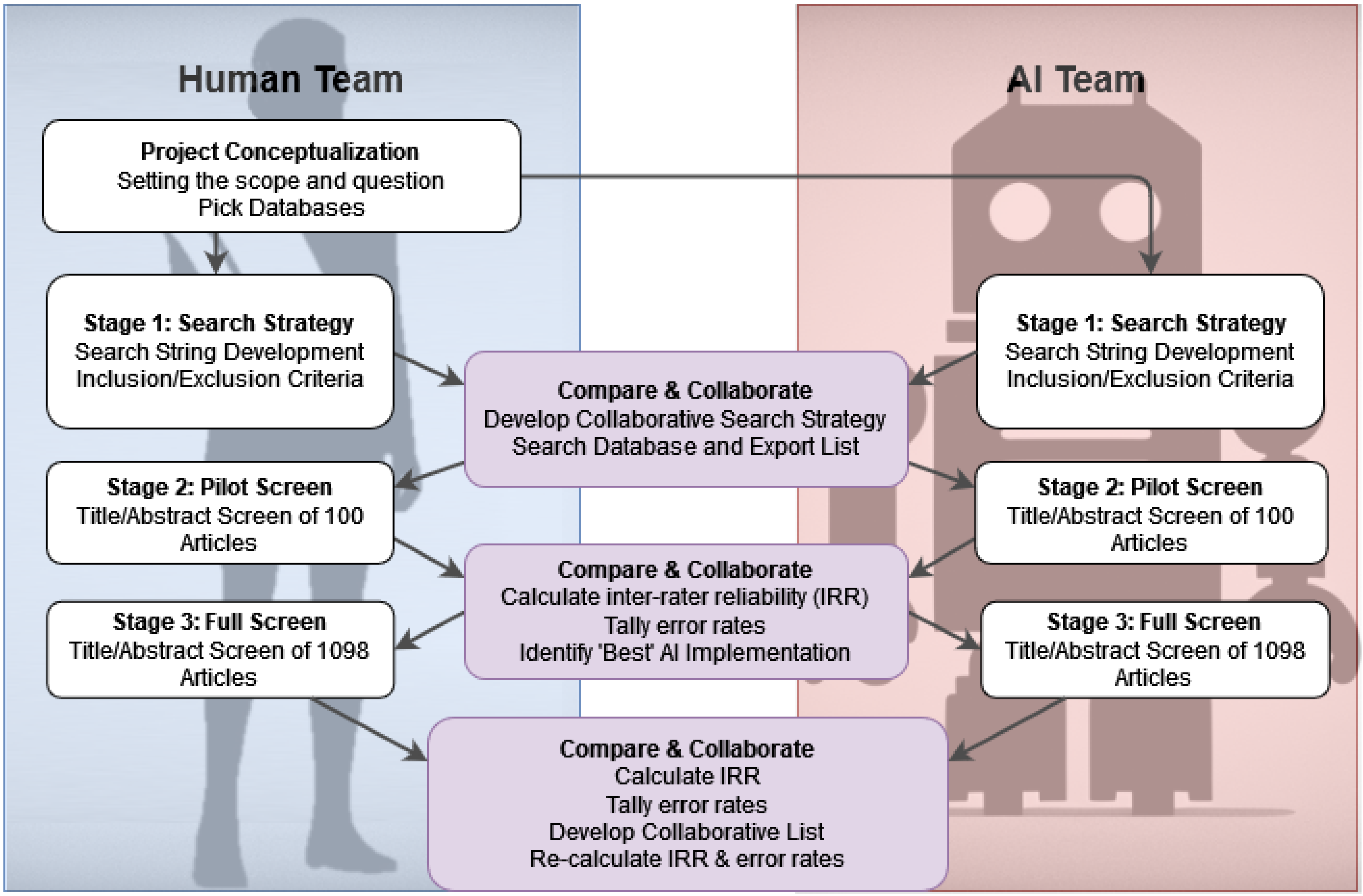
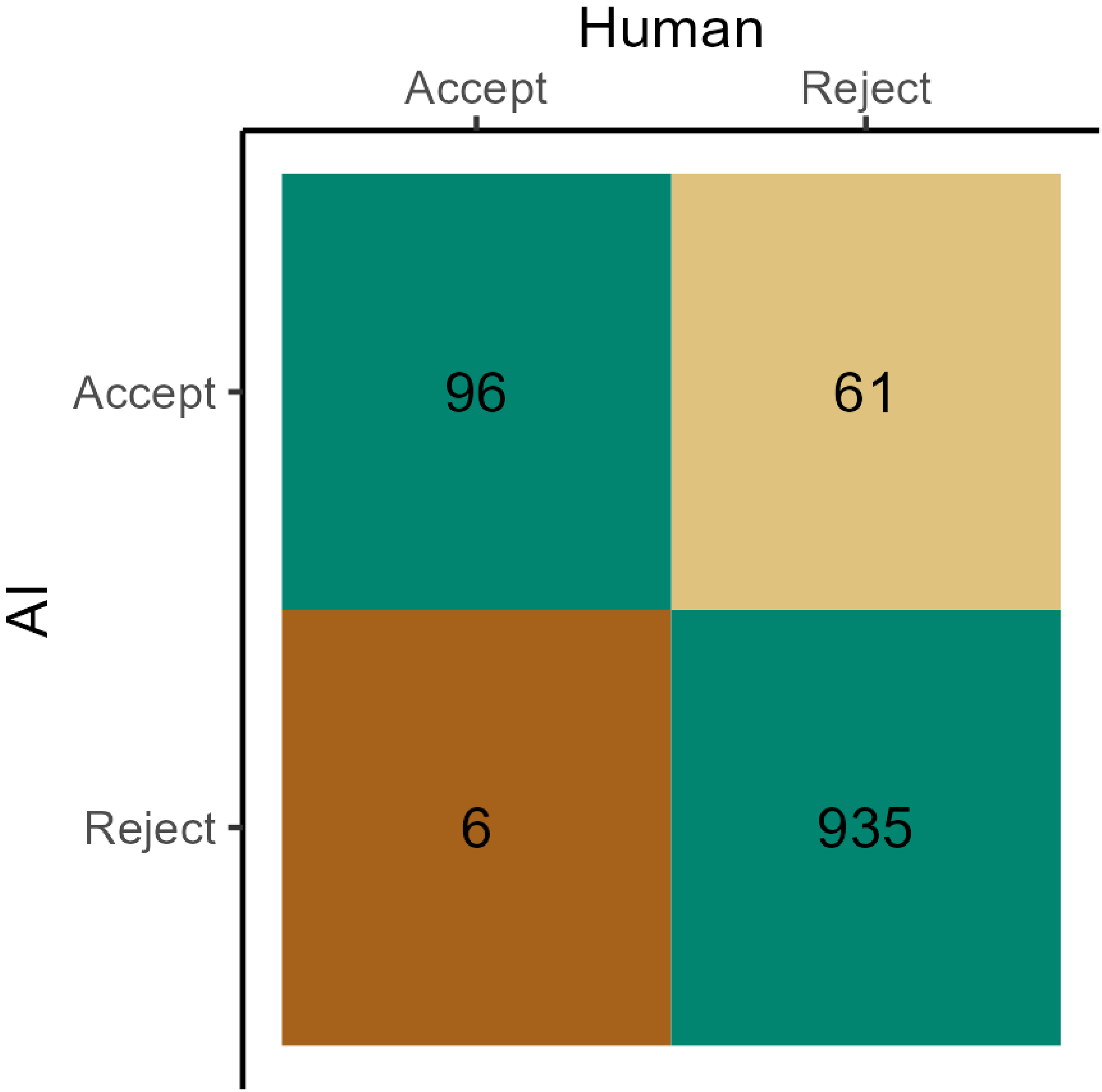

Human Team
AI Team
Project Conceptualization
Setting the scope and question
Pick Databases
Stage 1: Search Strategy
Search String Development
Inclusion/Exclusion Criteria
Compare & Collaborate
Develop Collaborative Search Strategy
Search Database and Export List
Stage 1: Search Strategy
Search String Development
Inclusion/Exclusion Criteria
Stage 2: Pilot Screen
Title/Abstract Screen of 100 Articles
Stage 2: Pilot Screen
Title/Abstract Screen of 100 Articles
Compare & Collaborate
Calculate inter-rater reliability (IRR)
Tally error rates
Identify 'Best' AI Implementation
Stage 3: Full Screen
Title/Abstract Screen of 1098 Articles
Stage 3: Full Screen
Title/Abstract Screen of 1098 Articles
Compare & Collaborate
Calculate IRR
Tally error rates
Develop Collaborative List
Re-calculate IRR & error rates

Figure 1

Methodological Approach. A team of researchers (‘human team’) (PT, JD, DK, RS, RAJ) systematically searched and screened literature relevant to the topic of Community-Based Fisheries Management. In parallel, a team of researchers used a large language model (‘AI Team’) (SS, MA, FB, RT) to facilitate the same process. After each stage of this process, the human team and AI team reconvened to compare and reflect on the outputs, and identify opportunities for collaboration to improve the robustness of the search and screen.



		Human	
		Accept	Reject
AI	Accept	96	61
	Reject	6	935

Figure 2

Number of titles included and excluded by human and AI screeners. The best performing AI implementation ‘Committee with Reflection and Any Acceptance’, which minimises the false negative rate, compared to the final collaborative list. Top left and bottom right boxes (green) represent agreement between Human and AI screeners. Top right and bottom left boxes represent disagreements. We characterise the bottom left (orange) as false negatives, and view them as worse errors than top right (tan) false positives, because false positives can be screened away in the next full-text stage, whereas false negatives represent lost relevant data.

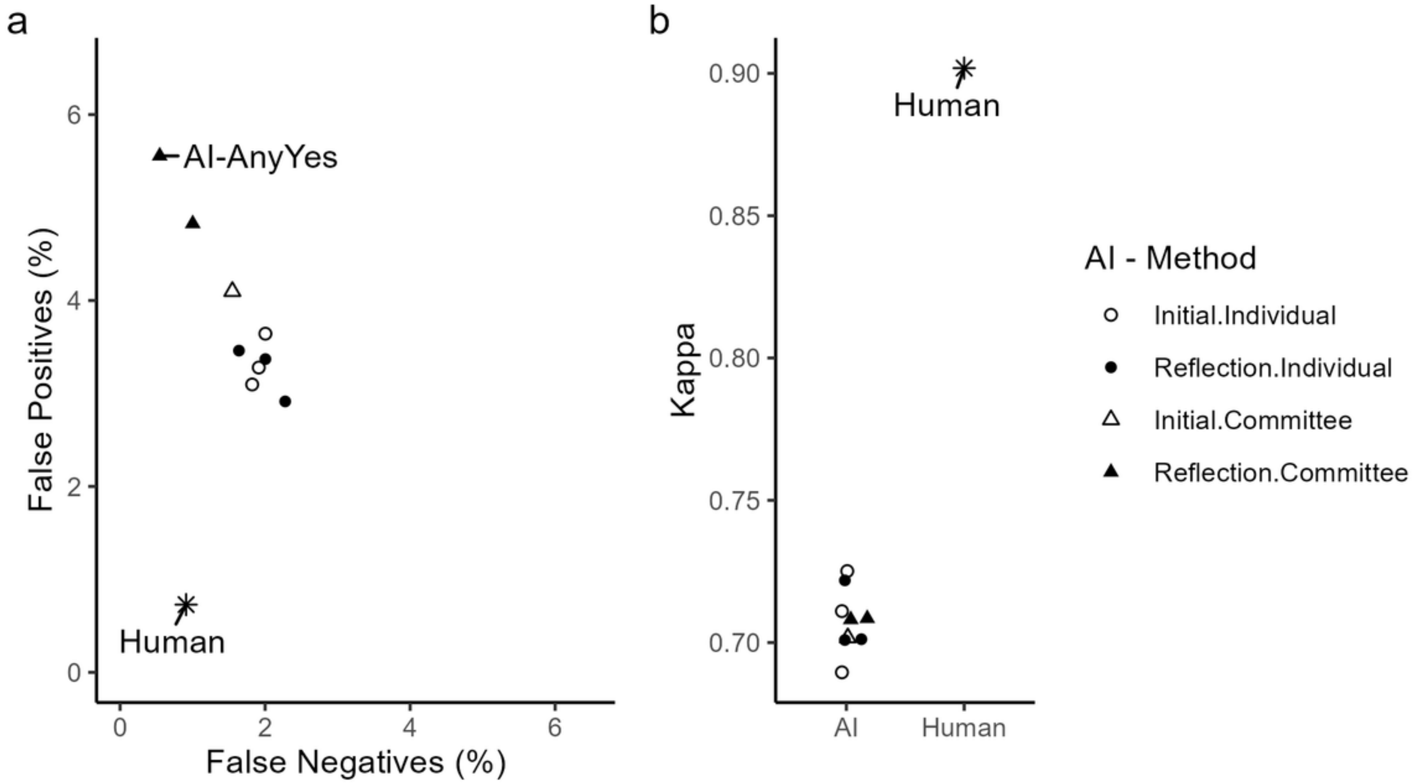


Figure 3

Implications of different AI processing techniques. (a) The trade-off between not missing relevant articles (False Negatives), and including articles that are not relevant (False Positives) when AI outputs are compared to the final collaborative list of included articles. Incorporating reflection for a single agent has little impact on the overall quality of the screen. Whereas, incorporating a committee of AI agents (n=3), decreases the number of relevant articles that are missed, at the cost of increasing the number of extra articles that will have to be screened at the next phase of the review. Human error rates compared to the final collaborative list are denoted by the (*). (b) Impact of incorporating reflection and 'committee' in screen decision-making. We include Cohen's kappa statistic calculated between the initial aggregated human list and the final collaborative list that incorporates the output of the best-performing AI implementation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1.png
- SupplementaryMaterials.docx