Using Elicit AI research assistant for data extraction in systematic reviews: a feasibility study across environmental and life sciences

Authors:

Malgorzata Lagisz[1,2], Ayumi Mizuno[2#], Kyle Morrison[1#], Pietro Pollo[1,3#], Lorenzo Ricolfi[1#], Yefeng Yang[1#], Shinichi Nakagawa[1,2#]

* Correspondence: M. Lagisz; e-mail: losialagisz@gmail.com

[#] These authors contributed equally and are listed in alphabetical order

Affiliations:

[1] Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of New South Wales, Kensington, NSW, 2052, Australia

[2] Department of Biological Sciences, University of Alberta, CW 405, Biological Sciences Building, Edmonton, AB T6G 2E9, Canada

[3]School of Environmental and Life Sciences, University of Newcastle, Newcastle, Australia

Short title: Evaluation of data extraction in Elicit AI

Data and code availability: project GitHub repository with all data and code can be found at https://github.com/mlagisz/elicit_extractions_testing (it will be archived to Zenodo when manuscript is accepted for publication).

ABSTRACT

Data extraction in systematic reviews, maps and meta-analyses is time-consuming and prone to human error or subjective judgment. Large Language Models offer potential for automating this process, yet their performance has been evaluated in a limited range of platforms, disciplines, and review types.

We assessed the performance of the *Elicit* platform across diverse data extraction tasks using journal articles from seven systematic-like reviews in life and environmental sciences. Human-extracted data served as the gold standard. For each review, we used eight articles for prompt development and another eight for testing. Initial prompts were iteratively refined to exceed 87% accuracy or up to five rounds. We then tested extraction accuracy, reproducibility across user accounts, and the effect of *Elicit*'s high-accuracy mode.

Of 90 considered prompts, 70 exceeded the 87% accuracy when compared to gold standard values but tended to be lower when tested on a new set of articles. Repeating data extractions with different *Elicit* user accounts resulted in 90% agreement on extracted values, though supporting quotes and reasoning matched in only 46% and 30% of cases, respectively. In high-accuracy mode, value matches dropped to 77%, with just 10% quote matches and 0% reasoning matches. Extraction accuracy did not differ by data types. *Elicit* also helped identify eight (<1%) errors in the gold standard data.

Our results show that *Elicit* can complement, but not replace, human data extractors. *Elicit* may be best used as a secondary reviewer and to evaluate the clarity of data extraction protocols. Prompts must be fine-tuned and independently validated.

HIGHLIGHTS

What is already known: Data extraction in systematic reviews is labour-intensive and prone to error. LLMs like *Elicit* are being explored as tools to automate this step, though evaluations remain limited in scope and robustness.

What is new: We assessed *Elicit*'s accuracy and repeatability across seven reviews in life and environmental sciences. While Elicit achieved high accuracy for some variables, performance varied and was sensitive to prompt design, user account and algorithm change.

Potential impact for Research Synthesis Methods readers: *Elicit* can support systematic reviews as a secondary extractor when paired with human oversight. Our study offers practical guidance on integrating LLM tools while highlighting current limitations in replicability and reasoning.

KEYWORDS

Artificial intelligence, evidence synthesis, systematic maps, meta-analysis, research methods, proof of concept

Research Synthesis Keywords

data extraction, AI, LLMs, automation

1 | INTRODUCTION

The use of Artificial Intelligence (AI) in automating evidence synthesis is growing and is expected to bring transformative changes to research evidence synthesis. Producing robust evidence synthesis of any type takes a significant amount of time, sometimes years [1]. This is due to the ever-expanding size of the evidence base, and the labour required to maintain high standards of work across all stages of the process.

Large Language Models (LLMs), a type of AI processing and learning from vast amounts of data, have been proposed as a promising way to automate or semi-automate tasks across all stages of systematic reviews. Application of increasingly efficient and sophisticated LLMs could help researchers to keep pace with the growing demands of funders and end-users for timely and efficient evidence syntheses [2,3]. However, there are also concerns that the use of AI-based solutions may be compromising the quality of evidence synthesis, potentially introducing errors and producing biased and not reproducible systematic reviews, evidence maps, meta-analyses, and other forms of evidence [3–5]. To address these concerns and evaluate applications of AI and LLMs in systematic reviews, case studies and reviews are being published at a rapid pace (e.g., [2,3,5–14]).

A recent scoping review [8] based on 37 articles on LLMs use in health research systematic reviews suggested that LLMs have been applied mainly in the three key stages of the systematic review process: literature searching (41% of articles), study selection (38%), and data extraction (30%). OpenAI's Generative Pre-trained Transformer (GPT) was the most frequently tested model, featured in 89% of articles. Findings on LLMs performance were mixed as around half of

the studies viewed LLMs as promising, a quarter were neutral, and one-fifth found them unhelpful.

Another recent systematic review [10] based on 172 articles on the use of LLMs in evidence review automation revealed similar trends. Most articles explored automation of a particular stage of review, focusing mainly on literature searching (35%), screening (33%) and data extraction (31%). The majority of articles expressed positive views on using LLMs in reviews (70%), while 43 articles (25%) reflected mixed or cautious perspectives, and 9 articles (5.2%) reported negative experiences with LLMs. Concerns included limited extraction accuracy for numeric data and low search and screening accuracy for bibliographic data, potentially linked to high hallucination rates (generating false references).

Over the years, studies and overviews on the use of AI in systematic reviews have consistently highlighted the need for further investigation to keep abreast with the rapidly evolving landscape of AI tools (e.g., [2,3,5,8–10,14]). More specifically, these publications emphasized the importance of evaluating the effectiveness, accuracy, and validity of individual AI tools and how they change over time. It is also critical to understand the limitations of these technologies and how they might influence the outcomes of evidence synthesis. In particular, the need to assess the impact of AI on reliability and reproducibility of systematic reviews remains a key area for research.

Elicit (elicit.ai; elicit.com) is one of such technologies holding a great promise of streamlining and accelerating systematic reviews [13,15]. It stands out from the generic tools like *OpenAI ChatGPT*, *Microsoft Copilot*, or *GoogleGemini*, because it has been designed specifically for academic use and especially for evidence synthesis of published academic literature (notably,

other similar platforms are being rapidly created, e.g., *SciSpace*, *PICOportal*, *Paperguide*). *Elicit* draws its data from an extensive *Semantic Scholar* database of over 100 M publications [16]. To assist with the searching and screening phases of evidence synthesis, *Elicit* uses LLM algorithm to evaluate semantic similarity of publication texts to find and rank publications for inclusion in a systematic review. *Elicit* also allows uploading and analysing PDF files of full-text documents that may be not available in its online database. Further, *Elicit* applies other LLMs that are trained on academic literature to generate article summaries and extract data that is commonly collected for systematic reviews, e.g., on the study subjects, interventions, exposures, reported outcomes and measurements [12]. Importantly, *Elicit* has a user-friendly interface, which makes it easy to generate and refine search and data extraction prompts [15]. However, two recent case studies point to limitations in accuracy and repeatability of *Elicit*-based literature searches and screening [12,17]. For data extractions, one study deemed data extraction from 33 papers on fisheries management to be on par with human ability, outperforming two GPT models [18]. Similarly, *Elicit* and *chatGPT* (GPT-4o model) extracted the right data about 90% of the time from 30 health-related research articles [11]. In contrast, another study deemed almost half of the values extracted by *Elicit* as valid but missing important details, and 4% as invalid, based on a sample of seven variables and 20 healthcare-related studies [13]. These case studies are limited by small sample sizes both in terms of numbers of tested variables and test articles, warranting more extensive and in-depth exploration, especially outside medical fields.

This paper explores the potential applications of *Elicit* platform in evidence synthesis with two specific aims:

1. To evaluate the effectiveness of LLMs integrated in *Elicit* for data extraction from full-text published research studies in predefined formats, and to assess the feasibility of using *Elicit* as a

supplementary platform for data extractions for systematic review in ecology, evolutionary biology and environmental sciences.

2. To evaluate the repeatability and consistency of data extraction results when using an independent *Elicit* user account or a different LLM algorithm.

We also provide recommendations on the use of this innovative technology.

2 | METHODS

We aligned our project with the recommendations on the responsible AI use in evidence synthesis [19]. Our project follows a protocol registered on OSF at https://doi.org/10.17605/OSF.IO/48PGE. Our reporting of author contributions adheres to the MERIT framework [20]. We have considered the practicality and affordability of the *Elicit* platform to a wider range of users globally when designing our project. We conducted all statistical analyses in the R v.4.5.0 statistical environment [21].

We set up the project to reflect *Elicit* functionality available via a *Plus* subscription plan (*Elicit Plus* was 12 USD per month or 120 USD per year to subscribe, https://support.elicit.com/en/articles/471617, as on 2024/09/30). *Plus* subscription plan may be a relatively affordable option for researchers who wish to evaluate suitability of *Elicit* platform for their data extraction (and other) needs before committing to more pricey plans with greater allowances for data volumes, or even perform actual data extractions on this plan. *Elicit Plus* plan is sufficient for conducting one-off small- to medium-scale evidence syntheses, with less than 300 PDFs to be extracted (ecological and evolutionary meta-analyses may typically include a median of 23 studies [22], 8 at a time, with up to 5 variables per extraction table. Thus, we have explicitly tailored our study plan to match these user subscription plan specifications.

2.1 | Datasets and variables

We based our project on the existing data from seven published systematic-like reviews (systematic reviews and maps, umbrella reviews and meta-analyses, hereafter "systematic reviews"; see [23]) on diverse topics from life and environmental sciences (**Table S1**; [24–30]; all reviews were pre-registered and published with accompanying data which was either double-extracted or cross checked by a second researcher). The lead authors of these systematic reviews are co-authors of this project. They provided underlying extracted raw data and meta-data for all full-text articles included in their reviews and contributed to the study planning and evaluation phases of this project.

When planning our data re-extractions for evaluating the *Elicit* platform, we focused on variables representing study scope, design and reporting quality. This mainly included qualitative data (e.g., study species, locations, types of exposure/intervention) and some quantitative data (e.g., number of included primary studies, chemical concentrations, and study durations), which are likely to be explicitly reported as text in the included articles. We did not extract numerical values used to calculate effect sizes because such values are often presented in figures or tables and *Elicit Plus* currently does not offer such extraction capabilities. We also considered variables related to reporting quality and presence of elements that could facilitate more detailed manual data extractions (e.g., presence of funding statement, conflict of interest statement, supplementary materials, raw data, analytical code). If variables related to reporting quality were not extracted in an original systematic review or meta-analysis, two researchers (ML and either AM, KM, LR, PP, SN, or YY) independently extracted them to create a gold standard answer for each additional variable. Overall accuracy of this additional human-based data extractions was

98.6% (1.4% error rate, i.e. 6 mismatched values between two human extractions out of 416 de-novo double-extracted values for 26 variables across all 7 original systematic reviews - these were usually variables related to reporting quality - presence of author contribution statement, conflict of interest statement, supplementary materials, data availability). In two cases one of the human extractors missed information on authors contributions, in another two cases on data sharing, once on the presence conflict of interest statement, and once on supplementary materials.

We followed a pre-defined extraction process in *Elicit* across all seven reviews, as feasible. All data extractions were based on the main full text of published articles uploaded as PDF files into *Elicit* user workspace. The *Elicit* workspace facilitates processing of uploaded files and data extraction via standardised input boxes (variable name, description and answer structure) and we considered this functionality when designing and recording our project workflow.

2.2 | Project workflow

Our main study workflow is presented in **Figure 1**. In brief, from each seven original systematic reviews, ML randomly sampled 8 included full-text studies for conducting prompt development (training set) and another 8 full-text studies for evaluation (test set). ML used the training set to create and iteratively refine data extraction prompts in *Elicit* for 10 variables per review that exceeded 87% accuracy (i.e. 7 / 8 correct answers per variable; as pre-defined in the study protocol) when compared to the gold standard answers, while allowing a maximum of five prompt development iterations per variable. Data extraction prompts that exceeded 87%

accuracy at this development phase were then used to extract data from a new set 8 included full-text studies (TEST phase). We then repeated the test extraction using a different *Elicit* user account (RETEST) and using *Elicit* in a high accuracy mode (HATEST).

ML has documented the prompt development process (DEV phase), noting any alterations to the original meta-data and data from the seven published systematic reviews (e.g., pooling or changing to free text option when the number of categories of a variable exceeded the limit. of 8 categories allowed in *Elicit*) and other encountered issues (**Table S2)**. This documentation denotes the number of prompt development iterations per variable, and lists variables that failed to reach the threshold of 87% accuracy, as well as new variables that were considered as replacement of the unsuccessful variables. The table also includes initial descriptions of the variables, based on the original meta-data from the reviews, as used to construct initial data extractions prompts in *Elicit*, and final prompts developed using *Elicit*.

In the main testing phase (TEST), ML used the test set of 8 studies (different from the DEV phase) per review to evaluate accuracy of *Elicit* extractions using the final prompts from the development phase for 10 variables per review. Here, again we compared the answers provided by *Elicit* to human-extracted gold standard data.

In order to evaluate repeatability of extractions across user accounts in *Elicit*, the other authors (AM, KM, LR, PP, SN, YY) repeated extractions from the TEST phase using their separate *Elicit* user accounts. This re-testing phase (RETEST) allowed us to compare extractions conducted by two different *Elicit* users using the same set of studies (full-text pdf files) and data extraction prompts (TEST-RETEST comparison).

However, before we analysed our data, *Elicit* upgraded its underlying algorithm resulting in our TEST results potentially being not representative of *Elict*'s performance and how it changes over time (we note that Elicit does not publicly share technical details of their models). In order to evaluate repeatability between different versions of *Elicit* algorithms, ML repeated all extractions from the TEST phase. This high accuracy-testing phase (HATEST) allowed us to compare extractions conducted by two different versions of *Elicit* algorithms using the same set of studies (full-text PDF files) and data extraction prompts (TEST-HATEST comparison).

ML exported tables with data extracted by *Elicit* as CSV files and manually compared them with the human-extracted values from the completed systematic reviews (gold standard). We did not automate these comparisons because it was necessary to account for deviations from the data formats, typos, partial matches and semantically equivalent answers (e.g., "not reported", "not explicitly mentioned", "none", "no", "-") and interpreting free-text answers provided by *Elicit*. We recorded the number of matching (equivalent) answers for each combination of variable and systematic review. We then used "Supporting quotes for …" and "Reasoning for …" fields from *Elicit* to elucidate the reasons for any discrepancies between *Elicit* extractions and the human-extracted gold standard data. If we detected actual errors made by human extractors, we corrected the gold standard data, as necessary, and adjusted our assessments of extraction accuracy accordingly.

2.3 | Analyses and reporting

We summarised results overall and for each of the seven original systematic reviews or meta-analyses used in this project. Further, we also considered results across the type of extracted data (Yes/No, number, categorical or free text string answers) and by the category of extracted information (study scope / design vs. reporting practice assessment).

We expressed "Accuracy" of extractions by *Elicit* as percent of the matching pairs of values out of the total pairs of values used in a given comparison (*Elicit* vs. gold standard). For the prompt development phase, we expressed "Success" as instances where accuracy exceeded the threshold of 87% (i.e. at least 7 out of 8 values per variable were extracted correctly in *Elicit*). For the testing phase (TEST) and two re-testing phases (RETEST, HATEST), we reported accuracy as well as reasons for potential mismatches.

2.4 | Derivations from the protocol

Our project deviated from the registered protocol in five ways, as outlined below.

First, we intended to implement *Elicit* extraction as two tables with 5 columns for each systematic review. However, in practice, it was more convenient to use one table with one column to work on the extraction variables one at a time and to export results for a single variable after each iteration of testing. This procedural deviation from the protocol does not influence the premise of the project or its results.

Second, we excluded studies where the PDF failed to parse on the upload to *Elicit*. These were usually old studies stored in PDF files containing scanned text or text in other non-parsable formats.

Third, rather than rating data extracted by *Elicit* as either match (fully matching human-extracted data), partial (matching some but not all available information extracted by humans), or mismatch (completely different answer from human-extracted data), we simplified the coding to binary values (match = 1) and (mismatch = 0). This change was necessary for interpreting free-text answers from *Elicit* and the cases where *Elicit* provided multiple answers instead of a required single answer for some of the categorical variables (this feature appeared to be outside of the user's control in *Elicit*). For variables with more than eight categories in the original studies, we had to pool some of the categories together during prompt development, because *Elicit* allows defining up to eight categorical answer options. Depending on the variable, partial matches were still informative (i.e. could be considered as a match), but not for others (mismatch), depending on the context. We documented such special considerations as comments on variables during prompt development (**Table S2**).

Fourth, on 16 May 2025, *Elicit* announced that all columns would now default to "high accuracy" to provide "the most reliable paper extractions across all plans". Previously, *Basic*/*Plus Elicit* plan users had limited access to high accuracy columns (one per table). Change to a high accuracy algorithm can be expected to significantly influence accuracy of data extractions. To test this expectation, on 26 June 2025 we repeated all extractions from the testing phase (HATEST phase; with the high accuracy algorithm) and compared the results to the earlier

extractions from the testing phase (TEST; without the high accuracy algorithm) and against the human-extracted gold standard answers.

Fifth, we had to exclude from our analyses three variables that failed in RETEST and HATEST phases due to human error in prompt specifications (misspecified answer structures). This deviation reduced our total sample sizes to 536 (out of planned 560) answer values for analyses in these two project stages only.

3 | RESULTS

3.1 | Prompt development effort and success rates in Elicit

The overall prompt development success rate was 78% (70 out of 90 considered variables reached accuracy > 87% within 5 iterations). Per review, this success rate varied from 71% to 91% (**Table S3**). In other words, we had to try 11 to 14 variables per review in order to get 10 "Successful" variables. While this was achieved within the first iteration for 30 of the 90 tested variables, 28 of the successful variables took 2 iterations, and remaining 12 variables were successful after 3 to 4 iterations. In contrast, we ran 5 iterations of prompt refinement for each of the 20 "non-successful" variables that failed to reach desired accuracy level. This means that 100 iterations of data extractions did not lead to success, out of a total of 231 iterations ran across the whole development phase (i.e. 43% of effort or time wasted).

Next, we categorised all our variables into 4 types, based on the expected structure of the answer (data type) (**Figure 2**). Variables that required a Yes / No answer comprised 49% of our data set in this phase (44 tested variables) and had an overall success rate of 82%. We had 27 variables with predefined categorical answers (2 to 11 categories, e.g. sex or age class of study subjects, type of study design), which had an overall success rate of 70%. We had 13 variables that required *Elicit* to extract names or answer with other non-predefined text ("any answer" option in Elicit; e.g., used to extract species, database or software names, or measures with their measurement units), achieving an overall success rate of 92%. We also had 6 variables that required extraction of a single number only (e.g., number of included primary studies, number of species, number of experimental doses or cues), but they succeeded only half of the time. There

was no statistically significant relationship between variable type and its chance of success or failure during the prompt development phase of the project (Pearson's Chi-squared test; Chi-squared = 5.54, $df = 3$, $p = 0.14$).

The majority (37 out of 50) of the variables we attempted to extract were specific to only one of our seven systematic reviews. We applied the remaining 13 variables in more than one review, with exactly the same initial prompt. Of these, we had four variables that were used and were successful across all 7 reviews (**Figure 3**). These four variables coded whether a study explicitly stated authors contributions, authors conflict of interests (or lack of such conflict), registered protocol, and supplementary materials (all as a Yes / No answer).

Elicit also successfully extracted mentions of the PRISMA flowchart in two reviews and mentions of the use of reporting guidelines in one review where this information was relevant (umbrella reviews, i.e. overviews of reviews). *Elicit* did not perform well extracting information on data availability (only succeeded for 2 out of 6 reviews) and disclosure of funding sources (only succeeded for 1 out of 6 reviews) (**Figure 3**). The remaining variables that we used were directly related to study scope or methods and were usually specific to a particular review (i.e. used only once) and had a mixed success during the prompt development phase (**Figure 3**). Variables related to reporting quality (authors contributions, conflict of interests statement, funding sources, supplementary materials, registered protocol, PRISMA flowchart, reporting guideline, data availability) were as likely as variables related to study scope or methods (e.g., study species, location, sample size, exposure dosage or duration, study design type) to be successful during the prompt development phase of the project (77% and 78% success rate, respectively; Chi-squared = 0, $df = 1$, $p = 1$).

3.2 | Test phase success and accuracy of Elicit

During the main testing phase (TEST) we used Elicit to extract 70 variables that were deemed successful during the development phase. We used a new sample of 8 studies from each of the seven systematic-like reviews, with 10 variables tested per review (successful variables from the prompt development phase).

Overall, 48 out of the 70 of the tested variables (69%) reached the expected accuracy threshold of at least 87% at this stage (i.e. at least 7 / 8 *Elicit* answers assessed as matching human-extracted gold standard answers). Across the seven reviews, the overall success rates ranged from 50% (Yang et al. 2024 and Ricolfi et al. 2024) to 100% (Lagisz et al. 2020) and the extraction accuracy varied across the variables (**Figure 4**).

The overall extraction accuracy during the prompt development phase and test phase were positively related ($r = 0.38$, $t = 3.38$, df = 68, $p = 0.001$), indicating that more accurate data extractions during the development phase were more likely to be also accurate during the testing phase. Similarly to the prompt development phase, there was no association between the type of the extracted variable and its success (i.e. extraction accuracy above 87%; Chi-squared = 3.88, $df = 2$, $p = 0.275$).

We compared the accuracy of *Elicit*'s data extraction with the accuracy of independent data re-extractions by two human researchers for the 26 variables that were added to the reviews in this project. Using the equivalent subsets of values from the main testing phase, we found that human

extractors had fewer mismatches with each other (4 / 208 = 1.9%) than *Elicit* had with human gold standard answers (22 / 208 = 10.6%; Chi-squared = 11.856, $df = 1$, $p = 0.0006$). For 13 out of 26 variables, Elicit performed as good as humans (no mismatches), for 12 variables it made more mistakes than humans, and for one it performed better (detecting one mention of sharing study data in Mizuno et al. 2024 review data set).

3.3 | Repeating Elicit data extractions with independent user accounts

To test whether data extractions are repeatable in *Elicit*, we repeated all extractions from the test phase (TEST) using a different *Elicit Plus* user account (RETEST phase). We found that almost 90% of the RETEST-extracted values matched exactly TEST-extracted values (476 out of 536; **Figure 5A**). Both extraction rounds also failed to extract 10 values (2%) and RETEST missed another value (which was correctly extracted in TEST).

Most (36 out of 50) mismatched values were due to *Elicit* interpretation error. These mismatches occurred because the TEST extraction was correct, but RETEST extraction was incorrect (19), the TEST extraction was incorrect, but RETEST extraction was correct (10), and both the TEST extraction and the RETEST extraction were incorrect (6). The remaining 15 cases of mismatches were due to differences in wording or formatting of the extracted values, but the answers could be considered as semantically equivalent and correct (**Figure 5B**).

For the supporting quotes provided by *Elicit* to justify its extracted values, 200 of the quotes matched between TEST and RETEST, but 248 did not (**Figure 5C**). Further 88 of the data

extractions had at least one of the supporting quotes missing (18 in TEST and 16 in RETEST, 54 both in TEST and RETEST).

For the reasoning narratives provided by *Elicit* to justify its extracted values, 158 of the cases did match between TEST and RETEST, but 377 did not (**Figure 5D**). Reasoning text was missing for only one of the data extractions (TEST).

Finally, in the RETEST phase, overall accuracy was 85.6% across 67 variables and seven reviews, when we compared *Elicit*-extracted data to human-extracted gold standard answers. This value was not statistically different from the accuracy of the data extractions for the equivalent set of variables in the TEST phase (86.6%, Chi-squared = 0.125, $df = 1$, $p = 0.724$).

3.4 | Repeating Elicit data extractions in high-accuracy mode

To test whether using high accuracy mode affected accuracy, we repeated all extractions from the test phase (TEST) after all *Elicit* accounts got upgraded to free high accuracy mode (HATEST phase). We found that almost 77% of the values re-extracted after the change to high accuracy matched exactly values that were extracted earlier (412 out of 536; **Figure 6A**). Both extraction rounds also failed to extract 9 values (2%) and HATEST missed another two values (which were correctly extracted in TEST).

Most (94 out of 122) mismatched values were due to *Elicit* interpretation error. These mismatches occurred because the TEST extraction was correct, but HATEST extraction was incorrect (60), the TEST extraction was incorrect, but HSTEST extraction was correct (26), and

both the TEST extraction and the HSTEST extraction were incorrect (8). We further subdivided these mismatches into cases where TEST extraction was correct but HATEST extraction incorrect (60), cases where TEST extraction was incorrect, but HATEST extraction correct (26), and cases where both TEST extraction and HATEST extraction were incorrect (8). The remaining 28 cases of mismatches were due to differences in wording or formatting of the extracted values, but the answers could be considered as semantically equivalent and correct (**Figure 6B**).

For the supporting quotes provided by *Elicit* to justify its extracted values, only 51 of the quotes matched between TEST and HATEST, while 412 did not (**Figure 6C**). Another 73 data extractions had at least one of the supporting quotes missing (71 in TEST and one in HATEST, one both in TEST and HATEST).

For the reasoning narratives provided by *Elicit* to justify its extracted values, none of the cases did match between TEST and HATEST (**Figure 6D**). Reasoning text was missing for only one of the data extractions (TEST).

Finally, in the HATEST phase, overall accuracy was 82.1% across 67 variables and seven reviews, when we compared *Elicit*-extracted data with human-extracted gold standard answers. This value was lower and almost statistically different (at $p = 0.05$ threshold) from the accuracy of the data extractions for the equivalent set of variables in the TEST phase (86.6%, Chi-squared = 3.734, $df = 1$, $p = 0.053$).

3.5 | Correcting gold standard values based on Elicit data extractions

Human extracted data, even if double-extracted or cross-validated, can contain errors (e.g., due to missing relevant information in the study text or coding errors). After cross-checking mismatches between *Elicit*'s data extractions and our gold standard values, we detected a total of 8 human-made errors (<1%; **Table S4**). We corrected these errors and accounted for them when calculating accuracy scores for the data extractions using *Elicit*.

4 | DISCUSSION

4.1 | Overview of results

Our work revealed four key aspects of using *Elicit* for data extractions for systematic reviews. First, during the prompt development phase, we were unable to reach our extraction accuracy threshold (87%) for 20 out of 90 tested extraction variables within five iterations of prompt refinement. Importantly, we found that meta-data (i.e., variable descriptions) from original reviews were often vague, requiring considerable effort to create clear and precise prompts for *Elicit*. This raises some concerns about the reusability and repeatability of data from published reviews and calls for more detailed documentation of data extractions performed by researchers. Second, the accuracy of nearly one-third of the variables declined when we applied the same prompts to a new set of studies during the testing phase. Third, when we repeated data extractions using independent *Elicit* user accounts with identical prompts and source files, 90% of the extracted values (476 out of 536) were consistent across accounts. However, supporting quotes and reasoning provided by *Elicit* matched in only 46% and 30% of cases, respectively. Fourth, when using *Elicit*'s high-accuracy algorithm mode, 77% of the extracted values (412 out of 536) exactly matched those from the earlier test. However, overall accuracy was slightly lower (82%). Supporting quotes matched only 10% of the time, and the wording of reasoning behind the extractions changed completely, with no textual overlap (0% match).

4.2 | Comparison with other studies of Elicit

Our findings are broadly consistent with other studies evaluating AI technologies for data extraction. In particular, our work complements three recent studies that assessed the use of *Elicit* as a data extraction tool for systematic reviews.

Spillas et al. (2025) conducted a pilot data extraction from 33 papers on fisheries management using 11 questions (variables) [18]. Ten out of 11 data extraction questions solicited open-text answers from Elicit, which were then subject to qualitative analysis in order to compare them with human extracted information. The remaining question was a categorical variable with a set of pre-defined categories. If the questions could not be answered based on the content of the paper being extracted, the paper was excluded from the evaluation pool, which differs from our approach (we allowed missing information). The quality of information extracted from the remaining 33 studies was manually graded using a three-point scale (Poor, Fair, Good), potentially introducing subjective bias. Overall, this approach revealed an acceptable level of performance of *Elicit*, similar to human ability. In line with our results, extraction quality varied among variables. This variation was not associated with the question difficulty as perceived by human extractors.

Bianchi et al. (2025) extracted seven variables from 20 randomized controlled trials across several healthcare-related topics [13]. They reported that Elicit significantly deviated from human extractions in 4% of cases, while 46% were classified as "partially equal"—meaning they were generally valid but missing important details. Accuracy varied across variables, with "Study design" with the best performance, and "Interventions" and "Intervention effects" with the worst performance. The authors emphasized that human verification remains essential when using *Elicit* for data extraction.

Helms Andersen et al. (2025) used *Elicit*'s high-accuracy mode to extract 180 values from 30 articles, achieving an overall accuracy of 91% [11]. Accuracy was highest for population characteristics (100%) and study design variables (100%) but dropped to 73% for review-specific variables. The authors identified five cases (6%) of "hallucinations", where the LLM generated values were not present in the full text or misrepresented them through incorrect labelling or rounding.

Together, these independent evaluations, along with our study, highlight both the strengths and limitations of using *Elicit* for research evidence synthesis. In our workflow, we encountered several types of challenges, including those related to developing effective extraction prompts, assessing the accuracy of extracted data, and identifying sources of error. Below, we outline concerns related to data extractions in *Elicit*.

4.3 | Potential hallucinations in Elicit

As noted above, Helms Andersen et al. (2025) have already reported cases of hallucinations in *Elicit* [11]. While we did not explicitly track which mismatches were caused by hallucinations, we observed several instances where *Elicit* extracted incorrect values that appeared to result from this issue. For example, *Elicit* occasionally detected the presence of conflict of interests statements or author contribution statements in studies in which such statements were absent.

4.4 | Misinterpretations and overinterpretations in Elicit

Elicit may interpret available information differently from human extractors. For example, we observed numerous cases where *Elicit* interpreted mentions of ethics approval as evidence of a registered study protocol. However, human researchers typically treat these as distinct categories of documentation. In most of the assessed studies, *Elicit* failed to identify funding statements, even though such statements are usually standardized and easily recognized by human extractors. As an example of overinterpretation, *Elicit* (in its high-accuracy mode) reasoned that if the methods section mentioned two authors conducting screening or data extraction, this justified coding that author contributions were explicitly stated. However, this does not constitute a full author contribution statement covering all aspects of the study (the intended meaning of the coded variable). Other extraction errors appeared to stem from the complexity of the published studies themselves, particularly those with intricate designs or multiple experiments reported within a single article. In such cases, *Elicit* often struggled to determine which parts of the study met the inclusion criteria for a systematic review and to correctly match relevant information across different sections (e.g., methods vs. results).

4.5 | Effects of answer structures and formats in Elicit

Elicit offers a range of default pre-trained extraction variables that return short free-text summaries, as well as options to define custom variables using one of three available answer structures: Any Answer (free text), Yes/No/Maybe (fixed categorical), and Specified (categorical with a maximum of 8 answer options). While free-text answers provide flexibility, they are problematic for systematic maps and meta-analyses, which require variables to be coded in strictly categorical or numeric formats to support data analysis and visualization. To achieve this, free-text responses must be either automatically or manually parsed and re-coded, introducing

extra steps and the potential for error. For example, in our study, we extracted the names of software or databases used. These variables were coded categorically in some of the reviews, but due to *Elicit*'s limitations (e.g., inability to specify more than 8 categorical options and naming variations in the studies), we had to switch to free-text answer structures. We then manually matched the extracted free-text responses to the gold standard data, accounting for partial matches, a process that introduced a degree of subjectivity. Additionally, for multilevel categorical variables, *Elicit* does not allow restricting responses to a single value (i.e. *Elicit's* answer can have one or more values per study, while it is possible to request from a human extractor to only select one most representative or relevant value). This sometimes resulted in multiple values being extracted by *Elicit* per variable per study, which increased the likelihood of partial matches rather than exact matches with the gold standard data.

4.6 | Limitations in Elicit's access to required data

In some cases, the information needed for extraction is not present in the main text of a study or is located in parts of the document that are inaccessible to *Elicit*. For example, *Elicit* currently cannot extract data from figures, and table extraction is unavailable to users on lower-tier accounts. Additionally, relevant information may be contained in supplementary materials, raw data files, or metadata hosted on external platforms, which *Elicit* cannot access. Even when data is available, it may require additional processing, such as filtering, calculations, or unit conversions, to produce accurate values (e.g., effect sizes for specific subsets). Human extractors are generally better equipped to locate, interpret, and process such obscure or complex information sources accurately.

4.7 | Gaps in article meta-data extracted by Elicit

Missing or incomplete meta-data retrieval from uploaded PDF files is another noticeable issue. We observed that *Elicit* often failed to extract key meta-data from uploaded PDF files. In particular, it was unable to retrieve DOIs or DOI links entirely and only partially extracted other fields such as journal names, publication years, and, in some cases, even article titles. As a result, we had to rely on stored file names to match data extractions to specific studies across all our trials in *Elicit*.

4.8 | Older or less informative documents

Older articles are sometimes available in file formats that contain less structured or accessible information. Over time, publication file standards, particularly PDF formats, have evolved to include richer metadata and improved text structure. While *Elicit* can process most modern PDF formats, it may struggle to extract information from older versions, such as scanned text-based PDFs. This can compromise the completeness and accuracy of the extracted data. When conducting case studies on *Elicit*'s performance, image-only or scanned PDFs were excluded in our and other works [11] and do not contribute to the performance evaluation scores.

4.9 | Limitations and strengths of our study

Our study has limitations related to sample size and selection of extracted variables. The number of studies evaluated per review in each phase was relatively small, which limits the precision of our estimates of success rates and extraction accuracy. Additionally, we did not attempt to

extract numerical values that represent (or that could be used to calculate) effect sizes, as these are often presented in figures or tables, which *Elicit* currently cannot process.

The strengths of our study lie in its robustness and transparency. We followed a pre-registered systematic workflow, including detailed cross-validation and thorough documentation across all phases. The systematic reviews used as the basis for our data extractions covered diverse topics and disciplines, and the variables extracted varied in answer structure and the level of interpretation required.

5 | CONCLUSIONS

Our study revealed the effort required to develop prompts that reach a predefined level of accuracy in data extractions, variation in accuracy of extractions across variables in an independent test set, repeatability of data extractions across user accounts, and effects of LLM algorithm change on *Elicit* platform. The human-usable variable descriptions (meta-data) often require iterative refinement and testing in order to achieve accuracy on par with human extractors. Despite its shortcomings, *Elicit* offers accessible advanced functionalities for extracting data from the full text of research studies and can support this stage of the systematic review process. However, challenges remain in ensuring the quality and replicability of extracted data, particularly when information is not explicitly reported, presented in inaccessible formats, or requires nuanced interpretation. We recommend that *Elicit* could be integrated into a modified systematic review workflow as a secondary extractor alongside a human reviewer. A

third reviewer could then reconcile discrepancies between human- and *Elicit*-extracted data, thereby improving efficiency while maintaining high accuracy.

DATA AND CODE AVAILABILITY

All data and code used in this study are available at https://github.com/mlagisz/elicit_extractions_testing and https://osf.io/ejyva/ [final version to be archived on Zenodo - link to be added after acceptance].

FUNDING

This study was supported by funding from ARC (Australian Research Council) Discovery Project grant DP210100812 and DP230101248 (ML, SN) and Canada Excellence Research Chair Program CERC-2022-00074 (SN). The original datasets used for deriving gold standard data (extracted manually by humans) were funded, as acknowledged in the relevant articles.

CONFLICT OF INTEREST

We acknowledge that we used temporary free *Elicit Plus* plan access provided by Elicit Research, PBC, to conduct tests from different user accounts. Representatives of Elicit Research, PBC, did not participate in conceptualisation or designing of this study. We did not receive any

financial payments from Elicit Research, PBC, and have no other relationships or activities that could have influenced our work on this project.

AUTHOR CONTRIBUTIONS

Conceptualisation: ML, SN

Data curation: ML

Formal analysis: ML

Funding acquisition: ML, SN

Investigation: ML, AM, KM, PP, LR, YY, SN

Methodology: ML, SN

Project administration: ML

Software: ML

Supervision: ML, SN

Visualisation: ML

Writing – original draft: ML

Writing – review & editing: ML, AM, KM, PP, LR, YY, SN

ORCID

Malgorzata Lagisz: 0000-0002-3993-6127

Ayumi Mizuno: 0000-0003-0822-5637

Kyle Morrison: 0000-0002-3700-2398

Pietro Pollo: 0000-0001-6555-5400

Lorenzo Ricolfi: 0000-0001-7101-3309

Yefeng Yang: 0000-0002-8610-4016

Shinichi Nakagawa: 0000-0002-7765-5182

USE OF AI STATEMENT

During the preparation of this work, the authors used *Elicit* to extract data from published articles, following a pre-registered protocol. The authors wrote the first draft of the manuscript and used OpenAI ChatGPT 4o to improve the readability of their own writing. The author(s) reviewed and edited the content and take full responsibility for the final manuscript.

REFERENCES

1. Haddaway NR, Westgate MJ. Predicting the time needed for environmental systematic reviews and systematic maps. *Conserv Biol J Soc Conserv Biol*. 2019;33(2):434-443. doi:10.1111/cobi.13231

2. Clark J, Barton B, Albarqouni L, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. *Res Synth Methods*. 2025;16(4):601-619.

doi:10.1017/rsm.2025.16

3. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. *JCPP Adv*. 2024;4(2):e12234. doi:10.1002/jcv2.12234

4. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8(1):163. doi:10.1186/s13643-019-1074-9

5. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ Open*. 2023;13(7):e072254. doi:10.1136/bmjopen-2023-072254

6. Cao C, Arora R, Cento P, et al. Automation of Systematic Reviews with Large Language Models. Published online June 19, 2025:2025.06.13.25329541. doi:10.1101/2025.06.13.25329541

7. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Syst Rev*. 2024;13(1):219. doi:10.1186/s13643-024-02609-x

8. Lieberum JL, Toews M, Metzendorf MI, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. *J Clin Epidemiol*. 2025;181:111746. doi:10.1016/j.jclinepi.2025.111746

9. Ofori-Boateng R, Aceves-Martins M, Wiratunga N, Moreno-Garcia CF. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artif Intell Rev*. 2024;57(8):200.

doi:10.1007/s10462-024-10844-w

10. Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J Am Med Inform Assoc JAMIA*. 2025;32(6):1071-1086. doi:10.1093/jamia/ocaf063

11. Helms Andersen T, Marcussen TM, Termannsen AD, Lawaetz TWH, Nørgaard O. Using Artificial Intelligence Tools as Second Reviewers for Data Extraction in Systematic Reviews: A Performance Comparison of Two AI Tools Against Human Reviewers. *Cochrane Evid Synth Methods*. 2025;3(4):e70036. doi:10.1002/cesm.70036

12. Bernard N, Sagawa Y, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicit. *BMC Med Res Methodol*. 2025;25(1):75. doi:10.1186/s12874-025-02528-y

13. Bianchi J, Hirt J, Vogt M, Vetsch J. Data Extractions Using a Large Language Model (Elicit) and Human Reviewers in Randomized Controlled Trials: A Systematic Comparison. *Cochrane Evid Synth Methods*. 2025;3(4):e70033. doi:10.1002/cesm.70033

14. Blaizot A, Veettil SK, Saidoung P, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. *Res Synth Methods*. 2022;13(3):353-362. doi:10.1002/jrsm.1553

15. Kung J. Elicit (product review). *J Can Health Libr Assoc J Assoc Bibl Santé Can*. 2023;44(1). doi:10.29173/jchla29657

16. Elicit's source for papers. Accessed July 25, 2025. https://support.elicit.com/en/articles/553025

17. Lau O, Golder S. Comparison of Elicit AI and Traditional Literature Searching in Systematic Reviews using Four Case Studies. Published online June 27, 2025:2025.06.17.25329772. doi:10.1101/2025.06.17.25329772

18. Spillias S, Ollerhead KM, Andreotta M, et al. Evaluating generative AI for qualitative data extraction in community-based fisheries management literature. *Environ Evid*. 2025;14(1):9. doi:10.1186/s13750-025-00362-9

19. Thomas J, Flemyng E, Noel-Storr A. Responsible AI in Evidence Synthesis (RAISE): guidance and recommendations. Published online August 29, 2024. doi:10.17605/OSF.IO/FWAUD

20. Nakagawa S, Ivimey-Cook ER, Grainger MJ, et al. Method Reporting with Initials for Transparency (MeRIT) promotes more granularity and accountability for author contributions. *Nat Commun*. 2023;14(1):1788. doi:10.1038/s41467-023-37039-1

21. R Core Team. R: A Language and Environment for Statistical Computing | BibSonomy. 2024. Accessed July 25, 2025. https://www.R-project.org/

22. Yang Y, Noble DW, Senior AM, Lagisz M, Nakagawa S. Interpreting prediction intervals and distributions for decoding biological generality in meta-analyses. *eLife*. 2025;14. doi:10.7554/eLife.103339.1

23. Moher D, Stewart L, Shekelle P. All in the Family: systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. *Syst Rev*. 2015;4:183. doi:10.1186/s13643-015-0163-7

24. Ricolfi L, Vendl C, Bräunig J, et al. A research synthesis of humans, animals, and

environmental compartments exposed to PFAS: A systematic evidence map and bibliometric analysis of secondary literature. *Environ Int*. 2024;190:108860. doi:10.1016/j.envint.2024.108860

25. Pollo P, Lagisz M, Yang Y, Culina A, Nakagawa S. Synthesis of sexual selection: a systematic map of meta-analyses with bibliometric analysis. *Biol Rev Camb Philos Soc*. 2024;99(6):2134-2175. doi:10.1111/brv.13117

26. Samarasinghe G, Lagisz M, Santamouris M, et al. A visualized overview of systematic reviews and meta-analyses on low-carbon built environments: An evidence review map. *Sol Energy*. 2019;186:291-299. doi:10.1016/j.solener.2019.04.062

27. Morrison K, Yang Y, Santana M, Lagisz M, Nakagawa S. A systematic evidence map and bibliometric analysis of the behavioural impacts of pesticide exposure on zebrafish. *Environ Pollut Barking Essex 1987*. 2024;347:123630. doi:10.1016/j.envpol.2024.123630

28. Mizuno A, Lagisz M, Pollo P, Yang Y, Soma M, Nakagawa S. A systematic review and meta-analysis of anti-predator mechanisms of eyespots: conspicuous pattern vs eye mimicry. *eLife*. 2024;13. doi:10.7554/eLife.96338.2

29. Lagisz M, Zidar J, Nakagawa S, et al. Optimism, pessimism and judgement bias in animals: A systematic review and meta-analysis. *Neurosci Biobehav Rev*. 2020;118:3-17. doi:10.1016/j.neubiorev.2020.07.012

30. Yang Y, Liu Q, Pan C, et al. Species sensitivities to artificial light at night: A phylogenetically controlled multilevel meta-analysis on melatonin suppression. *Ecol Lett*. 2024;27(2):e14387. doi:10.1111/ele.14387

FIGURE LEGENDS AND FIGURES

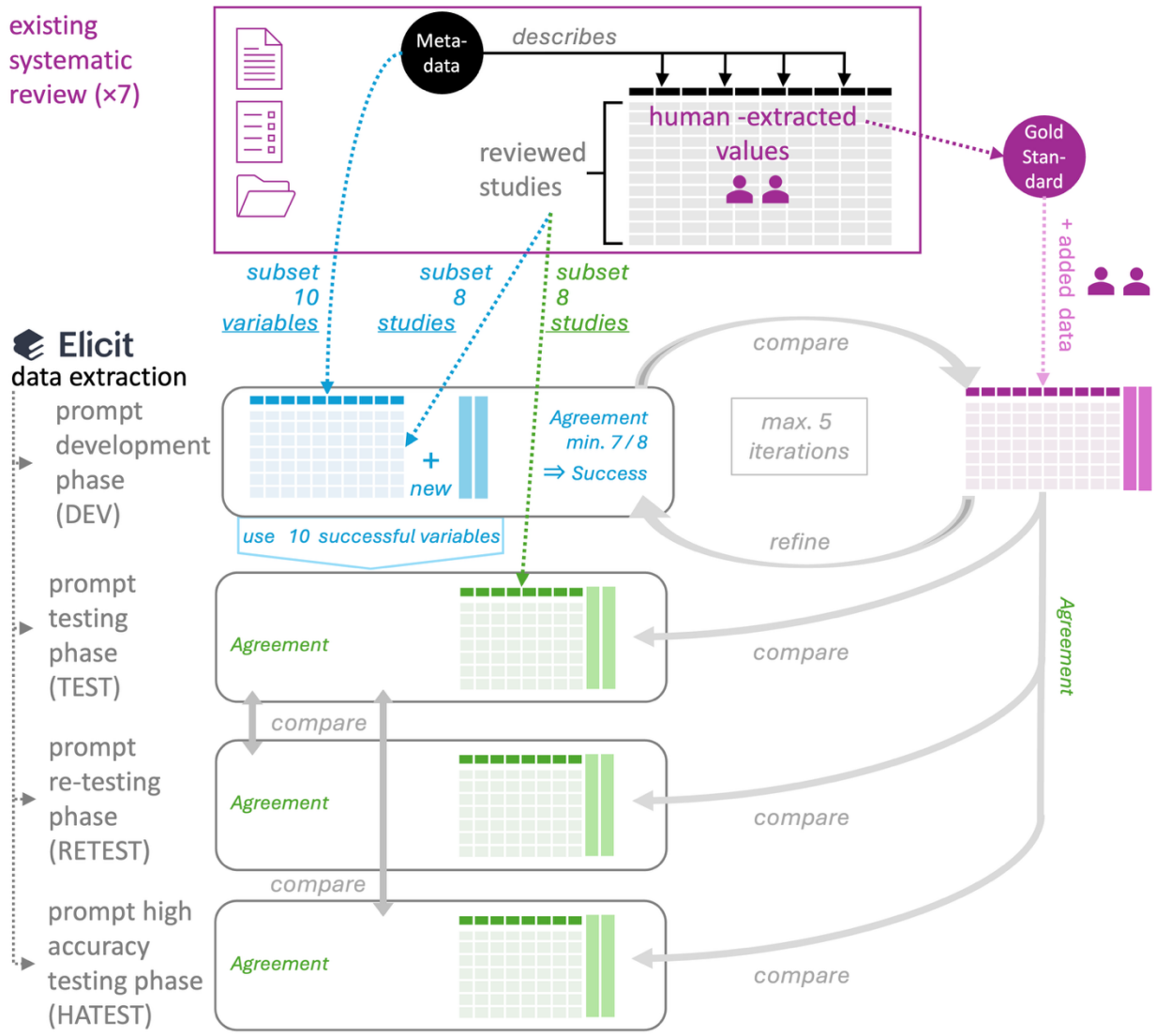

Figure 1.

Diagram of our approach used to develop and evaluate data extractions on the *Elicit* platform. We started the project from seven systematic-like reviews (systematic maps, meta-analyses, and umbrella reviews) representing different topics in ecology, evolution, and environmental sciences. We used human-extracted values from these reviews as our gold standard data and meta-data as a starting point for developing data extraction prompts in *Elicit*. From each review, we randomly selected eight included articles for the prompt development phase (DEV) and

another eight for the three testing phases (TEST, RETEST, HATEST). During the prompt development phase, we iteratively refined this prompt until reaching >87% agreement with the gold standard data or fifth iteration. We replaced extraction variables that did not reach this criterion with new variables until we had ten sufficiently accurate variables per review. Selected variables coded study design and methods, presence of supplementary materials, contributorship and conflict of interest statements, and other review-specific information. In the TEST phase, we evaluated accuracy of *Elicit* extractions on the set of eight studies (per review) that were not used for prompt development. In the RETEST phase, we re-ran the test using the same prompts and studies, but a different *Elicit* user account to test replicability of the extractions. In the HATEST phase, we ran another test using high accuracy mode to assess whether it improved accuracy of the data extractions. For detailed description of the underlying data sets and workflow phases see Methods section.

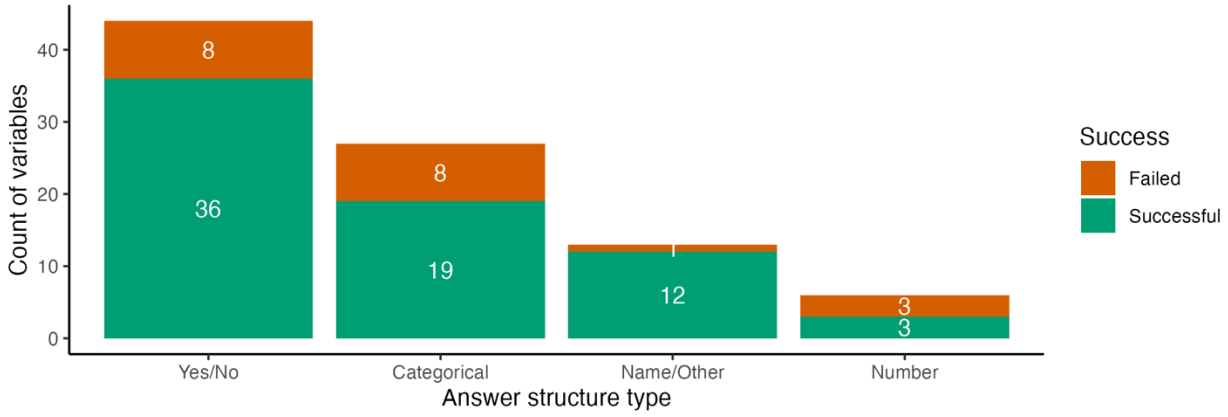

Figure 2.

Distribution and success of extraction variables by the type of expected answer during the prompt development phase of the project. A variable was considered as "Successful" if at least 7 out of 8 answer values were extracted correctly in *Elicit*, i.e. matched human-extracted gold standard values, within a maximum of 5 iterations of data extraction prompt refinement. The data underlying this prompt development stage comprises 90 variables considered across 7 systematic reviews, with 8 studies extracted per review. "Categorical" answer structure includes categorical variables (e.g., "Tissue measured" variable being coded as Blood, Pineal, SCN, Urinary, Retina, Water; "Age" being coded as Juvenile, Adult), except the binary Yes / No answers which are shown separately. We used "Yes//No" answer structure to code presence or absence of certain information or practice in a study (e.g., presence of a conflict of interests statement or whether raw data is shared). "Name/Other" answer structure includes variables where only a name (or names, if relevant) had to be extracted (e.g., species, software, database used in a study), or other atypical data (e.g. a measure and a unit quantifying exposure level or duration), which is equivalent to "free text" extraction specified "Any answer type" in *Elicit*. "Number" answer structure includes numeric variables (e.g., simple size, number of cues in a behavioural assay).

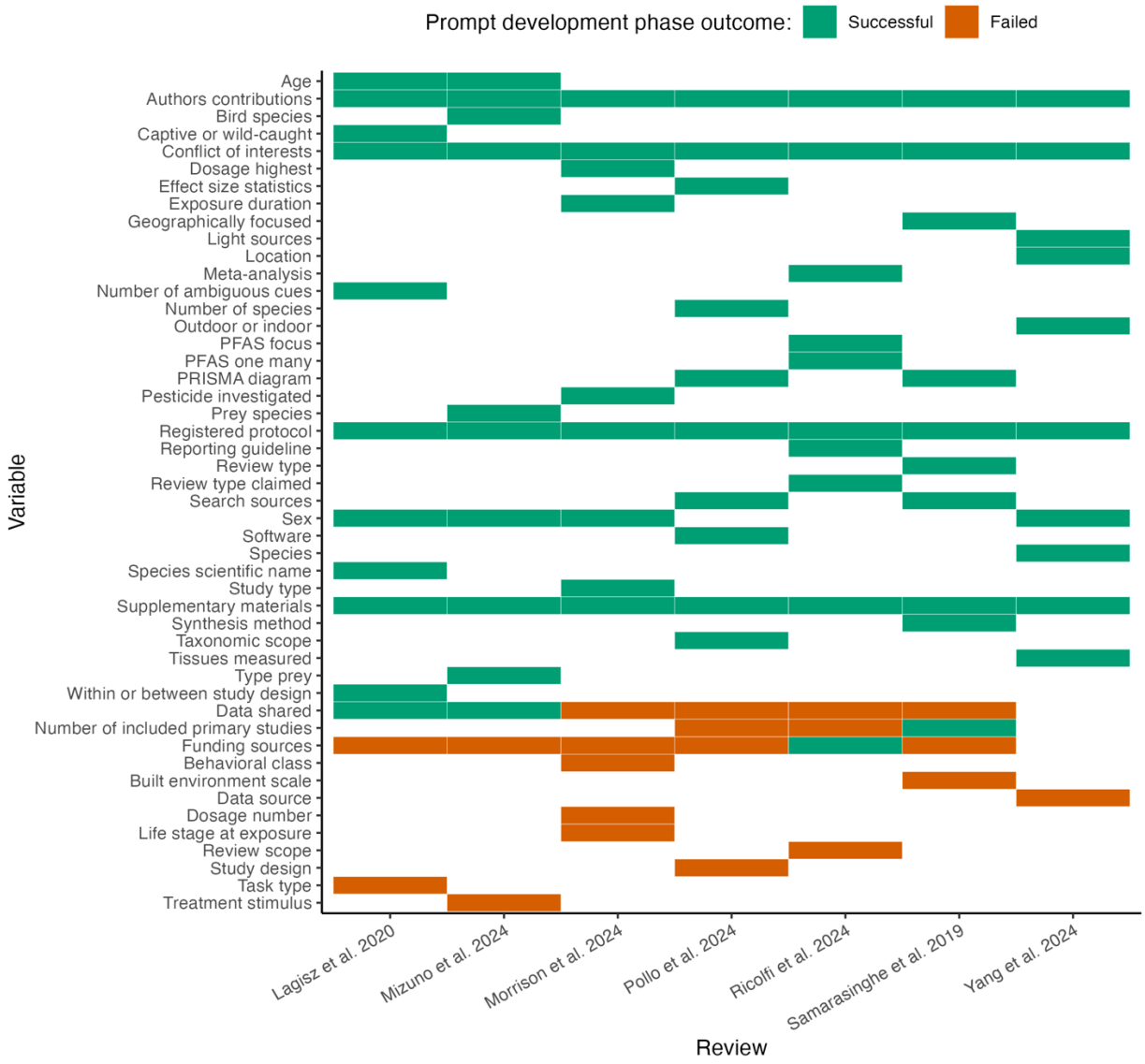

Figure 3.

Distribution and success of 90 considered extraction variables (y-axis) across 7 systematic reviews (x-axis) during the prompt development stage. Colour of cells in the grid indicates whether a variable failed (orange) or succeeded (green) during the prompt development phase of the project. A variable was considered as "Successful" if at least 7 out of 8 answer values were extracted correctly in *Elicit*, i.e. matched human-extracted "gold standard" values, within a maximum of 5 iterations of data extraction prompt refinement. White cells indicate that a given variable was not used in each systematic review. Where variable names and initial prompts were identical for different reviews, they are shown on the same line.

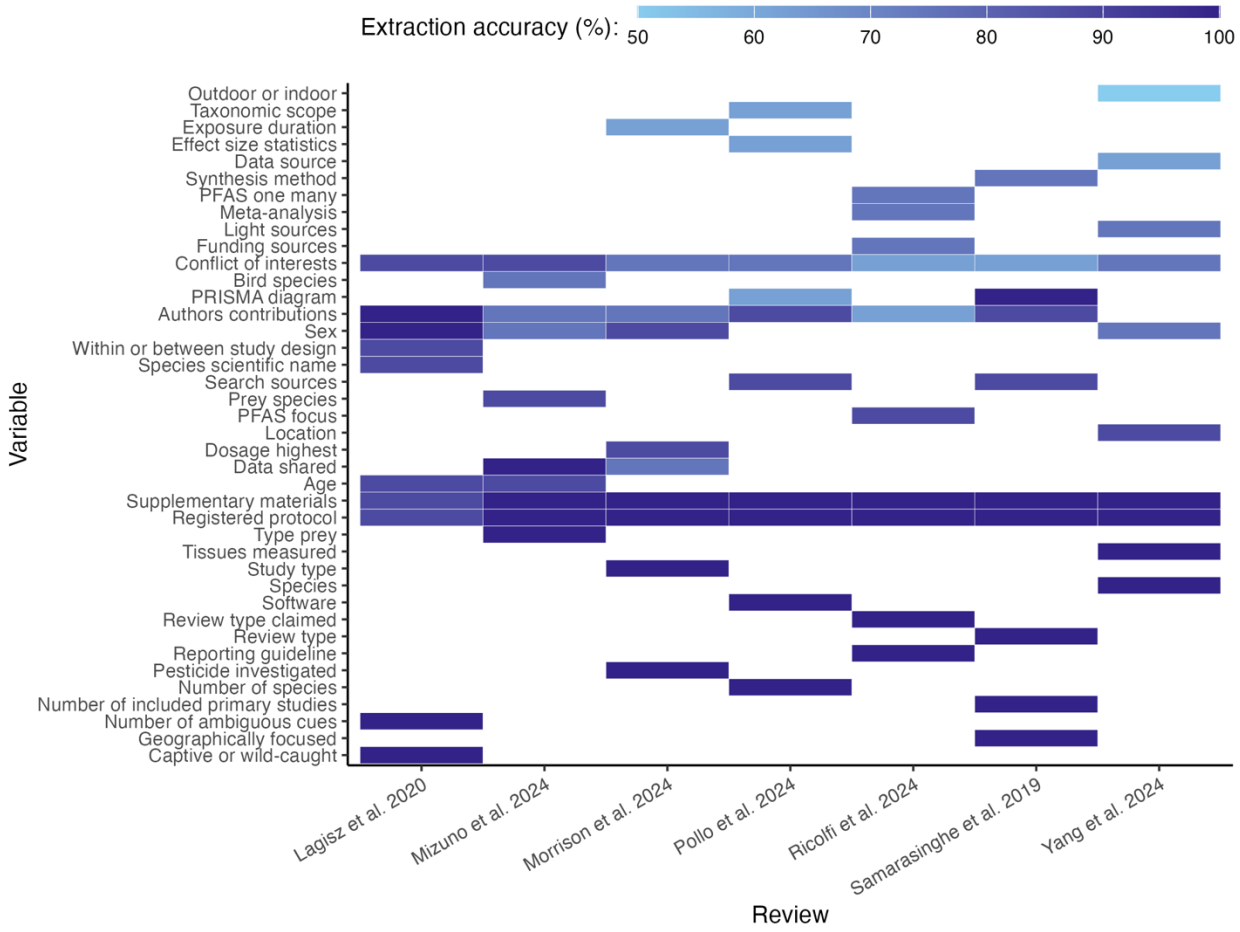

Figure 4.

Accuracy of data extractions performed using *Elicit* platform when compared to human-extracted gold standard answers. We tested extraction variables that passed the prompt development stage with at least 87% accuracy (7 / 8 answers correct) rating (y-axis) for each of the seven systematic reviews (x-axis) using a new test set of 8 studies distinct from the prompt development stage per review. Colour of cells in the grid indicates accuracy (proportion of correct answers) of *Elicit* data extractions during the testing stage of the project. White cells indicate that a given variable was not tested for a given review.

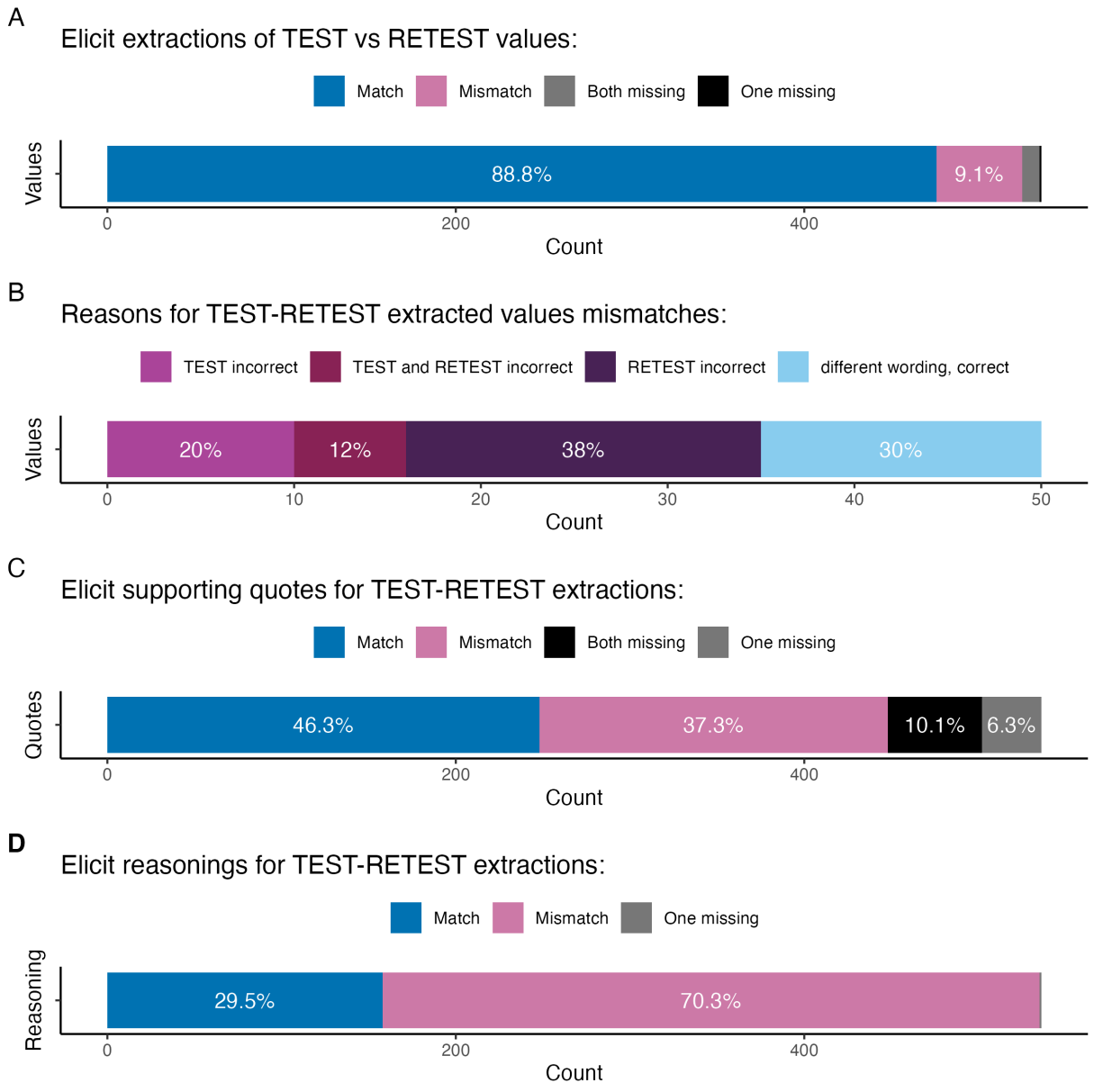

Figure 5.

Comparison of results from using two different accounts in *Elicit* to extract 10 test variables for each of the 7 original systematic reviews (TEST-RETEST phases). Plots represent exact matching of 536 extracted values (A), classification of the reasons of mismatched values (B), comparisons of corresponding supporting quotes (C) and reasoning (D) provided by *Elicit*.

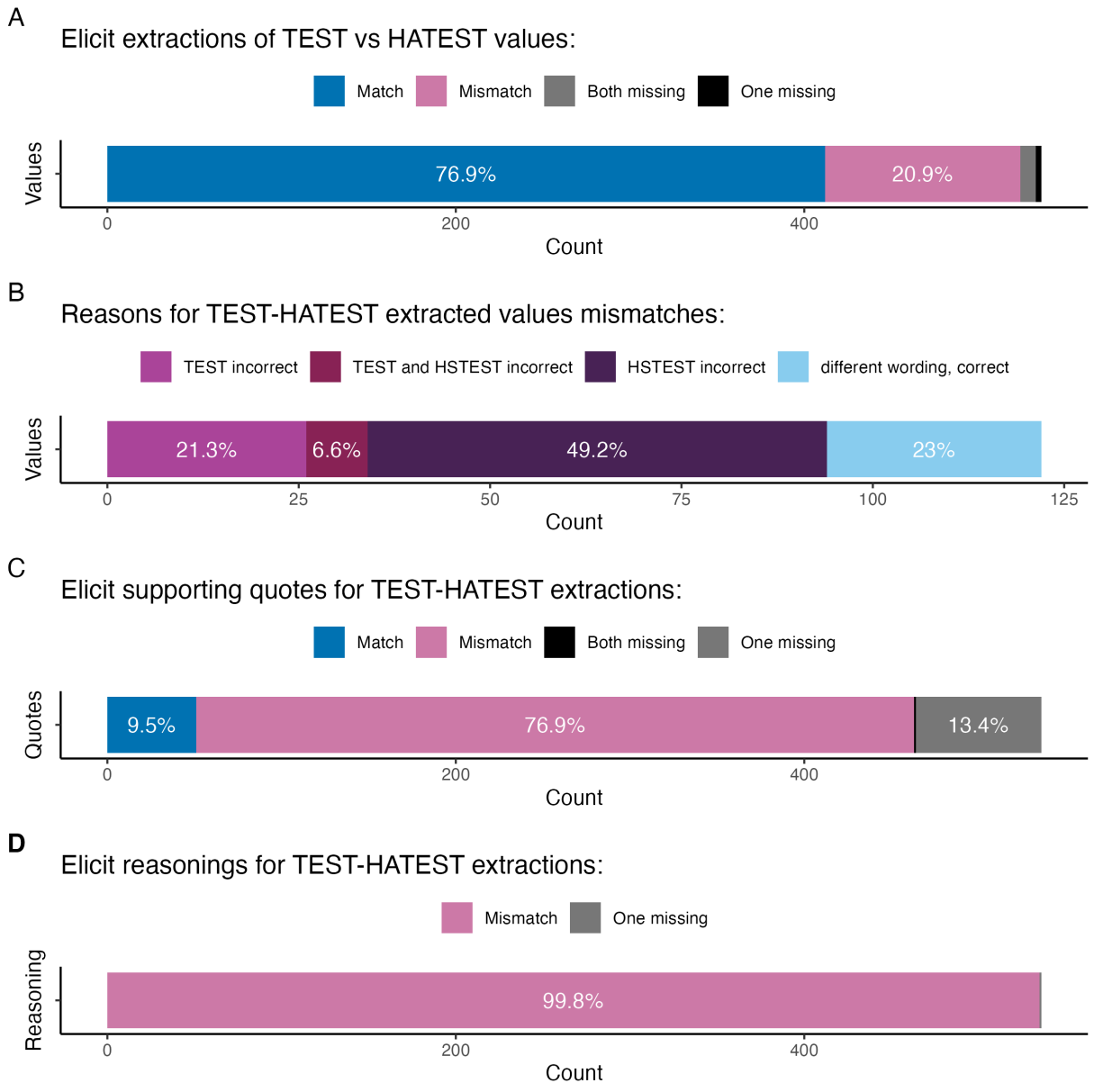

Figure 6.

Comparison of results from re-running test extractions (10 test variables for each of the 7 original systematic reviews) in *Elicit* after the platform enabled a free high accuracy mode for all accounts and plans (TEST-HATEST phases). Plots represent exact matching of 536 extracted values (A), classification of the reasons of mismatched values (B), comparisons of corresponding supporting quotes (C) and reasoning (D) provided by *Elicit*.

Using Elicit AI research assistant for data extraction in systematic reviews: a feasibility study across environmental and life sciences

SUPPLEMENTARY FILE 1

Authors:

Malgorzata Lagisz[1,2], Ayumi Mizuno[2#], Kyle Morrison[1#], Pietro Pollo[1,3#], Lorenzo Ricolfi[1#], Yefeng Yang[1#], Shinichi Nakagawa[1,2#]

* Correspondence: M. Lagisz; e-mail: losialagisz@gmail.com

[#] These authors contributed equally and are listed in alphabetical order

Affiliations:

[1] Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of New South Wales, Kensington, NSW, 2052, Australia

[2] Department of Biological Sciences, University of Alberta, CW 405, Biological Sciences Building, Edmonton, AB T6G 2E9, Canada

[3]School of Environmental and Life Sciences, University of Newcastle, Newcastle, Australia

Short title: Evaluation of data extraction in Elicit AI

Data and code availability: project GitHub repository with all data and code can be found at https://github.com/mlagisz/elicit_extractions_testing (it will be archived to Zenodo when manuscript is accepted for publication).

SUPPLEMENTARY TABLES

Table S1

Completed systematic reviews used as a source of data for testing data extractions in *Elicit*. "Review code" is the abbreviated reference used to denote each of the seven systematic reviews used as a source of data in this project. "Reference" is the full bibliographic reference to a published systematic review article. "Type" is the type of systematic review: SM = Systematic Map, MA = Meta-Analysis. "Discipline" is a broad discipline represented by each of the seven systematic reviews: enviro = environmental sciences, biomed = biomedical sciences, ecoevo = ecology and evolutionary biology. "N" is the number of studies originally included in a given systematic review.

Review code	Reference	Type	Discipline	N
Samarasinghe_2019	Samarasinghe, G.#, Lagisz, M.#, Santamouris, M., Yenneti , K., Upadhyay. A.K., De La Peña Suarez, F., Taunk, B., Nakagawa, S. (2019) A visualized overview of systematic reviews and meta-analyses on low-carbon built environments: An evidence review map. Solar Energy 186: 291-299. https://doi.org/10.1016/j.solener.2019.04.062 ; www.researchweaving.com	SM	enviro	131
Lagisz_2010	Lagisz, M., Zidar, J., Nakagawa, S., Neville, V., Sorato, E., Paul, E.S., Bateson, M, Mendl, M., Løvlie, H. (2020) Optimism, pessimism and judgement bias in animals: a systematic review and meta-analysis. Neuroscience & Biobehavioral Reviews 118: 3-17. https://doi.org/10.1016/j.neubiorev.2020.07.012	MA	biomed	71
Yang_2024	Yang, Y., Liu, Q., Chen, J., Pan, C., Xu, B., Liu, K., Pan, J., Lagisz, M., Nakagawa, S. (2024) Species sensitivities to artificial light at night: A phylogenetically controlled multilevel meta-analysis on melatonin suppression. Ecology Letters 27: e14387. https://doi.org/10.1111/ele.14387	MA	ecoevo	38
Pollo_2024	Pollo, P., Lagisz, M., Yang, Y., Culina, A., Nakagawa, S. (2024) Synthesis of sexual selection: a systematic map of meta-analyses with bibliometric analysis. Biological Reviews, 10.1111/brv.13117	SM	ecoevo	152
Ricolfi_2024	Ricolfi, L, Vendl, C., Bräunig, J., Taylor, M., D, Hesselson, D., Neely, G., G., Lagisz, M. & Nakagawa, S. (2024) A research synthesis of humans, animals, and environmental compartments exposed to PFAS: A systematic evidence map and bibliometric analysis of secondary literature. Environment	SM	enviro	175

	International. 190: 108860, 10.1016/j.envint.2024.108860			
Morrison_2024	Morrison, K., Yang, Y., Santana, M., Lagisz, M.#, & Nakagawa, S.# (2024). A systematic evidence map and bibliometric analysis of the behavioural impacts of pesticide exposure on zebrafish. Environmental Pollution 347:123630. https://doi.org/10.1016/j.envpol.2024.123630	SM	enviro	83
Mizuno_2024	Mizuno A, Lagisz M, Pollo P, Yang Y, Soma M, Nakagawa S. (2024) A systematic review and meta-analysis of anti-predator mechanisms of eyespots: conspicuous pattern vs eye mimicry. eLife: 10.7554/eLife.96338	MA	ecoevo	33

Table S2

Summary of variables used during development (DEV) and testing (TEST, RETEST, HATEST) phases of the for seven systematic reviews. Review code indicates the systematic review (as in Table S1). Variable name matches "Column name" in Elicit. "[ ]" denotes column name in the original raw data file. Variables with * were not explicitly coded in the original systematic review and were coded manually in parallel (before and independently) to their assessment using Elicit. "Description" matches the "Instructions" field in Elicit and "Coding options" matches the "Answer structure" field in Elicit. Initial variable descriptions (Development phase) and coding options are based on the meta-data and other descriptions published in systematic reviews and used in the human-extracted data.

Review code / Variable name	**Development phase**	**Testing phase**
Samarasinghe_2019 / Built environment scale [Built environment scale]	Description: "Type of built environments reviewed in the review article" Coding options: "Predefined list (select one answer): Global / Country / Region, Urban area / Urban system, Building system, Material / Device, Community / Population group" Number of iterations: 5 Initial accuracy: 3/8 Final accuracy: 3/8 Success: no Comment: Elicit is electing 1+ answers - ignoring instruction to select only one. Coded as match with GS if at least one matched	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Samarasinghe_2019 / Geographically focused [Geographically focused]	Description: "Whether it is focused on a particular geographic region" Coding options (select one answer): Yes, No Number of Iterations: 3 Initial accuracy: 4 /8 Final accuracy: 8/8 Success: yes Comment: Elicit uncovered 2 potential errors in human-extracted data (adjusted when scoring)	Description: "Whether it is focused on a particular geographic region.This applies to the aims of the review and its inclusion criteria, not to the coverage of actually included and reviewed studies. If aims are to collect evidence globally or do not specify that review is focused on particular geographic regions (country, continent, climatic zone), the answer should be coded as "no". If the aims and/or If inclusion criteria state that the only studies from particular geographic area are of interest or eligible to be included, the answers should be coded as Yes." Coding options: Yes, No Accuracy: 8/8 Comment: NA

Samarasinghe_2019 / Number of included primary studies [No. of original sources]	Description: "Number of original studies reviewed. Code as a total number of studies included in the systematic review or meta-analysis" Coding options: Number, Not mentioned Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: "Number of original studies reviewed. Code as a total number of studies included in the systematic review or meta-analysis" Coding options: Number Accuracy: 8/8 Comment: NA

Samarasinghe_2019 / Synthesis method [Synthesis method]	Description: “How the evidence was combined” Coding options (select one answer): Qualitative, quantitative, map, qualitative + quantitative Number of Iterations: 4 Initial accuracy: 4/8 Final accuracy: 8/8	Description: “How the evidence was combined. By “evidence” we mean the result of individual primary studies. The approach to summarise evidence can be quantitative, qualitative or combination of quantitative and qualitative. Maps visually summarise information about studies but not their results. Select "qualitative" answer if no statistical synthesis of results of individual studies has been conducted in the review and only a narrative summary of the results of included studies has been provided. Providing some numbers from primary studies or counting them does not qualify as quantitative analysis. Select "quantitative" answer if statistical synthesis of results of individual studies has been conducted in the review (usually a meta-analysis or eta-regression) and no extensive narrative summary has been provided of the results of included studies. Select "map" answer if the scopes of included studies were mainly summarised as in graphs and/or tables, without statistical synthesis of the results of individual

	Success: yes Comment: Elicit selects more than one answer gets selected, so "qualitative, quantitative" was be considered as equivalent to "qualitative + quantitative"	studies (usually a review will state it is a map). Select "qualitative + quantitative" answer if both statistical synthesis of results of individual studies has been conducted in the review and a narrative summary of the results of included studies has been." Coding options: qualitative, quantitative, map, qualitative + quantitative Accuracy: 6/8 Comment: NA

Samarasinghe_2019 / Search sources [Search sources]	Description: “Online databases and other source used in searches” Coding options: Names, No search sources found Number of Iterations: 2 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: Elicit finds more sources, includes minor ones and grey literature, propagates author mistakes (PsycINFO not PsychINFO), but cannot distinguish alternative names (WoS/WoK). Inconsistent answer formatting.	Description: “Online databases and other sources used in searches. Code as a list of names of sources separated by commas if sources are listed in the review. Ignore sources of grey literature. Code Web of Knowledge as Web of Science (correct alternative name). Code as "No search sources found", without additional comments, if the names of search sources cannot be found in the review.” Coding options: Any answer (free text) Accuracy: 7/8 Comments: coded one mismatch where Elici did not find any search sources (in supplements)

Samarasinghe_2019 / PRISMA diagram [Prisma diagram used]	Description: "Whether PRISMA diagram is presented" Coding options (select one answer): Yes, No Number of Iterations: Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: Elicit missed one Yes due to the data being presented in a supplementary file. Extra iteration added to clarify instructions for human coders.	Description: "Whether a PRISMA-like diagram is presented. PRISMA-like diagrams (PRISMA diagrams) illustrate the process of searching and screening of literature (usually primary studies, articles, etc.). PRISMA-like diagrams are also called flow diagrams of the searching screening process." Coding options: Yes, No. Accuracy: 8/8 Comment: NA
Samarasinghe_2019 / Conflict of interests [Conflict of interest]	Description: "Was the conflict of interests disclosed by the authors" Coding options (select one answer): Yes, Not declared specifically Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: "Was the conflict of interests disclosed by the authors." Coding options: Yes, Not declared specifically Accuracy: 5/8 Comment: Elicit coded "The authors have declared that no competing interests exist" as No.

Samarasinghe_2019 / Funding sources [Funding sources]	Description: "Organisations and bodies that funded the study" Coding options (select one answer): Yes, No funding sources recorded Number of Iterations: 5 Initial accuracy: 1/8 Final accuracy: 3/8 Success: no Comment: Hunter_2015.pdf does not include Acknowledgements sections which is available online	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase
Samarasinghe_2019 / Supplementary materials*	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options (select one answer): Yes, No Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 8/8 Success: yes Comment: extra iteration added to to	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links. Supplementary materials are sometimes called Supporting Material or Appendix or Additional File." Coding options: Yes, No Accuracy: 8/8 Comment: NA

	include term “Appendix” in the prompt	

Samarasinghe_2019 / Data shared*	Description: “Whether the raw data used for analyses presented in the article has been shared either in supplementary materials, or as a link to an external file repository. If data is available on request, it will be coded as No” Coding options (select one answer): Yes, No Number of Iterations: 5 Initial accuracy: 5/8 Final accuracy: 5/8 Success: no Comment: one study only shares partial data (assessments) and another has all data in the main text as a table. It is hard to distinguish raw and partial data unless there is an explicit statement that all data is shared at a specified location.	Description: NA Coding options: NA Accuracy: NA Comments: variable not used in the testing phase

Samarasinghe_2019 / Review type	Description: "Claimed review type (systematic review or meta-analysis)" Coding options(select one answer): systematic review, meta-analysis Number of Iterations: 3 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: extra to iterations added trying to get Elicit to select only one answer (unsuccessfully).	Description: "Claimed review type. If review claims to contain a meta-analysis or meta-regression, select "meta-analysis". Select only one answer. For example, if a review appears to be both a systematic review and a meta-analysis, only select "meta-analysis"." Coding options: systematic review, meta-analysis Accuracy: 8/8 Comment: NA
Samarasinghe_2019 / Registered protocol	Description: "Whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan." Coding options(select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: all No	Description: "Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan." Coding options: Yes, No Accuracy: 8/8 Comment: all No

Samarasinghe_2019 / Authors contributions	Description: "Whether the study contains a description of the roles and/or contributions of the study authors" Coding options(select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: "Whether the study contains a description of the roles and/or contributions of the study authors" Coding options: Yes, No Accuracy: 7/8 Comment: Fenwick_2013 has in-text contributions to screening tasks only - overinterpreted by Elicit as Yes
Lagisz_2020 / Species scientific name [ScientificName]	Description: "Scientific name of an animal species used in the experiment" Coding options: Name Number of Iterations: 3 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: synonymous names should be acceptable as a match. Two extra iterations added to reduce the answer in Elicit to a name only.	Description: "Scientific name of an animal species used in the experiment. Do not add any comments or explanations in the answer, only return the name of the species" Coding options: Name Accuracy: 7/8 Comment: Lacune not considered as a synonym to sheep, esp. given that word "sheep" is used in the article text (Lacune sheep).

Lagisz_2020 / Captive or wild-caught [Captive_Wild-caught]	Description: "Source of animals used in the experiment, as reported in the paper: captive = all used animals were captive, or source not reported; wild-caught = all used animals were wild-caught" Coding options (select one answer): captive, wild-caught Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: 1 Comment: all captive	"Source of animals used in the experiment, as reported in the paper: captive = all used animals were captive, or source not reported; wild-caught = all used animals were wild-caught" Coding options: captive, wild-caught Accuracy: 8/8 Comment: all captive

Lagisz_2020 / Age [Age]	Description: “Age of animals used in the experiment: juvenile = all used animals were not sexually mature; adult = all used animals were sexually mature, mixed age, or age not reported” Coding options (select one answer): juvenile, adult Number of Iterations: 2 Initial accuracy: 5/8 Final accuracy: 8/8 Success: yes Comment: two studies were on young adults that should be classified as adults	Description: "Age of animals used in the experiment: juvenile = all used animals were not sexually mature or not close to becoming sexually mature; adult = all used animals were sexually mature or close to sexual maturity (young adults), mixed age, or age not reported. Select only one answer." Coding options: juvenile, adult Accuracy: 7/8 Comment: NA

Lagisz_2020 / Within or between study design [WithinBetween]	Description: "Whether between-individual or within-individual study design was used: between = two or more groups of animals were simultaneously subject to different treatments (or treatment vs. control/benign); within = same group of animals was subject sequentially to different treatments (or treatment vs. control/benign), includes cross-over design" Coding options (select one answer): between, within Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: "Whether between-individual or within-individual study design was used: between = two or more groups of animals were simultaneously subject to different treatments (or treatment vs. control/benign): within = same group of animals was subject sequentially to different treatments (or treatment vs. control/benign), includes cross-over design" Coding options: between, within Accuracy: 7/8 Comment: NA

Task type [TaskType]	Description: “Type of the task used during behavioural trials: active choice = go/go tasks in which an animal is required to make an active response to cues perceived as positive and to cues perceived as negative; go/no-go = tasks in which an animal is required to suppress a response to cues perceived as negative and actively respond only to cues perceived as positive” Coding options (select one answer): active choice, go/no-go Number of Iterations: 5 Initial accuracy: 6/8 Final accuracy: 5/8 Success: no Comment: often tricky to distinguish even for researchers	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Lagisz_2020 / Sex [Sex]	Description: “Sex of tested animals in the compared groups: female = only female animals were used; male = only male animals were used; both = both female and male animals were used” Coding options (select one answer): female, male, both Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: “Sex of tested animals in the compared groups: female = only female animals were used; male = only male animals were used; both = both female and male animals were used” Coding options: female, male, both Accuracy: 8/8 Comment: NA
Lagisz_2020 / Conflict of interests*	Description: “Was the conflict of interests disclosed by the authors” Coding options (select one answer): Yes, Not declared specifically Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “Was the conflict of interests disclosed by the authors” Coding options: Yes, Not declared specifically Accuracy: 7/8 Comment: NA

Lagisz_2020 / Funding Source*	Description: "Whether the organisations and bodies that funded the study are mentioned" Coding options (select one answer): Yes, No funding sources recorded Number of Iterations: 5 Initial accuracy: 1/8 Final accuracy: 1/8 Success: no Comment: Elicit unable to detect funding statements.	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase
Lagisz_2020 / Supplementary materials*	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options (select one answer): Yes, No Number of Iterations: 2 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links. Supplementary materials are sometimes cal;led Supporting Material or Appendix or Additional File." Coding options: Yes, No Accuracy: 7/8 Comment: No reasoning for one "-" which should be No.

	Comment: refined prompt to include term “Appendix”	
Lagisz_2020 / Data shared*	Description: “Whether the raw data used for analyses presented in the article has been shared either in supplementary materials, or as a link to an external file repository. If data is available on request, it will be coded as No” Coding options (select one answer): Yes, No Number of Iterations: Initial accuracy: 8/8 Final accuracy: 7/8	Description: NA Coding options: NA Accuracy: NA Comment: decided not to use and added other standardised variables instead

	Success: yes Comment: NA	
Lagisz_2020 / Registered protocol*	Description: "Whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan." Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: all No	Description: "Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan. Coding options: Yes, No Accuracy: 7/8 Comment: all No

Lagisz_2020 / Authors contributions*	Description: "Whether the study contains a description of the roles and/or contributions of the study authors" Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: "Code whether the study contains a description of the roles and/or contributions of the study authors" Coding options: Yes, No Accuracy: 8/8 Comment: NA
Lagisz_2020 / Number of ambiguous cues	Description: "Number of different ambiguous cues used in the judgement bias trials" Coding options: Number Number of Iterations: 2 Initial accuracy: 6/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: "Number of different ambiguous cues used in judgement bias trials. Do not report how many times animals were tested during judgement bias trials. Report only as an integer number." Coding options: Yes, No Accuracy: 8/8 Comment: NA

Yang_2024 / Species [Species]	Description: “The Latin binomial name (e.g., scientific name) of the wild animal studied in the paper” Coding options: Name Number of Iterations: 2 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: Accepting name synonyms as correct answers. An additional iteration needed to refine answers (removing extra text)	Description: “The Latin binomial names (e.g., scientific name) of the wild animal studied in the paper. Provide scientific names of the species only, without context. or comments. If more than one species was used in the project, list all species names separated by a comma.” Coding options: NA Accuracy: 8/8 Comment: Elicit helped to identify a typo in a human-extracted species name

Yang_2024 / Outdoor or indoor [Outdoor_or_Indoor]	Description: “Was the paper an Outdoor study or an Indoor study. For an observational study (e.g., free-living, outdoor enclosures/ cages), it is coded as an Outdoor study. If the animal was collected from the wild, but the experiments was conducted in the lab, it is an Indoor study)” Coding options (select one answer): Outdoor, Indoor, Unclear/Other Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: Renthlei_2019 has 2 experiments in different conditions, only one was used in Yang_2024	Description: “Was the paper an Outdoor study or an Indoor study. For an observational study (e.g., free-living, outdoor enclosures/ cages), it is coded as an Outdoor study. If the animal was collected from the wild, but the experiments was conducted in the lab, it is an Indoor study)” Coding options: Outdoor, Indoor, Unclear/Other Accuracy: 4/8 Comment: Two values reported when part of the experiment conducted outdoors.

Yang_2024 / Location [Location]	Description: "Country where the experiment was conducted for a field study or animals" Coding options: Name Number of Iterations: 2 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: one location missing in gold standard answers. An additional iteration needed to refine answers (removing extra text)	Description: "Country where the experiment was conducted for a field study or animals were collected for a lab study, if reported. If not reported, the country of the first author was assumed to be the study location. Provide country names only, without context or comments" Coding options: Name Accuracy: 7/8 Comment: Elicit coded South Australia as a country
Yang_2024 / Light sources [Light_sources]	Description: "Types of lamps used in the paper" Coding options (select one answer): LED, Fluorescent lamp, Incandescent lamp, Halogen lamp, Streetlight, Unclear/Other Number of Iterations: 2 Initial accuracy: 4/8 Final accuracy: 7/8 Success: yes Comment: two gold standard values missing - not reported in the papers or unclear	Description: "Types of lamps used in the experiment when exposing animals to artificial light at night" Coding options: LED, Fluorescent lamp, Incandescent lamp, Halogen lamp, Streetlight, Unclear/Other Accuracy: 6/8 Comment: after cross-checking, interpreted empty gold standard values as unclear

Yang_2024 / Sex [Sex]	Description: “The sex of the animals” Coding options (select one answer): male, female, mixed/unclear Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 8/8 Success: yes Comment: refined prompt to exclude sex of the authors	Description: “The sex of the animals. Ignore sex of the authors” Coding options: male, female, mixed/unclear Accuracy: 6/8 Comment: interpreted empty gold standard values as unclear
Yang_2024 / Data source [Data_source]	Description: “Where the extracted melatonin data was reported in the paper. Five options: text (e.g., text_p4), figure (e.g., fig2), table (e.g., table3), supplementary material (e.g., supplement_fig2)” Coding options: Free text Number of Iterations: 5 Initial accuracy: 3/8 Final accuracy: 6/8 Success: no Comment: scored leniently as Yes if answer included gold standard values	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Yang_2024 / Conflict of interests*	Description: "Whether the conflict of interests was disclosed by the authors" Coding options (select one answer): Yes, Not declared specifically Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: reused the final prompt from the other reviews	Description: "Whether the conflict of interests was disclosed by the authors" Coding options: Yes, Not declared specifically Accuracy: 6/7 Comment: "Not declared specifically" should mean that there is on statements, not that there is no CoI
Yang_2024 / Funding Source*	Description: "Whether the organisations and bodies that funded the study are mentioned" Coding options (select one answer): Yes, No funding sources recorded Number of Iterations: 5 Initial accuracy: 5/8 Final accuracy: 5/8 Success: no Comment: reused the final prompt from the other reviews	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Yang_2024 / Supplementary materials*	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: reused the final prompt from the other reviews	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links. Supplementary materials are sometimes cal;led Supporting Material or Appendix or Additional File" Coding options: Yes, No Accuracy: 8/8 Comment: NA

Yang_2024 / Data shared*	Description: "Whether the raw data used for analyses presented in the article has been shared either in supplementary materials, or as a link to an external file repository. If data is available on request, it will be coded as No" Coding options (select one answer): Yes, No Number of Iterations: 5 Initial accuracy: 6/8 Final accuracy: 6/8 Success: no Comment: reused the final prompt from the other reviews	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase
Yang_2024 / Registered protocol*	Description: "Whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan." Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: all No	Description: "Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan." Coding options: Yes, No Accuracy: 8/8 Comment: all No

Yang_2024 / Authors contributions*	Description: "Whether the study contains a description of the roles and/or contributions of the study authors" Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: "Whether the study contains a description of the roles and/or contributions of the study authors" Coding options: Yes, No Accuracy: 7/8 Comment: Fenwick_2013 has in-text contributions to screening tasks only - overinterpreted by Elicit as "Yes"
Yang_2024 / Tissues measured	Description: "What tissue was used for measurement of melatonin levels: Blood, Pineal, SCN (= Suprachiasmatic Nucleus), Urinary, Retina (= ocular), Water (= melatonin in tank water was measured)? Coding options (select one answer): Blood, Pineal, SCN, Urinary, Retina, Water Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: one miscoded value identified by Elicit.	Description: "What tissue was used for measurement of melatonin levels: Blood, Pineal, SCN (= Suprachiasmatic Nucleus), Urinary, Retina (= ocular), Water (= melatonin in tank water was measured)? Coding options: Blood, Pineal, SCN, Urinary, Retina, Water Accuracy: 8/8 Comment: NA

Pollo_2024 / Effect size statistics [effect_size_statistics]	Description: "Type of effect size statistics used in the study" Coding options: Unclear, Mean Standardised Difference, Odds Ratio, Correlations, lnRR, Other Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: extra iteration added trying to refine the prompt to consider log effect sizes	Description: "Type of effect size statistics used in the study. Treat log-transformed effect sizes the same as their non-transform d versions (e.g. log-transformed odds ratio (LOR) is the same as odds ration (OR))" Coding options: unclear, mean standardised difference, Odds Ratio, correlations, lnRR, other Accuracy: 5/8 Comment: Elicit extracted all mentioned, not just the ones used in a meta-analysis
Pollo_2024 / Number of included primary studies [number_studies]	Description: "Number of individual studies used in meta-analysis/ meta-regression" Coding options: Number Number of Iterations: 5 Initial accuracy: 6/8 Final accuracy: 6/8 Success: no Comment: Elicit confusing number of datasets or effect sizes with numbers of studies.	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Pollo_2024 / Taxonomic scope [taxonomic_scope]	Description: "Taxonomic scope of the study" Coding options (select one answer): Single species, Multiple species from a specific taxon, Multiple species from all taxa Number of Iterations: 2 Initial accuracy: 5/8 Final accuracy: 8/8 Success: yes Comment: actual extracted values used do not match their meta-data, but are similar, so can be matched accordingly	Description: "Taxonomic scope of the study. "single species" - focused on a single species. "multiple species from a specific taxon" - focus on a taxonomic group but not all possible organisms. "multiple species from all taxa" - not focused on specific species or taxonomic group." Coding options: single species, multiple species from a specific taxon, multiple species from all taxa Accuracy: 5/8 Comment: NA
Pollo_2024 / Sex [sex_focused]	Description: "Which sex is investigated in the meta-analysis" Coding options (select one answer): Males, Females, Both, Hermaphrodites, Unclear Number of Iterations: NA Initial accuracy: NA Final accuracy: NA Success: NA Comment: variable described in the meta-data but	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

	absent from the gold standard dataset - omitted	
Pollo_2024 / Study design [study_design]	Description: “Type of study design used in the meta-analysis” Coding options (select one answer): Experimental, Observational, Mixed, Unclear Number of Iterations: 5 Initial accuracy: 5/8 Final accuracy: 4/8 Success: no Comment: NA	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Pollo_2024 / PRISMA diagram [prisma]	Description: "PRISMA-like flowchart is provided" Coding options (select one answer): Adequate, Insufficient, Not applicable Number of Iterations: 2 Initial accuracy: 3/8 Final accuracy: 7/8 Success: yes Comment: coding options according to PRISMA-EcoEvo	Description: "Whether PRISMA-like flowchart is provided. Select "not applicable" answer if searches and screening of literature were not conducted as a source of the data used in the paper" Coding options: adequate, insufficient, not applicable Accuracy: 5/8 Comment: coding options according to PRISMA-EcoEvo
Pollo_2024 / Conflict of interests [Conflict of interests]	Description: "Was the conflict of interests disclosed by the authors" Coding options (select one answer): Yes, Not declared specifically Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: reused the final prompt from the other reviews	Description: "Was the conflict of interests disclosed by the authors" Coding options (select one answer): Yes, Not declared specifically Accuracy: 6/8 Comment: NA

Pollo_2024 / Search sources [search_sources]	Description: "Sources to conduct literature searches are listed" Coding options: Unclear, Web of Science, Scopus, Other databases that cover only published studies, Other databases that also cover grey literature, Google Scholar, Backward citations of key papers, Forward citations of key papers, Backward citations of initially selected papers, Forward citations of initially selected papers, Other Number of Iterations: 3 Initial accuracy: 1/8 Final accuracy: 7/8 Success: yes Comment: changed answer format from categorical to free text	Description: "Online databases and other source used in searches. Code as a list of names of sources separated by commas if sources are listed in the review. Ignore sources of grey literature. Code Web of Knowledge as Web of Science (correct alternative name). Code as "No search sources found", without additional comments, if the names of search sources cannot be found in the review. Do not add text on context or comments." Coding options: free text (Any answer) Accuracy: 7/8 Comment: some cleaning needed; matched "Not mentioned" as equivalent to "Unclear"

Pollo_2024 / Supplementary materials*	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options (select one answer): Yes, No Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: extra iteration added trying to further refine the prompt	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links. Supplementary materials are sometimes called Supporting Material or Appendix or Additional File." Coding options: Yes, No Accuracy: 8/8 Comment: NA

Pollo_2024 / Data shared [main_data]	Description: “Data necessary to reproduce results are provided” Coding options (select one answer): Adequate, Substandard, Insufficient, Not applicable Number of Iterations: 5 Initial accuracy: 5/8 Final accuracy: 5/8 Success: no Comment: one study only shares partial data (assessments) and another has all data in the main text as a table	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase
Pollo_2024 / Funding sources*	Description: “Organisations and bodies that funded the study are listed.” Coding options (select one answer): Yes, No funding sources recorded Number of Iterations: 5 Initial accuracy: 1/8 Final accuracy: 5/8 Success: no Comment: not extractable	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Pollo_2024 / Registered protocol*	Description: “Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan” Coding options (select one answer): Yes, No Number of Iterations: Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: all No	Description: “Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan.” Coding options: Yes, No Accuracy: 8/8 Comment: all No
Pollo_2024 / Authors contributions*	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors.” Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors” Coding options: Yes, No Accuracy: 7/8 Comment: NA

Pollo_2024 / Number of species [Number of species]	Description: "Number of species used in meta-analysis/ meta-regression" Coding options: Number Number of Iterations: 3 Initial accuracy: 5/8 Final accuracy: 7/8 Success: yes Comment: answers needed cleaning to be number-only	Description: "Number of species used in meta-analysis/ meta-regression. If the article is focused on a single species, answer is "1". Answer by providing only a single number, without additional text, comments or context (i.e. do not answer with full sentences)" Coding options: Number Accuracy: 8/8 Comment: Elicit identified one wrong number
Pollo_2024 / Software [software]	Description: "Which software is used for analysis (NA if unclear)" Coding options: free text (Name) Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: extra iteration to refine the prompt for answer cleaning	Description: "Number of species used in meta-analysis/ meta-regression. If the article is focused on a single species, answer is "1". Answer by providing only a single number, without additional text, comments or context (i.e. do not answer with full sentences)" Coding options: free text (Any answer) Accuracy: 8/8 Comment: considered as match if answer included gold standard values (partial match)

Ricolfi_2024 / Review type claimed [Review_type_claimed]	Description: "Which type of review do the authors claim their review to be? (record as stated in the included review)" Coding options (select one answer): Critical review, Meta-analysis, Systematic review, Comprehensive review, Review, Scoping review, Systematic evidence map, Systematic review and meta-analysis Number of Iterations: 5 Initial accuracy: 3/8 Final accuracy: 8/8 Success: yes Comment: changed from categorical answer to free text	Description: "Which type of review or meta-analysis do the authors claim their review to be? Record only the key term used in the article, without context or comments (i.e. do not use full sentences in the answer)." Coding options: free text (Any answer) Accuracy: 8/8 Comment: NA
Ricolfi_2024 / Meta-analysis [Meta-analysis]	Description: "Does the review include a meta-analysis (quantitative synthesis of information from multiple sources)?" Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: "Does the review include a meta-analysis (quantitative synthesis of information from multiple sources)?" Coding options: Yes, No Accuracy: 6/8 Comment: NA

Ricolfi_2024 / PFAS focus [PFAS_focus]	Description: “Is the review focused on PFAS? (or does it also investigate other chemicals)” Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: “Is the review focused on PFAS? (or does it also investigate other chemicals)” Coding options: Yes, No Accuracy: 7/8 Comment: NA
Ricolfi_2024 / Review scope [Human_animal_environment]	Description: “Does the study deal with a PFAS-topic related to humans, animals or the environment?” Coding options (select one answer): Human, Animal, Environment, Mixed Number of Iterations: 5 Initial accuracy: 2/8 Final accuracy: 4/8 Success: no Comment: accepted Mixed alongside other answers	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Ricolfi_2024 / Number of included primary studies [N_studies]	Description: “How many studies are listed as included in the review?” Coding options: Number Number of Iterations: 5 Initial accuracy: 6/8 Final accuracy: 6/8 Success: no Comment: NA is for missing information. Some numbers are inferred from figures or supplementary files	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase
Ricolfi_2024 / Reporting guideline [Reporting_guideline]	Description: “Whether review claims to follow a specific reporting or conduct guideline/checklist? ” Coding options: Name Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: answer options (in the protocol not matching extracted data; extra iteration to get rid of critical appraisal checklists	Description: “Whether review claims to follow a specific reporting or conduct guideline/checklist? Ignore checklists used for critical appraisal of relevant studies” Coding options: Yes, No Accuracy: 8/8 Comment: NA

Ricolfi_2024 / Conflict of interests [COI_statement]	Description: “Do the authors provide ‘conflict of interest’ or an equivalent statement?” Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “Was the conflict of interests disclosed by the authors” Coding options: Yes, No Accuracy: 5/8 Comment: Elicit coded No when authors have declared no conflict of interests (Yes)

Ricolfi_2024 / Funding source [Funding_statement]	Description: “Do the authors provide a statement about the funding of the study?”	Description: "Whether there is any indication that the research was funded or supported by an external organization, institution, grant, or sponsor. Code as a "yes" answer any explicit or implicit mention of external funding, support, or acknowledgment of resources or assistance that enabled the research to be conducted, for example acknowledgments of: financial support (e.g., grants, fellowships, awards, institutional funds), Research funding agencies (e.g., national science foundations, health research councils, private foundations), Government funding (e.g., ministries, departments, public research funding programs), Industry or corporate sponsorship, Non-profit organization support. Code as a "yes" answer any common phrases and variations that may indicate funding or support, for example: "This work was supported by...", "Financial support was provided through...", "This project was funded by...", "Research related to this article was funded by...", "This research is part of a project supported by...",

	Coding options (select one answer): Yes, No Number of Iterations: 5 Initial accuracy: 4/8 Final accuracy: 7/8 Success: yes Comment: added detailed explanations and many examples to make this work better	"The authors received funding from...", "Supported by the [Name of Organization/Agency]", "Funding was received from...", "This work was carried out under the funding provided by...", "[Author] was supported by...", "The study was made possible by funding from...". Code as a "yes" answer any mentions of grant numbers, project IDs, or funding references. Code as a "yes" answer any acknowledgments of research programs or consortia noted to be externally funded. Code as a "yes" answer any mentions that researchers were affiliated with funded programs or centers during the work. Funding acknowledgments may appear in dedicated sections (e.g., Acknowledgments or Funding), in author affiliations or footnotes, or even in the main text or footnotes of the article." Coding options: Yes, No Accuracy: 6/8 Comment: one human mistake revealed by Elicit

Ricolfi_2024 / Supplementary materials*	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options: Yes, No Number of Iterations: 2 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: refined the prompt to match the final one from other reviews	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links. Supplementary materials are sometimes called Supporting Material or Appendix or Additional File." Coding options: Yes, No Accuracy: 8/8 Comment: NA
Ricolfi_2024 / Data shared [Raw_data]	Description: "Is the raw data of the study provided?" Coding options (select one answer): Yes, No Number of Iterations: 5 Initial accuracy: 6/8 Final accuracy: 6/8 Success: no Comment: NA	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Ricolfi_2024 / PFAS one many [PFAS one many]	Description: “What types of PFAS does the review focus on? If it is one PFAS type only, answer "One", in the case of several PFAS being mentioned in the included studies, state 'Multiple'” Coding options (select one answer): One, Multiple Number of Iterations: 3 Initial accuracy: 5/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “What types of PFAS does the review focus on in terms of evidence synthesis? If it is focused on one PFAS type, answer "One", even if it briefly mentions other PFAS types or chemicals. If it is focused on more than one types of PFAS substances, answer "Multiple". Ignore other chemical substances that are not classifies as PFAS (e.g. polybrominated diphenyl ethers, PBDEs)” Coding options: One, Multiple Accuracy: 6/8 Comment: NA
Ricolfi_2024 / Registered protocol [Protocol]	Description: “Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan” Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: “Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan” Coding options: Yes, No Accuracy: 8/8 Comment: NA

Ricolfi_2024 / Authors contributions*	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors” Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors” Coding options: Yes, No Accuracy: 5/8 Comment: NA
Morrison_2024 / Study type [study_type]	Description: “Record the study type (if unclear or not reported select Not reported)" Coding options (select one answer): observational, experimental, other, not reported Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: all “experimental”	Description: “Record the study type (if unclear or not reported select "not reported")” Coding options: observational, experimental, other, not reported Accuracy: 8/8 Comment: all “experimental”
Morrison_2024 / Dosage number [dosage_number]	Description: “Record whether the experiment uses one or multiple pesticide doses (if unclear or not reported state Not reported)" Coding options: Number Number of Iterations: 5 Initial accuracy: 3/8 Final accuracy: 4/8	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

	Success: no Comment: NA	
Morrison_2024 / Life stage at exposure [life_stage_exposure]	Description: "Record the sexual maturity of the zebrafish (e.g. juvenile, adult or larvae) that was exposed to a pesticide (if unclear or not reported select "not reported"; if multiple, select multiple options)" Coding options (select one answer): juvenile, adult, larvae, embryo, other, not reported Number of Iterations: 5 Initial accuracy: 5/8 Final accuracy: 5/8 Success: no Comment: NA	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Morrison_2024 / Sex [Sex]	Description: “Record the sex of the zebrafish used (Both option requires males and females to be measured separately whilst Mixed means that males and females are included and measured together)” Coding options (select one answer): male, female, both, mixed, other, not reported Number of Iterations: 2 Initial accuracy: 4/8 Final accuracy: 8/8 Success: yes Comment: merged "both" and "mixed" - considered them to be equivalent to simplify extraction	Description: “Record the sex of the zebrafish used in the pesticide exposure experiment to measure effects on behaviour” Coding options: male, female, both, mixed, other, not reported Accuracy: 7/8 Comment: NA

Morrison_2024 / Behavioral class [behavioral_class]	Description: “State the behavioural class being used to monitor the behavioural change in response to pesticide exposure” Coding options: Activity or movement, Courtship and/or mating behaviour, Post-copulation and/or parental care, Aggression, Sociality, Cognition and/or learning, Boldness or anxiety, Foraging, Antipredator, Lateralization, Other, Not reported Number of Iterations: 5 Initial accuracy: 3/8 Final accuracy: 4/8 Success: no Comment: Had to collapse last few original levels into Other (max. 8 allowed)	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase
Morrison_2024 / Conflict of interests*	Description: “Whether the conflict of interests was disclosed by the authors” Coding options (select one answer): Yes, Not declared specifically Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes	Description: “Was the conflict of interests disclosed by the authors” Coding options: Yes, Not declared specifically Accuracy: 6/8 Comment: one article has COi in supplementary files

	Comment: reused prompt from other reviews	
Morrison_2024 / Funding Source*	Description: "Whether the organisations and bodies that funded the study are mentioned" Coding options (select one answer): Yes, No funding sources recorded Number of Iterations: 5 Initial accuracy: 2/8 Final accuracy: 5/8 Success: no Comment: not extractable	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Morrison_2024 / Supplementary materials*	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options (select one answer): Yes, No Number of Iterations: Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options: Yes, No Accuracy: 8/8 Comment: NA

Morrison_2024 / Data shared*	Description: "Whether the raw data used for analyses presented in the article has been shared either in supplementary materials, or as a link to an external file repository. If data is available on request, it will be coded as No" Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: refined the prompt to match other reviews	Description: "Whether the raw data used for analyses presented in the article has been publicly shared. Select "Yes" answer if there is any mention of the raw collected data being shared, for example, in supplementary materials or appendices, as downloadable tables or files, via links to external repositories or websites, as full datasets embedded in the article (e.g., detailed data tables in the main text). Raw data refers to the original collected data that underlies the analyses - not just summary statistics, graphs, or processed outputs. Select "No" answer if the data is only available upon request, the data is only shared in summarized or aggregated form, there is no clear mention of shared raw data." Coding options: Yes, No Accuracy: 6/8 Comment: Elicit scored yes when author says that data is in supplementary files (but no actual data there), data is available on request, or just the sequences were deposited (not data)

Morrison_2024 / Exposure duration ["duration" and "duration units"]	Description: “Record the duration of the pesticide exposure used in the experiment (if multiple select the longest). Also record the unit used to measure the pesticide exposure time. If unclear or not reported select Not reported” Coding options (select one answer): Number and unit (free text) Number of Iterations: 2 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: extra iteration needed to clean up the answers	Description: “Record the duration of the pesticide exposure used in the behavioural experiment. If multiple pesticide exposure durations were used in the behavioral experiment, select the longest one. Also record the unit used to measure the pesticide exposure time. If unclear or not reported, select "not reported". Answer by providing only a single number and unit, without additional text, comments or context (i.e. do not answer with full sentences)” Coding options: free text (Any answer) Accuracy: 5/8 Comment: Two extraction mistakes found by Elicit

Morrison_2024 / Pesticide investigated [pesticide_investigated]	Description: “Record the type of pesticide tested within the study (if multiple, separate with semicolon, if cocktail state "cocktail" and if a general group state the "group general")” Coding options (select one answer): name (free text) Number of Iterations: 2 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: extra iteration to clean the answers	Description: “Record the type of pesticide tested within the study (if multiple, separate with semicolon, if cocktail state "cocktail" and, if a general group, state the "group general"). Answer by providing only pesticide name, without additional text, comments or context (i.e. do not answer with full sentences)” Coding options: free text (Any answer) Accuracy: 8/8 Comment: Two extraction mistakes found by Elicit
Morrison_2024 / Registered protocol*	Description: “Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan” Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: All no	Description: “Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan” Coding options: Yes, No Accuracy: 8/8 Comment: All no

Morrison_2024 / Authors contributions*	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors” Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors” Coding options: Yes, No Accuracy: 6/8 Comment: NA

Morrison_2024 / Dosage highest ["dosage_highest" and "dosage_highest_unit "]	Description: “Record the dose of the highest exposure group (that is not used as a control). Record the units of concentration used for the highest dose of pesticide exposure (if unclear or not reported select "not reported")” Coding options (select one answer): Value and units (free text) Number of Iterations: 3 Initial accuracy: 6/8 Final accuracy: 7/8 Success: yes Comment: in khotimah_2015 at max. concentration all fish died so not used for behavioral tests data extractions - so original extraction was correct too	Description: “Record the highest dose of pesticide fish were exposed to during experiments measuring effects on behaviour. Record the units of concentration used for the highest dose of pesticide exposure. If unclear or not reported, answer "not reported". Answer by providing only a single number and unit, without additional text, comments or context (i.e. do not answer with full sentences)” Coding options: free text (Any answer) Accuracy: 7/8 Comment: two mismatches due to additional criteria, not errors.

Mizuno_2024 / Bird species [Bird_species]	Description: “Bird species Latin binomial name (if specified)” Coding options: Name (free text) Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: extra iteration to clan the answers	Description: “Report species Latin binomial name of birds used in the experiment on the effects of conspicuous wing patterns such as eye-spots (if specified). If not specified, code as "none". If more than one bird species used, list all bird species names, separated by a comma. Answer by providing only Latin binomial name of birds used in the experiment on the effects of conspicuous wing patterns such as eye-spots, without additional text, comments or context (i.e. do not answer with full sentences)” Coding options: free text (Any answer) Accuracy: 6/8 Comment: NA
Mizuno_2024 / Bird sex [Bird_sex]	Description: “Subject bird sex (if specified)” Coding options (select one answer): male, female, both Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “Subject bird sex (if specified). If not specified, code as "NA". If females and males used, code as "both" only” Coding options: male, female, both, NA Accuracy: 6/8 Comment: NA

Mizuno_2024 / Bird age [Bird_age]	Description: “Subject bird age (if specified)” Coding options: adult, chick Number of Iterations: 2 Initial accuracy: 3/8 Final accuracy: 7/8 Success: yes Comment: had to add “NA” and "nestling" to the answers (not in meta-data, but coded)	Description: “Subject bird age (if specified). If not specified, code as "NA". Only consider experiments on wing spots and eyespots” Coding options: adult, chick, nestling, NA Accuracy: 7/8 Comment: NA
Mizuno_2024 / Treatment stimulus [Treatment_stimulus]	Description: “Type of presented stimulus pattern (eyespots or non-eyespot but conspicuousness pattern)” Coding options: eyespot, non_eyespot Number of Iterations: 5 Initial accuracy: 0/8 Final accuracy: 0/8 Success: no Comment: not extractable	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase

Mizuno_2024 / Type prey [Type prey]	Description: “Prey material type. 'Real' if a real/imitation of a particular butterfly was used as prey, otherwise artificial” Coding options: real, artificial Number of Iterations: 2 Initial accuracy: 6/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “Prey material type. Answer 'Real', if the authors used a real butterfly/moth/caterpillar, or patterning of a real butterfly/moth/caterpillar species, or a modified patterning of a real butterfly/moth/caterpillar species. If the authors used geometrical or simplified patterns, answer "artificial"” Coding options: real, artificial Accuracy: 8/8 Comment: NA
Mizuno_2024 / Conflict of interests*	Description: “Whether the conflict of interests was disclosed by the authors” Coding options (select one answer): Yes, Not declared specifically Number of Iterations: Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: “Was the conflict of interests disclosed by the authors. Predefined list (select one answer): yes, not declared specifically” Coding options: Yes, Not declared specificall Accuracy: 7/8 Comment: NA

Mizuno_2024 / Funding Source*	Description: "Whether the organisations and bodies that funded the study are mentioned" Coding options (select one answer): Yes, No funding sources recorded Number of Iterations: 5 Initial accuracy: 1/8 Final accuracy: 2/8 Success: no Comment: not extractable	Description: NA Coding options: NA Accuracy: NA Comment: variable not used in the testing phase
Mizuno_2024 / Supplementary materials*	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links" Coding options (select one answer): Yes, No Number of Iterations: 2 Initial accuracy: 7/8 Final accuracy: 8/8 Success: yes Comment: extra iteration to refine the prompt	Description: "Whether the article contains mentions of online supplementary materials (additional details of methods, data, code, etc.). Supplementary materials can be either appended at the end or stored in separate files or accessible online links. Supplementary materials are sometimes called Supporting Material or Appendix or Additional File" Coding options: Yes, No Accuracy: 8/8 Comment: NA

Mizuno_2024 / Data shared*	Description: "Whether the raw data used for analyses presented in the article has been shared either in supplementary materials, or as a link to an external file repository. If data is available on request, it will be coded as No" Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: NA	Description: "Whether the raw data used for analyses presented in the article has been shared either in supplementary materials, or as a link to an external file repository. If data is available on request, it should be coded as No" Coding options: Yes, No Accuracy: 8/8 Comment: NA
Mizuno_2024 / Registered protocol*	Description: "Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan" Coding options (select one answer): Yes, No Number of Iterations: 1 Initial accuracy: 8/8 Final accuracy: 8/8 Success: yes Comment: All no	Description: "Code whether the study has been registered, pre-registered or is based on a pre-defined protocol or study plan" Coding options: Yes, No Accuracy: 8/8 Comment: All no

Mizuno_2024 / Authors contributions*	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors” Coding options (select one answer): Yes, No Number of Iterations: 2 Initial accuracy: 6/8 Final accuracy: 7/8 Success: yes Comment: NA	Description: “Code whether the study contains a description of the roles and/or contributions of the study authors. Do not include acknowledgements of the people who helped but are not authors of the study.” Coding options: Yes, No Accuracy: 6/8 Comment: NA

Table S3

Summary of the prompt development phase in Elicit for 7 systematic reviews. Full references to published reviews are provided in the reference list. For each review, prompt development in Elicit started from 10 pre-defined variables and initial data extraction prompts based on meta-data and descriptions from the reviews. Prompts were iteratively modified for up to five interactions of data extractions in Elicit. For each iteration, extracted variables were compared to the values from the original review using a random subset of 8 studies. Variables that did not achieve a pre-specified accuracy threshold of 87% (0.87 = 7/8 correct answers) within 5 prompt refinement iterations, were replaced with other variables selected from the review or generic variables related to reporting until 10 variables were successfully developed for each review.

Review	**Variables tested to get 10 successful variables**		**Number of prompt refinement iterations needed to achieve success per variable**				**Final accuracy of 10 successful variables**			
	Total	Success rate	Min	Max	Mean	Median	Min	Max	Mean	Median
Lagisz_2020	13	0.77	1	5	1.8	1.0	0.88	1.00	0.95	1.00
Mizuno_2024	12	0.83	1	5	2.1	2.0	0.88	1.00	0.92	0.88
Morrison_2024	14	0.71	1	3	1.7	2.0	0.88	1.00	0.97	1.00
Pollo_2024	14	0.71	1	3	1.9	2.0	0.88	1.00	0.91	0.88
Ricolfi_2024	13	0.77	1	5	1.2	1.5	0.88	1.00	0.92	0.88
Samarsinghe_2019	13	0.77	1	4	2.0	2.0	1.00	1.00	0.99	1.00
Yang_2024	11	0.91	1	2	1.4	1.0	0.88	1.00	0.94	0.94

Table S4

Summary of the errors in human-extracted data detected via *Elicit* extractions.

Nr	Review / Variable	Study	Error
1	Samarsinghe_2019 / Geographically focused	Pomponi_2016	One miscategorised value found by Elicit (should be Yes, but originally coded as No): study focuses on temperate regions, so could be considered as geographically focused
2	Samarsinghe_2019 / Geographically focused	Mackenbach_2014	ne miscategorised value found by Elicit (should be Yes, but originally coded as No): study focuses on high income countries, so could be considered as focused on particular regions
3	Yang_2024 / Sex	Renthlei_2019	One missing value found by Elicit (male study subjects)
4	Yang_2024 / Tissues measured	Iigo_2003	One inconsistently coded value identified by Elicit ("Eye" instead of "Retina")
5	Morrison_2024 / Pesticide investigated	Liu_2020	One incorrectly extracted value identified by Elicit: rotenone vs. carbofuran
6	Morrison_2024 / Pesticide investigated	Boyda_2021	One incorrectly extracted value identified by Elicit: deltamethrin vs. diazinon
7	Pollo_2024 / Number of species	Dougherty_2016	One incorrect value identified by Elicit - study on two species (Lygaeus equestris L. and L. simulans), but it is possible that only data for one was used in a meta-analysis
8	Ricolfi_2024 / Funding sources	Ferguson_2013	One miscategorised value found by Elicit (should be Yes, but originally coded as No) - study acknowledges funding sources