Research Synthesis Methods (2026), **00**: 1–18
doi:10.1017/rsm.2026.10080

Research Synthesis Methods

RESEARCH ARTICLE

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

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

[1]Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, UNSW Sydney, Australia
[2]Department of Biological Sciences, University of Alberta, Canada
[3]School of Environmental and Life Sciences, The University of Newcastle, Australia

Corresponding author: Malgorzata Lagisz; Email: m.lagisz@unsw.edu.au

Received: 25 July 2025; **Revised:** 11 January 2026; **Accepted:** 16 January 2026

Keywords: artificial intelligence; evidence synthesis; meta-analysis; proof of concept; research methods; systematic maps

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 the potential for saving time, 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 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, 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 for sanity checks 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.

This article was awarded Open Data and Open Materials badges for transparent practices. See the Data availability statement for details.

A.M., K.M., P.P., L.R., and Y.Y. have contributed equally to this work and are listed in alphabetical order.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

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 RSM readers
Elicit can support data extraction in systematic reviews, but should be used cautiously and with human oversight. Our study offers practical guidance on integrating LLM tools, while highlighting current limitations in replicability and reasoning.

1. Introduction

The use of 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 that processes and learns 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.[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.[2,3,5,8–10,14] More specifically, these publications emphasised 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 the reliability and reproducibility of systematic reviews remains a key area for research.

Elicit (elicit.ai; elicit.com) is one such technology that holds great promise for streamlining and accelerating systematic reviews.[13,15] It stands out from the generic tools like *OpenAI ChatGPT*, *Microsoft Copilot*, or *Google Gemini*, 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] *Elicit* also allows uploading and

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

analysing PDF files of full-text documents that may not be available in its online database. A built-in LLM algorithm evaluates the semantic similarity of publication texts to find and rank publications for inclusion in a systematic review. Further, other LLMs that are trained on academic literature can generate article summaries and extract data that is commonly collected for systematic reviews, for example, 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]

Despite its promise, two recent case studies point to limitations in the 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 article explores the potential applications of the *Elicit* platform in evidence synthesis with two specific aims:

1. To evaluate the effectiveness of LLMs integrated into *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 use of AI 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] as feasible. 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 of 2024/09/30). The *Plus* subscription plan may be a relatively affordable option for researchers who wish to evaluate the suitability of the *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. The *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 reviews and maps, umbrella reviews, and meta-analyses (hereafter "systematic reviews" as they all belong to a "systematic reviews family")[23] on diverse topics from life and environmental sciences (Supplementary Table S1[24–30]; all reviews were preregistered and published with accompanying data, which were either double-extracted or cross-checked by a second researcher). The lead authors of these systematic reviews are co-authors of

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

this project. They provided underlying extracted raw data and metadata 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 the 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, and analytical code). If variables related to reporting quality were not extracted in an original systematic review or meta-analysis, two researchers (M.L. and either A.M., K.M., L.R., P.P., S.N., or Y.Y.) independently extracted them to create a gold standard answer for each additional variable. The overall accuracy of these additional human-based data extractions was 98.6% (1.4% error rate, i.e., six mismatched values between two human extractions out of 416 de novo double-extracted values for 26 variables across all seven 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 predefined 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 the *Elicit* user workspace. The *Elicit* workspace facilitates the 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 of the seven original systematic reviews, we randomly sampled eight included full-text studies for conducting prompt development (training set) and another eight full-text studies for evaluation (test set). We 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 predefined 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 of eight 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).

We documented the prompt development process (DEV phase), noting any alterations to the original metadata and data from the seven published systematic reviews (e.g., pooling or changing to a free text option when the number of categories of a variable exceeded the limit of eight categories allowed in *Elicit*) and other encountered issues (Supplementary 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 metadata from the reviews, as used to construct initial data extraction prompts in *Elicit*, and final prompts developed using *Elicit*.

In the main testing phase (TEST), we used the test set of eight studies (different from the DEV phase) per review to evaluate the accuracy of *Elicit* extractions using the final prompts from the development

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

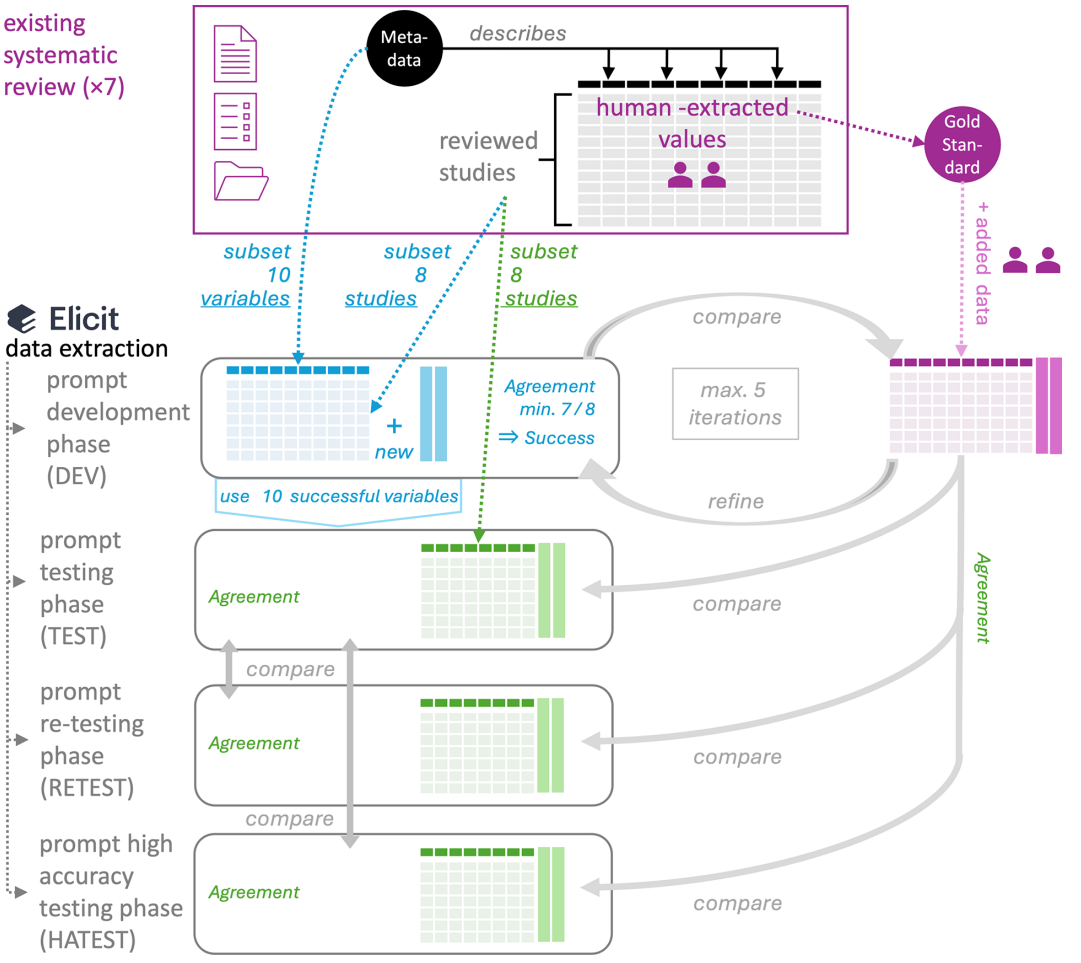


Figure 1. *Diagram of our approach used to develop and evaluate data extractions on the* Elicit *platform. We started the project with seven systematic reviews (systematic maps, meta-analyses, and umbrella reviews) representing different topics in ecology, evolution, and environmental sciences. We used human-extracted values from these reviews as our gold standard data and metadata 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, and 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 the 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 with a different* Elicit *user account to test the replicability of the extractions. In the HATEST phase, we ran another test using high-accuracy mode to assess whether it improved the accuracy of the data extractions. For a detailed description of the underlying datasets and workflow phases, see the Methods section.*

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

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 the repeatability of extractions across user accounts in *Elicit*, we 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 not being representative of *Elicit*'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, we repeated all extractions from the TEST phase. This high-accuracy-testing phase (HATEST) allowed us to compare extractions conducted by two different versions of the *Elicit* algorithms using the same set of studies (full-text PDF files) and data extraction prompts (TEST–HATEST comparison).

We 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 to interpret 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 the percentage 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 by *Elicit*). For the testing phase (TEST) and the two re-testing phases (RETEST, HATEST), we reported accuracy as well as reasons for potential mismatches.

2.4. Deviations 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 five 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 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

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

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 (Supplementary 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). A change to a high-accuracy algorithm can be expected to significantly influence the 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 the 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 five iterations). Per review, this success rate varied from 71% to 91% (Supplementary Table S3). In other words, we had to try 11–14 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 the remaining 12 variables were successful after 3–4 iterations. In contrast, we ran five iterations of prompt refinement for each of the 20 "non-successful" variables that failed to reach the desired accuracy level. This means that 100 iterations of data extractions did not lead to success, out of a total of 231 iterations run across the whole development phase (i.e., 43% of effort or time was wasted).

Next, we categorised all our variables into four types, based on the expected structure of the answer (data type) (Figure 2). Variables that required a Yes/No answer comprised 49% of our dataset in this phase (44 tested variables) and had an overall success rate of 82%. We had 27 variables with predefined categorical answers (2–11 categories, e.g., sex or age class of study subjects, and 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 six 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 seven 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 in extracting information on data availability

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

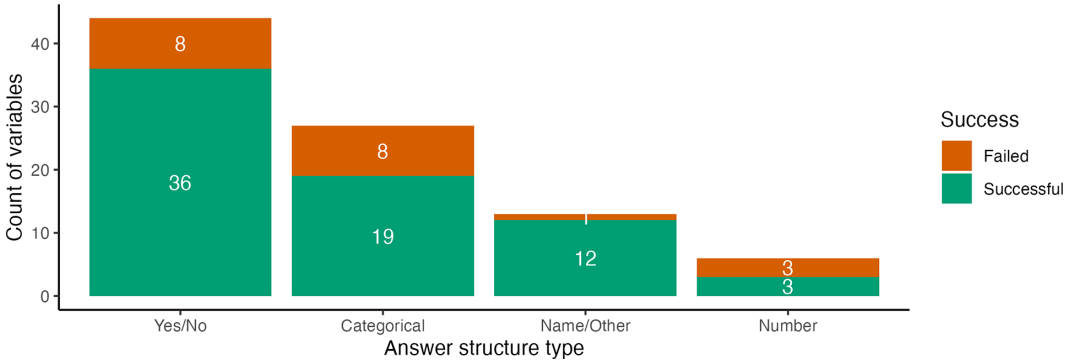


Figure 2. *Distribution and success of extraction variables by the type of expected answer during the prompt development phase of the project. A variable was considered "Successful" if at least seven out of eight answer values were extracted correctly in* Elicit*, that is, matched human-extracted gold standard values, within a maximum of five iterations of data extraction prompt refinement. The data underlying this prompt development stage comprise 90 variables considered across seven systematic reviews, with eight studies extracted per review. The "Categorical" answer structure includes categorical variables (e.g., "Tissue measured" variable being coded as Blood, Pineal, SCN, Urinary, Retina, and Water; "Age" being coded as Juvenile, and Adult), except the binary Yes/No answers, which are shown separately. We used a "Yes/No" answer structure to code the presence or absence of certain information or practices in a study (e.g., the presence of a conflict of interest statement or whether raw data are shared). The "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 as "Any answer type" in* Elicit*. The "Number" answer structure includes numeric variables (e.g., simple size, number of cues in a behavioural assay).*

(only succeeded for two out of six reviews) and disclosure of funding sources (only succeeded for one out of six reviews) (Figure 3). The remaining variables that we used were directly related to the 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, and data availability) were as likely as variables related to study scope or methods (e.g., study species, location, sample size, exposure dosage or duration, and 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 eight studies from each of the seven systematic reviews, with 10 variables tested per review (successful variables from the prompt development phase).

Overall, 48 out of the 70 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.[30] and Ricolfi et al.[24]) to 100% (Lagisz et al.[29]), and the extraction accuracy varied across the variables (Figure 4).

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press



Figure 3. *Distribution and success of 90 considered extraction variables (*y*-axis) across seven systematic reviews (*x*-axis) during the prompt development stage. The 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 "Successful" if at least seven out of eight answer values were extracted correctly in* Elicit*, that is, matched human-extracted "gold standard" values, within a maximum of five 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.*

The overall extraction accuracy during the prompt development phase and test phase was positively related ($r = 0.38$, $t = 3.38$, df = 68, $p = 0.001$), indicating that more accurate data extractions during the development phase were also more likely to be accurate during the testing phase. Similarly to the prompt development phase, there was no association between the type of 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

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

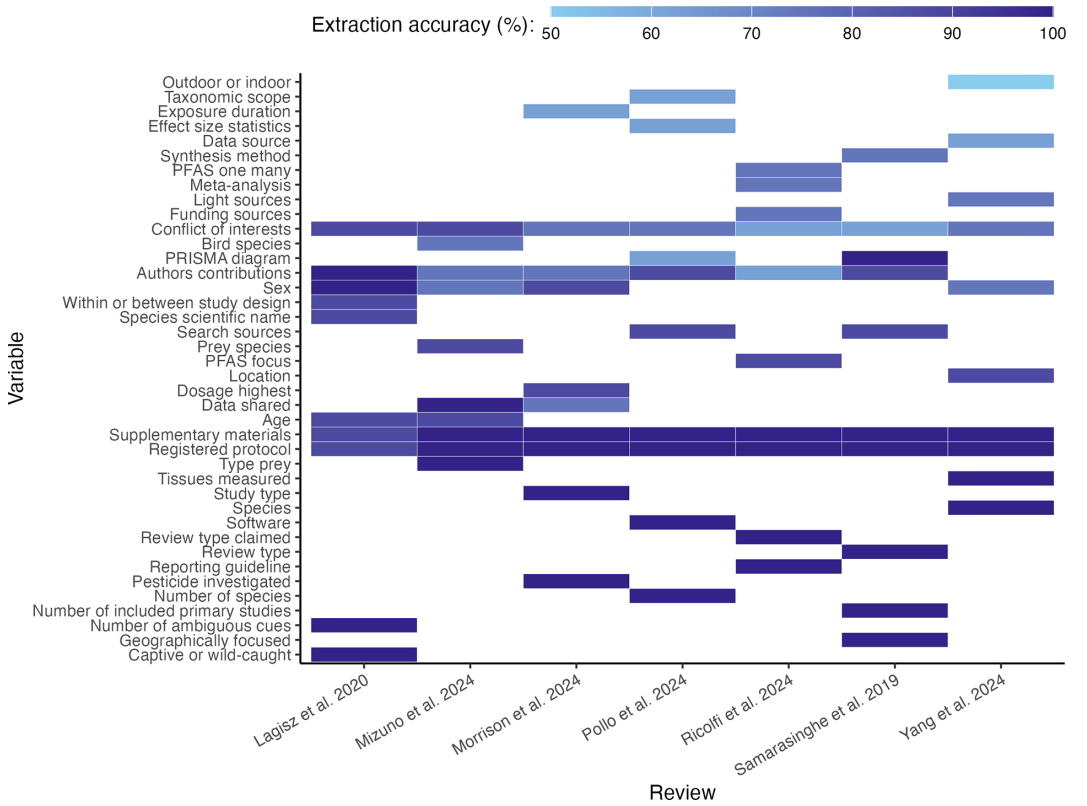


Figure 4. *Accuracy of data extractions performed using the* Elicit *platform when compared to human-extracted gold standard answers. We tested extraction variables that passed the prompt development stage with at least an 87% accuracy (7/8 answers correct) rating (*y*-axis) for each of the seven systematic reviews (*x*-axis) using a new test set of eight studies distinct from the prompt development stage per review. The colour of cells in the grid indicates the 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.*

(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.[28] review dataset).

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 the 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 the *Elicit* interpretation error. These mismatches occurred because the TEST extraction was correct, but the RETEST extraction was incorrect (19), the TEST extraction was incorrect, but the 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 semantically equivalent and correct (Figure 5b).

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

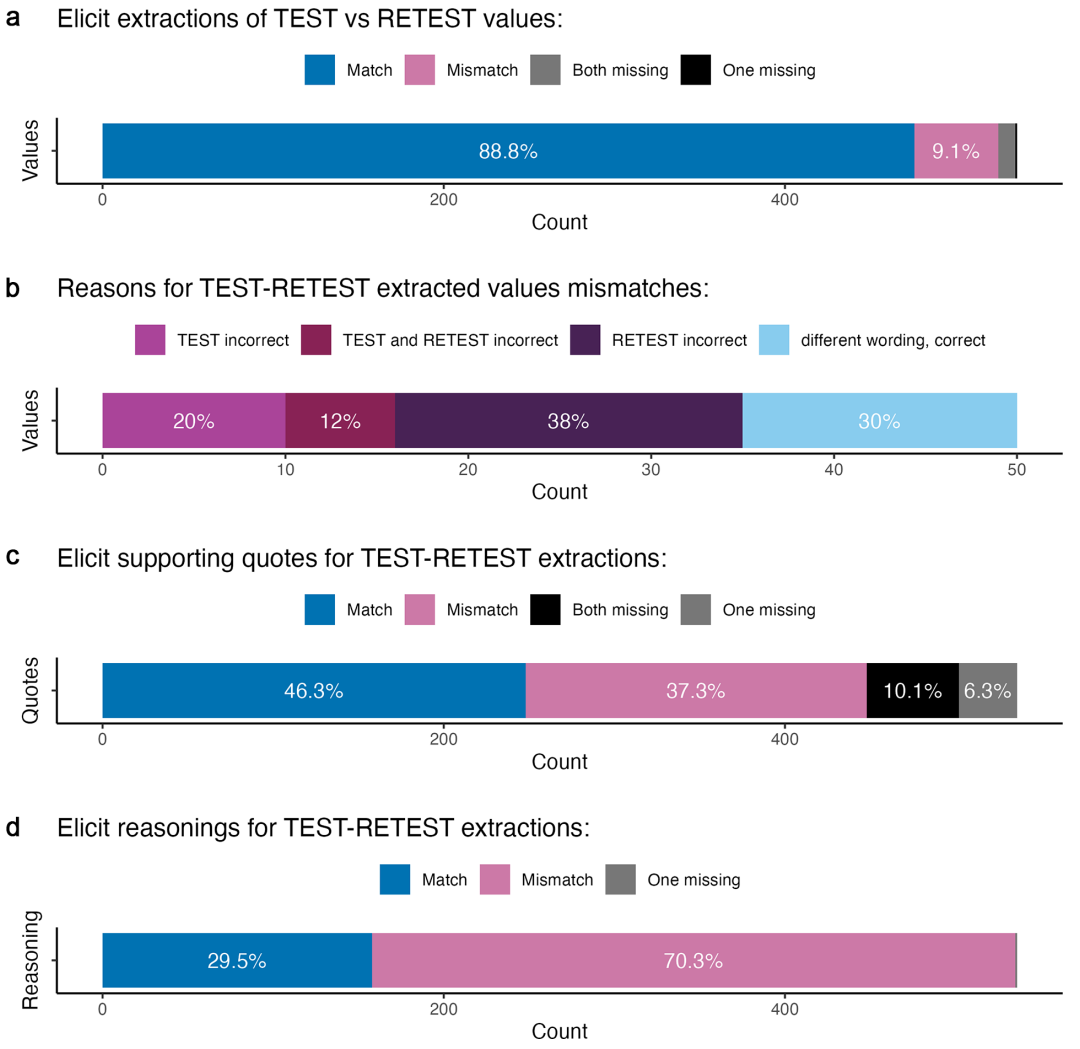


Figure 5. *Comparison of results from using two different accounts in* Elicit *to extract 10 test variables for each of the seven original systematic reviews (TEST–RETEST phases). Plots represent exact matching of 536 extracted values (a), classification of the reasons of mismatched values (b), comparisons of corresponding supporting quotes (c) and reasoning (d) provided by* Elicit.

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, and 54 both in TEST and RETEST).

For the reasoning narratives provided by *Elicit* to justify its extracted values, 158 of the cases matched 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, $df = 1$, $p = 0.724$).

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

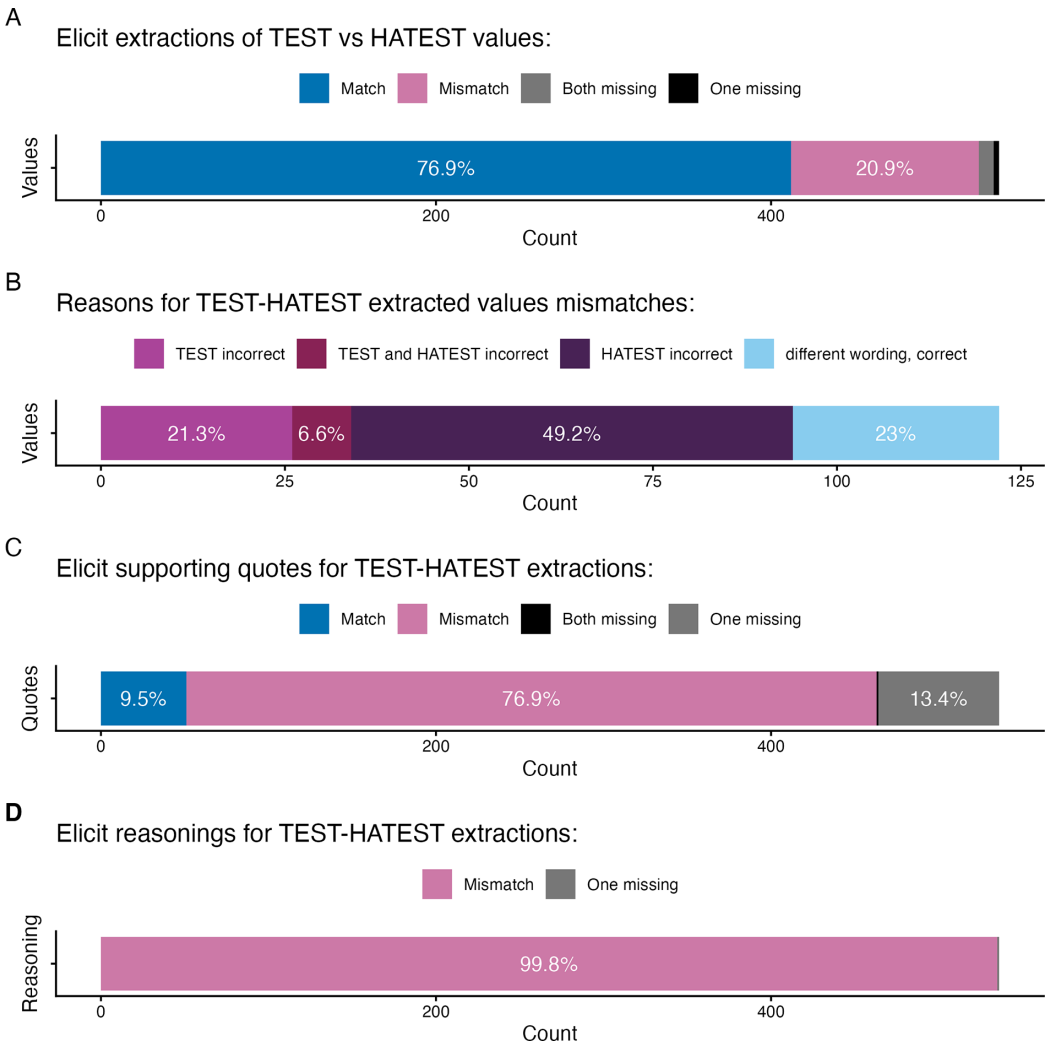


Figure 6. *Comparison of results from re-running test extractions (10 test variables for each of the 7 original systematic reviews) in* Elicit *after the platform enabled a free high-accuracy mode for all accounts and plans (TEST–HATEST phases). Plots represent exact matching of 536 extracted values (a), classification of the reasons of mismatched values (b), comparisons of corresponding supporting quotes (c), and reasoning (d) provided by* Elicit.

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 were 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 the values that were extracted earlier (412 out of 536; Figure 6a). Both extraction rounds also failed to extract nine 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 errors. These mismatches occurred because the TEST extraction was correct, but HATEST extraction was incorrect (60), the TEST extraction was incorrect, but HATEST extraction was correct (26), and both the TEST extraction

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

and the HATEST extraction were incorrect (8). We further subdivided these mismatches into cases where TEST extraction was correct, but HATEST extraction was incorrect (60), cases where TEST extraction was incorrect, but HATEST extraction was 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 in both TEST and HATEST).

For the reasoning narratives provided by *Elicit* to justify its extracted values, none of the cases matched 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 eight human-made errors (<1%; Supplementary 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 extraction 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 metadata (i.e., variable descriptions) from original reviews (conducted exclusively by human reviewers) 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 using a new data subset. 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. Surprisingly, overall accuracy was slightly lower (82%). Supporting quotes matched only 10% of the time, and the wording of the 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.

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

Spillias et al. 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 predefined 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, and 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. extracted seven variables from 20 randomised 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 emphasised that human verification remains essential when using *Elicit* for data extraction.

Helms Andersen et al. 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 *Elicit* 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. 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 interest 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 standardised and easily recognised 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 the 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).

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

4.5. Effects of answer structures and formats in Elicit

Elicit offers a range of default pretrained 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 eight 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 visualisation. 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 eight 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 Material, raw data files, or metadata hosted on external platforms, which *Elicit* cannot access. Even when data are available, they 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. Using LLM tools like *Elicit* requires uploading full-text PDFs, raising concerns around copyright, licensing, and data privacy (e.g., for paywalled or proprietary content). While *Elicit* states that "all PDFs you upload remain private to your account and are not shared or accessible to any other users" (https://support.elicit.com/en/articles/723521), broader issues remain regarding long-term storage, licensing terms, and ethical use.

4.7. Gaps in article metadata extracted by Elicit

Missing or incomplete metadata retrieval from uploaded PDF files is another noticeable issue. We observed that *Elicit* often failed to extract key metadata 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

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

Elicit's performance, image-only or scanned PDFs were excluded in our and other works[11] and thus did not contribute to the presented performance evaluation scores.

4.9. The trade-off of time investment

Our results indicate that setting up and testing data extractions is not always successful and can be time-consuming. Some variables fail to reach the desired level of accuracy, while others require several iterations to achieve satisfactory results. It is also difficult to predict in advance which variables will perform well, necessitating testing across all variables under consideration and limiting the use of *Elicit* to those yielding acceptable accuracy.

The key question, therefore, is whether using *Elicit* ultimately saves time in data extraction for systematic reviews while delivering performance comparable to that of human reviewers. The answer depends on the balance between the time invested in the setup and testing phases and the time saved during the main data extraction phase. For instance, investing a week to optimise extractions in *Elicit* is worthwhile if it reduces more than a week of manual labour during the main extraction stage, which is an outcome that may be particularly relevant for large-scale systematic reviews involving hundreds of studies, or more. It is noteworthy that *Elicit* can also reveal mistakes made by human extractors. Moreover, human extractors also require time investment into process optimisation and training, which needs to be considered when assessing the advantages of automated systems. Automated coding may also be useful for the scoping of research questions and piloting data extractions.

4.10. Limitations and strengths of our study

Our study has limitations related to sample size and the 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. We could not compare extraction times between automated and manual workflows due to a lack of timing data at the variable level for human extractors in the original reviews.

The strengths of our study lie in its robustness and transparency. We followed a preregistered 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 that, in the *Elicit* platform, (1) developing prompts that reach a predefined level of accuracy in data extractions requires considerable effort, (2) accuracy of extractions varies substantially, (3) data extractions across user accounts are highly repeatable, and (4) distinct LLM algorithms for data extraction produce different results. The variable descriptions intended for human researchers (metadata) often require iterative refinement and testing in order to achieve accuracy on par with human extractors, which imposes additional time cost and limits the use of *Elicit* for data extractions in systematic reviews. Despite its shortcomings, *Elicit* offers accessible advanced functionalities for extracting simple data from the full text of research studies and can offer some support at this stage of the systematic review process, especially when simple data are extracted from hundreds or thousands of studies. However, challenges remain in ensuring the quality and replicability of extracted data, particularly when information is not explicitly reported, is presented in inaccessible formats, or requires nuanced interpretation. Further, upgrades to the underlying algorithms implemented in the *Elicit* online platform may affect performance and compromise reproducibility. Under these conditions, we recommend integrating *Elicit* into a modified systematic review workflow for sanity checks or as a

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

secondary extractor alongside a human reviewer for large-scale systematic reviews. A third reviewer could then reconcile discrepancies between human- and *Elicit*-extracted data, thereby improving efficiency while maintaining high accuracy.

Author contributions. Conceptualisation: M.L. and S.N.; Data curation: M.L.; Formal analysis: M.L.; Funding acquisition: M.L. and S.N.; Investigation: M.L., A.M., K.M., P.P., L.R., Y.Y., and S.N.; Methodology: M.L. and S.N.; Project administration: M.L.; Software: M.L.; Supervision: M.L. and S.N.; Visualisation: M.L.; Writing—original draft: M.L.; Writing—review and editing: M.L., A.M., K.M., P.P., L.R., Y.Y., and S.N.

Competing interest statement. We acknowledge that we used a 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 the conceptualisation or design 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.

Data availability statement. The project GitHub repository with all data and code can be found at https://github.com/mlagisz/elicit_extractions_testing and its archived release v.1.0.0. on Zenodo at https://doi.org/10.5281/zenodo.18636510.

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

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

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/rsm.2026.10080.

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1136/bmjopen-2023-072254.

[6] Cao C, Arora R, Cento P, et al. Automation of systematic reviews with large language models. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1186/s12874-025-02528-y.

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press

[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. https://doi.org/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. https://doi.org/10.1002/jrsm.1553.
[15] Kung J. Elicit (product review). *J Can Health Libr Assoc J Assoc Bibl Santé Can*. 2023;44(1). https://doi.org/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. https://doi.org/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. https://doi.org/10.1186/s13750-025-00362-9.
[19] Thomas J, Flemyng E, Noel-Storr A. Responsible AI in evidence synthesis (RAISE): guidance and recommendations. 2024;29. https://doi.org/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. https://doi.org/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:RP103339. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1111/ele.14387.

Cite this article: Lagisz M, Mizuno A, Morrison K, Pollo P, Ricolfi L, Yang Y, Nakagawa S. Using *Elicit* AI research assistant for data extraction in systematic reviews: A feasibility study across environmental and life sciences. *Research Synthesis Methods*. 2026;00: 1–18. https://doi.org/10.1017/rsm.2026.10080

https://doi.org/10.1017/rsm.2026.10080 Published online by Cambridge University Press