AI-assisted Living Evidence Databases for Conservation Science

Sadiq Jaffer*[1], William Morgan[2], Sam Reynolds[2], Alec Christie[2,3], Anil Madhavapeddy[1], William Sutherland[2]

*Corresponding author

Affiliations:

[1] *Department of Computer Science & Technology, University of Cambridge, Cambridge UK*
[2] *Conservation Science Group, Department of Zoology, University of Cambridge, Cambridge, UK*
[3] *Centre for Environmental Policy, Imperial College London, London, UK*

Living evidence databases offer a robust and dynamic alternative to static systematic reviews but require a resilient technical infrastructure for continuous evidence processing. This working paper describes the architecture and implementation of a complete, end-to-end pipeline for this purpose, developed initially for the conservation science domain. Designed to operate on local infrastructure using self-hosted models, the system ingests and normalizes documents from academic publishers, screens them for relevance using a multi-stage process, and extracts structured data according to a predefined schema. Key features include a hybrid retrieval model; a human-AI collaborative process for refining inclusion criteria from complex protocols, and the integration of an established, statistically-principled stopping rule to ensure efficiency. In a baseline evaluation against a prior large-scale manual review, the fully automated pipeline achieved 97% recall and identified a significant number of relevant studies not included in the original review, demonstrating its viability as a foundational tool for maintaining living evidence databases.

This is a living document: version 0.0.1 last updated 3rd October 2025.

1. Introduction

The established paradigm of conducting discrete, static systematic reviews has been key to the development of evidence-based policy and practice in fields such as medicine, education, international development and environmental management [1]. However, such reviews are costly and time-consuming to conduct, difficult to reproduce, and quickly go out of date as new studies are published and existing ones retracted. These lengthy timelines often produce answers outside the narrow windows required for policy decisions [2], [3]. Furthermore, the narrow scope of systematic reviews renders their findings effectively “single-use”: the focal question is answered comprehensively, yet answering even highly related questions requires starting from scratch and re-searching the entire literature [4]. These challenges may lead to decision makers resorting to Artificial Intelligence (AI) models that deliver almost instant answers. However, these lack transparency about their evidence base and potential biases. Artificial Intelligence (AI) can be used to provide incremental improvements in the efficiency of the systematic review process, but the inflexible structural limitations of the final review remain.

Current advances in AI-enabled evidence synthesis are often presented as a false dichotomy between opaque, blind automation and marginal productivity gains. Instead, we argue that

advances in AI have brought us to an inflection point, where a fundamentally different approach becomes viable: subject-wide, living evidence databases, powered by human-in-the-loop AI. These databases will continuously collate, index and catalogue evidence across a domain to enable delivery of rapid answers to a variety of related research and policy questions. Crucially, these systems should be transparent and traceable, allowing a 'human-in-the-loop' to scrutinise model decisions and associated reasoning at all stages.

The main barrier to the vision of living evidence databases is the lack of practical, end-to-end technical infrastructure designed for this purpose, especially under real-world operational constraints (i.e. integrating with academic publishers). While systems exist that could be used to implement AI into elements of the evidence synthesis process, such as screening or data extraction, the lack of an end-to-end system means there is no clear traceability or ability to propagate confidence through the system from ingestion to output.

1.1 From static to living reviews

While evidence reviews have proven crucial in fields like medicine for practitioners to keep up with evidence, they face a number of challenges. Research suggests that they go out of date quickly [5] with half of the medical reviews examined needing material updates within 5.5 years. The comprehensiveness of evidence reviews is threatened by a rapidly growing literature base - the number of articles indexed by Scopus and Web of Science grew 47% between 2016 and 2022 [6] - a trend likely to accelerate with the aid of AI [7]. Furthermore, redundancy and duplication of efforts in the process of conducting reviews affects cost-efficiency and timeliness; with each review searching, categorising and synthesising the same literature. For some topics, the number of evidence syntheses even exceeds the number of primary studies [8].

To address reviews becoming rapidly out of date, Living Systematic Reviews (LSRs) have been proposed as a solution [9]. These are online databases that are continuously updated and re-reviewed by the original authors when new relevant studies are identified. Their uptake has been gradual but has increased rapidly since 2019 [10]. However, they do not address the duplication of efforts between reviews that cover much of the same literature and do not overcome the issue of being able to directly apply extracted information to highly relevant alternative questions.

Subject-wide evidence synthesis [11], [12] is a method of evidence synthesis that is well designed for the production of Living Evidence Databases (LEDs). Subject-wide evidence synthesis is a systematic method for finding, summarising, and assessing all relevant research evidence on a broad topic to inform decisions and practices. Similarly, LEDs are based on broad subject areas and feature many different interventions and outcomes. A populated LED would allow reviews to be conducted extremely quickly, by simply filtering for the evidence relevant to your question. Crucially the work of searching, data extraction and critical appraisal is shared amongst all produced reviews - rather than repeated for each one, as in systematic reviews.

1.2 The need for automation

With the deluge of publications [13] and the dramatic improvements in AI capabilities in recent years, proponents of Systematic Reviews and Living Evidence Databases have proposed

accelerating one or more steps in the evidence review process with AI. Existing research has focused on different stages in a typical systematic review process. For example, research protocol and search strategy[1], search string generation [14], [15], screening [16], [17], [18], [19] and data extraction [20], [21], [22], [23]. However, tooling is fragmented and necessitates using different software packages together in order to conduct a review. This poses a challenge for living reviews as there is limited traceability between stages, which prevents the propagation of confidence measures and limits auditability. It is also dependent on the ongoing support for a range of software packages. Solving these issues is crucial for long-running maintenance of LEDs and reliable evidence synthesis.

1.3 Our Contribution

We present a complete, end-to-end AI-enabled evidence synthesis pipeline designed as a practical solution to the challenges facing evidence synthesis. Our contributions are as follows:

1. A system that can operate fully self-hosted, with built in traceability at all stages.
2. We propose a novel human-in-the-loop process for criteria refinement that uses synthetic data and enables domain experts to rapidly and precisely refine complex inclusion criteria.
3. We demonstrate real-world feasibility through integrating the pipeline with multiple academic publishers, moving beyond the limitations of typical academic projects that are restricted to open-access corpora only.
4. Finally, we rigorously evaluate the pipeline against the large-scale, manually curated Conservation Evidence database[2], demonstrating 97% recall against the prior human manual review and showing that our system enhances the review's comprehensiveness by identifying hundreds of relevant studies the original process overlooked.

2. A Pipeline for Living Evidence

2.1 System Overview

The system is designed to execute configurable data processing pipelines. A pipeline is a directed graph of nodes, where each node performs a specific function. The primary node types are: `criteria` nodes, which test articles against a natural language inclusion criteria; `filter` nodes, which apply a conditional filter using a user-provided code snippet; `extract` nodes, which extract structured data; and `output` nodes, which aggregate results. This modular structure allows for the creation of complex, multi-stage workflows for processing scholarly documents.

2.2 Data Ingestion and Normalization

A principle difference between conventional systematic review methodology automation, and our subject-wide evidence synthesis based evidence pipeline is the lack of keyword-based filtering in our approach. The process begins by identifying a target set of documents using

[1] https://mesh.med.yale.edu/, https://picoportal.org/
[2] https://conservationevidence.com/

their Digital Object Identifiers (DOIs) [24]. Article DOIs can be sourced from data sources such as the OpenAlex dataset [25] or Semantic Scholar [26]. In collaboration with the university library, agreements have been established with the following academic publishers to permit access to both open and closed access full-text articles, either via API or by scraping their websites: Public Library of Sciences (PLOS), Oxford University Press, Elsevier, Springer, Frontiers, The Royal Society, Wiley, Taylor & Francis, Sage, CSIRO and BioOne.

The acquired documents, arriving in heterogeneous formats, are then normalized. PDFs are processed using Grobid [27] to generate XML (eXtensible Markup Language) TEI (Text Encoding Initiative). The XML format creates files that encode data in a format that is both human and machine readable and following the TEI guidelines ensures that the text is encoded in a consistent, semantic way, focusing on meaning rather than just presentation. Source files already in XML format were converted to XML TEI using `pub2TEI` [28]. The evaluation described in this paper did not include articles for which only a JSON format was available. Figures were not included in the TEI conversion.

2.3 Relevance Screening

Relevance screening is an iterative process designed to efficiently identify relevant articles from the large initial corpus. It employs a Continuous Active Learning process using a Large Language Model, human or hybrid as the oracle.

First, a feature vector is constructed for every article in the dataset using the concatenation of its title and abstract. This is a one-time process. The vector is a concatenation of sparse and dense features. The sparse component is generated through uni- and bi-gram tokenization into a fixed-size (2^18) vector using feature hashing. The dense component is a 256-dimension Matryoshka embedding [29] from the `nomic-ai/modernbert-embed-base` [30] model.

For the Large Language Model, we selected `Deepseek-R1-Llama-3.3-70B` as, at the time of evaluation (January 2025), this was the highest performing reasoning model which could be self-hosted on a single NVIDIA H100 [31].

The screening process then proceeds in batches:

- **Initial Batch:** On the first iteration, a random sample of 50 articles is selected and passed to the inclusion testing workflow.
- **Subsequent Batches:** For all subsequent iterations, a logistic regression classifier is trained to guide article selection. Thirty synthetic articles that meet the inclusion criteria are generated by an LLM. These synthetic positives, along with the confirmed negatives from the previous batch, form the training set. A logistic regression model is trained on these examples. This model then ranks all un-processed articles in the corpus, and the top 50 are selected for the current batch.

Articles selected for a batch undergo a two-phase inclusion test. First, an LLM assesses the title and abstract against the inclusion criteria; this is done five times for each article leveraging model self-consistency [32] through a majority vote of the five outputs. Articles that pass this initial check proceed to a full-text verification stage, which is structured as a boolean data extraction node.

To mitigate the prohibitive cost of processing the entire corpus with the expensive full-text LLM stages, we implemented a statistically principled stopping rule. This approach is based on modelling the discovery of relevant documents as an inhomogeneous Poisson process, drawing heavily from prior work in technology-assisted reviews [33], [34]. That is, we assume the rate of discovery of relevant documents follows a poisson distribution where the rate varies over time through a *rate function*. As the review progresses, the system collects data on which articles pass the full-text relevance check, pairing a binary relevance score (1 for relevant, 0 for not) with the article's rank. We use the hyperbolic rate function:

$$\lambda(x) = \frac{a}{(1 + bcx)^{\frac{1}{b}}}$$

Where x is ranking index and a, b and c are parameters controlling the shape of the function with $0 \leq b \leq 1$. This is fitted, with scipy.optimize.curve_fit, to the observed relevance probabilities (smoothed using a sliding window of 100 ranks). The integral of this function is used to estimate the total number of relevant articles in the corpus. The screening process is then stopped when the number of relevant documents found so far reaches a predefined percentage of this estimated total, with a confidence interval calculated using a Poisson distribution. We set our recall target to 95%, as is common in automated screening for systematic reviews [35], [36]. This allows for a more robust stopping decision, such as terminating the review only when the lower bound of the 95% confidence interval on recall exceeds the target threshold.

2.4 Data Extraction

The extraction process is designed to focus the LLM on the most relevant parts of a document. For a given extraction node (e.g., 'Habitat type'), the article is first broken down into chunks of approximately twenty sentences. A short description of the data to be extracted is used to identify potentially relevant sections within these chunks. These relevant sections are then reassembled in their original document order to form a condensed text.

The LLM is then prompted to extract the structured data from this condensed text, using the full, detailed prompt for that extraction task. For all `criteria` and `extract` nodes, constrained decoding is employed to ensure the output is valid JSON conforming to the predefined schema. The JSON output includes structured reasoning: for `criteria` nodes, a concise `reasoning` field intended for end-users is generated, while `extract` nodes are prompted to provide one or more supporting statements, each linked back to the specific text sections from which the information was derived.

3. Human-AI Criteria Refinement

Translating a high-level research protocol into a precise, machine-operable inclusion criteria is a critical bottleneck. To address this, we use a human-in-the-loop process to collaboratively refine the criteria with domain experts. The generic process is as follows:

- **Initial Draft:** Source documents for the review (e.g., existing protocols, project aims) are collated. An LLM (for this work, Anthropic's Claude 3.5 Sonnet was used) processes these documents to generate an initial candidate inclusion criteria.

- **Iterative Refinement:** A group of domain experts is convened for an iterative feedback session. The LLM is prompted to generate synthetic but plausible titles and abstracts that are designed to be ambiguous under the current version of the inclusion criteria. The domain experts review these ambiguous examples, deciding whether each should be included or excluded and providing a brief justification for their decision. The LLM is then provided with the experts' classifications and justifications, and instructed to improve the inclusion criteria to better resolve the identified ambiguities. This cycle is repeated until the experts agree that the generated examples are no longer revealing significant ambiguities and the criteria is sufficiently precise. For the butterfly and moth synopsis review, we generated three synthetic examples per iteration over five iterations.

4. Evaluation

We evaluated the performance of the Pipeline by attempting to reproduce the Conservation Evidence (CE) series synopsis for global butterfly and moth conservation [37]. CE uses subject-wide evidence synthesis [12] to collate the global evidence for the effectiveness of conservation actions [38]. Authors manually screen titles and abstracts of all articles from conservation relevant journals, and those that evaluate the impacts of any conservation action for any wild taxon (e.g. birds, reptiles, mammals) or habitat (e.g. forests, grasslands, wetlands) are retained for full text screening. Article full texts are then screened, and those that meet inclusion criteria are grouped based on the action they test and a ~200 word summary is written detailing the main results and other key information, including study design and the location and habitat where the action was carried out. The butterfly and moth synopsis collated evidence published up to and including 2018, and while the synopsis covered nearly 300 English-language journals, over 300 non-English journals and 16 report series (Bladon et al., 2022), for the purposes of this exercise, we restricted our evaluation to English-language journals only.

We attempted to download all articles from all journals, and years, searched and assessed for the manually curated butterfly and moth synopsis (Appendix 1-3 from Bladon et al. 2022). Due to certain restrictions, such as the number of scientific publishers with whom we have text-mining agreements, accidental black-listing or rate-limiting of our scraping software, and the non-standard formatting of some journals meaning we were unable to convert them into the required XML TEI format, we could only download 151,727 articles. Within these 151,727 articles, 167 articles included in the butterfly and moth synopsis were covered. These 151,727 articles were used as the basis for the evaluation. The specific pipeline configuration used for the evaluated Conservation Evidence Butterfly and Moth synopsis [39] is detailed in Table 1.

Table 1. The configuration of the living evidence database pipeline.

Pipeline stage (`node`)	**Model**	**Sampler**	**Temp-erature**	**Prompts**	**Input data**	**Output data**
Ranking (`criteria`)	nomic-ai/modern bert-embed-base logistic regression	-	-	-	Title and abstracts	Probability of relevance
Synthetic ranking papers (`criteria`)	Deepseek-R1-Llama-3.3-70B	Topk = 40 TopP = 0.9 MinP = 0.1	2.0	See Table S4	Inclusion criteria	Synthetic titles and abstracts of relevant papers
Inclusion (`criteria`)	Deepseek-R1-Llama-3.3-70B	Topk = 40 TopP = 0.9 MinP = 0.1	0.7	See Table S5	Title and abstract	Include/Exclude (with reasoning)
Full-text (`criteria`)	Deepseek-R1-Llama-3.3-70B	Topk = 40 TopP = 0.9 MinP = 0.1	0.7	See Table S6	Full paper text	Include/Exclude (with reasoning)
Study design (`extract`)	Deepseek-R1-Llama-3.3-70B	Topk = 40 TopP = 0.9 MinP = 0.1	0.7	See Table S7	Full paper text and schema for data	Data matching schema
Geography (`extract`)	Deepseek-R1-Llama-3.3-70B	Topk = 40 TopP = 0.9 MinP = 0.1	0.7	See Table S8	Full paper text and schema for data	Data matching schema
Habitat (`extract`)	Deepseek-R1-Llama-3.3-70B	Topk = 40 TopP = 0.9 MinP = 0.1	0.7	See Table S9	Full paper text and schema for data	Data matching schema

To develop our inclusion criteria, we started with the existing Conservation Evidence criteria (Table S1) and undertook the process of Human-AI Collaborative Refinement (Section 3) to develop a final pipeline-specific inclusion criteria (Table S2). Six members of the Conservation Evidence team and a technical facilitator took part in a workshop to carry out this refinement process.

4.1 Article screening and inclusion

We evaluated article screening (title and abstract) and inclusion (full text) stages by comparing outputs with the existing synopsis of Butterfly and Moth Conservation. Specifically we assessed the ability to correctly include articles that appeared in the synopsis (*recall*) and exclude those that did not (*precision*), where:

$$Recall = \frac{True\ positives}{(True\ positives + False\ negatives)}$$

$$Precision = \frac{True\ positives}{(True\ positives + False\ positives)}$$

In addition to this, we manually assessed the full texts of a random subset of *false positives* (i.e. articles included by the pipeline that do not appear in the synopsis) to check whether any of these articles did in fact meet our inclusion criteria. This manual checking was carried out by six members of the Conservation Evidence team, who all have experience in screening articles for inclusion in the database and authoring Conservation Evidence synopses. We calculated the proportion of the *false positives* that could actually have been included in the synopsis under their existing inclusion criteria (so called "*pseudo false positives*") and used this to make an updated estimate of pipeline *precision*.

Finally, we assessed how repeated article screening (at title and abstract) and inclusion (at full text) by LLMs multiple times affected the rate of *false positives*. We fitted logistic regression models where *false positive* (Y or N) was a binary response variable and the number of inclusions was included as a categorical explanatory variable. We assessed differences between levels of the categorical variable by comparing the estimated marginal means and carrying out post-hoc pairwise comparisons (using Tukey-adjusted p values for multiple comparisons) using the emmeans R package [40]. We fitted separate models for screening at title and abstract stage and inclusion at full text stage, and all models were fitted with a binomial error distribution and logit link function.

4.2 Data extraction

We evaluated the ability of the pipeline to extract data on experimental design, geographical location and habitat type from all included articles that appear in the current synopsis.

4.2.1 Experimental design

Article summaries in the Conservation Evidence database contain a standard set of study design terms that denote whether they include replication, randomisation, pre-impact sampling, comparisons with control sites/individuals or pairing of impact and control sites (Table S3). To compare pipeline classifications with real study designs, we first extracted the true study design terms from each summary paragraph. As the study design is described in the first sentence of the summary paragraph, we retained only those cases where the term appeared within the first 150 characters. For cases where there were multiple summary paragraphs for a single article, we combined all study design terms.

For individual study design terms, we calculated pipeline *recall* and *precision* based on the ability of the pipeline to correctly classify each term. For the full study design, we calculated the accuracy of pipeline predictions, where:

$$Accuracy = \frac{Correct\ predictions}{Total\ predictions}$$

As the pipeline was required to classify the study design for all included articles, *total predictions* was simply the total number of articles assessed.

4.2.2 Geographical location and habitat

The Conservation Evidence database stores data on the geographical location and habitat type in which actions took place. Habitats are classified using a system that is closely aligned to the IUCN habitats classification system Version 3.1 [41]. We assessed the ability of the pipeline to extract the correct country and "level 1" habitat type from all included articles. As the pipeline could predict multiple countries or habitat types per article, we calculated *recall* and *precision* using both micro- and macro-averaging:

$$Micro\ recall = \frac{N_{Correct\ classifications}}{N_{instances}}$$

$$Macro\ recall = \frac{\sum_1^{N_{studies}} recall}{N_{studies}} \qquad \textit{(mean per study recall)}$$

$$Micro\ precision = \frac{N_{Correct\ classifications}}{N_{Classifications}}$$

$$Macro\ precision = \frac{\sum_1^{N_{studies}} precision}{N_{studies}} \qquad \textit{(mean per study precision)}$$

4.3 Evaluation results

We were able to access and download 151,727 article full texts, of which 167 appear in the current synopsis of Butterfly and Moth Conservation.

4.3.1 Article screening and inclusion

Using our stopping rule (when the number of relevant documents found reached 95% of the estimated total number of relevant documents), we took forward 18,238 articles to be screened at title and abstract stage. Of these, 16,106 were rejected and 2,132 were taken forward and screened at the full-text stage. 148 articles were removed due to errors in the pipeline

accessing full-texts. Of the remaining 1,984 articles, 699 were rejected at full-text stage and 1433 were included as relevant articles. A comparison of the relative inclusion rates at different stages are shown in Table 2. The pipeline rejected a substantially lower proportion of articles than in the original synopsis at title and abstract screening stage (88.31% versus 99.96%) and was also relatively more lenient at full-text stage (72.23% versus 82.09%; Table 2).

Table 2. Comparison of the number of articles included and excluded at each stage of the review process for the original manual synopsis review and the pipeline. *148 full texts could not be accessed due to errors. #12 articles were removed as full texts could not be accessed, and one article was a meta-analysis.

	Original synopsis		**Pipeline**			
	All articles		**All articles**		**Articles from original synopsis**	
Screening stage	Number	Percentage included/ excluded	Number	Percentage included/ excluded	Number	Percentage included/ excluded
Total articles for screening	1,186,661	-	151,727	-	167	-
Total articles removed after ranking	NA	-	133,489	87.98	0	0.00
TA total screened	1,186,661	-	18,238	12.02	155#	100.00
TA rejected	1,186,192	99.96	16,106	88.31	0	0.00
TA included	469	0.04	2132	11.69	155	100.00
FT total screened	469	-	1,984*	-	155	-
FT rejected	84	17.91	551	22.77	5	0.00
FT included	385	82.09	1,433	72.23	150	96.77

Within the 151,727 articles screened by the pipeline (a subset of the total corpus), a total of 167 articles (of the 385 articles; Table 2) included in the butterfly and moth synopsis were present and therefore represented the maximum possible recall. All 167 of the articles that appear in the synopsis made it through the initial ranking stage of the pipeline. After removing cases where full text examination returned an error and one meta-analysis (which was excluded under our criteria), 155 articles that appear in the synopsis were screened and assessed for inclusion (title and abstract and full text), and 150 were included, equating to a recall of 96.8% (Table 2).

The pipeline included an additional 1,283 *false positives* (i.e. articles that were not included in the existing synopsis), equating to an initial precision estimate of 10.5%. Randomised manual checking of 140 full texts revealed that 20% (28) of articles were *pseudo false positives*, meaning they did in fact meet the inclusion criteria. Assuming this *pseudo false positive* rate of 20%, our updated estimate of precision was 28.4%.

Articles that were included during all five repetitions of the pipeline were less likely to be *false positives* than those included only three (Title & abstract: Z ratio = 4.83, $p < 0.0001$; Full text: Z ratio = 3.02, $p = 0.007$) or four times (Title & abstract: Z ratio = 5.35, $p < 0.0001$; Full text: Z ratio = 2.77, $p = 0.015$). Articles included in three or four repetitions had a similar chance of being false positives (Title & abstract: Z ratio 0.87, $p = 0.657$; Full text: Z ratio = 0.72, $p = 0.753$; Figure 1).

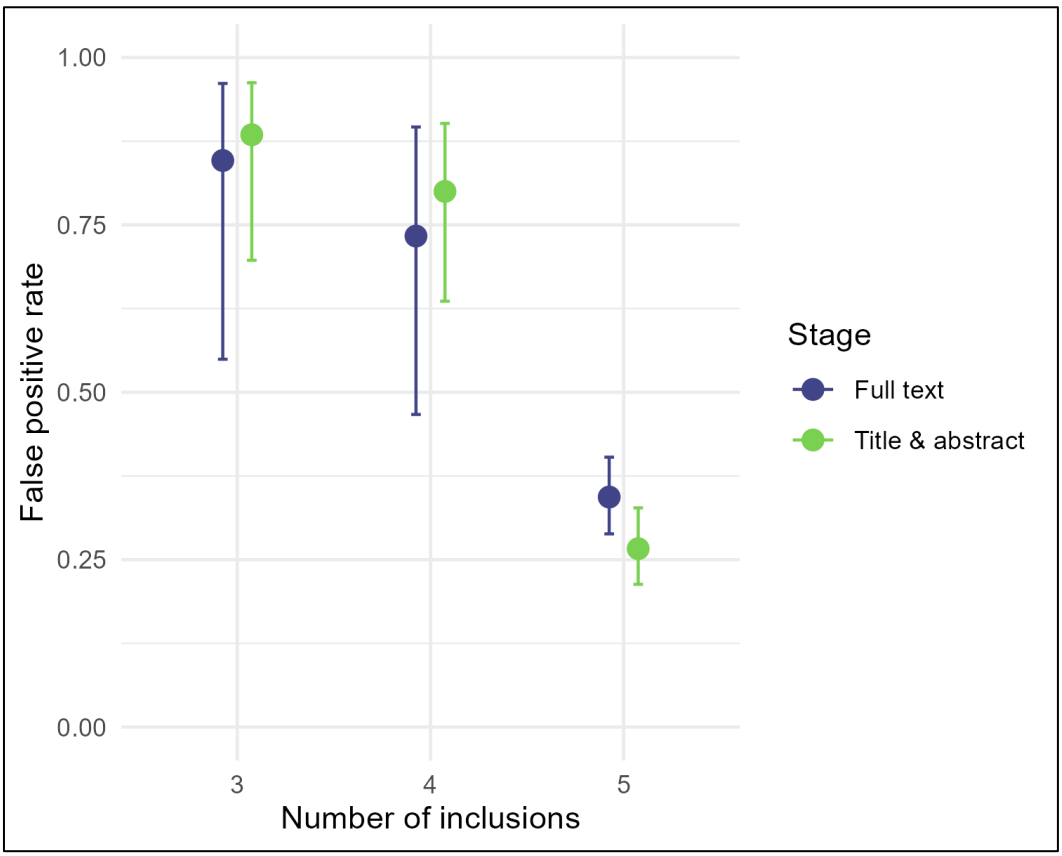

Figure 1. *The probability of included articles being false positives at both the title and abstract screening stage and the full text inclusion stage depending on the number of times they were classified (out of 5 runs of the pipeline)*.

4.3.2 Data extraction

Both *recall* and *precision* were lowest for articles defined as “studies”, which contained no formal study design terms (Table 3). For all other study design terms, *recall* ranged from 0.35 for paired designs to 1.0 for before and after designs, and *precision* ranged from 0.21 for before and after designs to 0.89 for replicated designs (Table 3). For classification of the full study design (i.e. combining all components, of which there are 128 different unique combinations), *accuracy* was 0.12.

Table 3. *Recall* and *precision* for classification of each individual study design component.

Study Design	Recall	Precision
Before and after	1.00	0.21
Controlled	0.95	0.56
Paired	0.35	0.68
Randomized	0.88	0.43
Replicated	0.71	0.89
Site comparison	0.71	0.74
Study	0.25	0.14

When classifying the country in which actions took place, *recall* was >0.99 (for both micro- and macro-averaging), and of 142 countries present in the dataset, 141 were correctly classified. *Precision* for country classification and both *recall* and *precision* for habitat classification suggest moderate to good performance (Table 4).

Table 4. *Recall* and *precision* for classification of country and habitat data, calculated by both micro- and macro-averaging.

Variable	Recall		Precision	
	Micro	***Macro***	***Micro***	***Macro***
Country	0.99	1.00	0.78	0.90
Habitat	0.77	0.85	0.76	0.83

5. Discussion

Successful development and evaluation of a Living Evidence Database pipeline

Our evaluation demonstrates that our end-to-end, self-hosted AI pipeline with humans-in-the-loop can screen articles for inclusion in a Living Evidence Database on a par with human experts, but for a fraction of the cost in terms of time and effort. Our evaluation results also point to better than human expert-level screening performance, given that 20% of false positives were likely to be *pseudo false positives* - articles that were missed by human experts but found by the AI pipeline. These *pseudo false positives* were likely found because the AI pipeline is not susceptible to human error (e.g., caused by fatigue from reviewing large numbers of articles) and the ability of the pipeline to screen a larger number of articles at the full text stage (1,984 articles versus 469; Table 2) by using a more liberal threshold at the title and abstract screening stage than human experts (12% lower rejection rate; inclusion rate was 12% of articles screened versus just 0.04% for the original reviewers; Table 2).

The ability of the pipeline to extract data had more mixed results, but these were encouraging given the future improvements and iterations that can be made. For example, the pipeline performed exceptionally well for more concrete data types such as geographical location and habitat classification. However, performance on nuanced, multi-component data types such as study design was more variable and requires further refinement.

We also successfully piloted a novel component for improving evidence synthesis that is applicable to existing evidence synthesis approaches: Human-AI Criteria Refinement. This was able to translate an evidence synthesis protocol, suitable for human experts, into precise, machine-operable inclusion criteria, which would arguably improve agreement between human reviewers if used in existing review methods.

Our pipeline also implements an automated statistically principled stopping rule through a combination of LLM synthetic data generation, active learning and statistical process modelling. This is in contrast with approaches typically used in evidence synthesis which

necessitate initially screening a random set of articles [42], [43], relying on heuristics [44], [45] or budget-based approaches.

Implications for evidence synthesis at scale to build Living Evidence Databases

Our work provides a proof-of-concept for the technical feasibility of shifting from static, resource-intensive systematic reviews to building dynamic, subject-wide Living Evidence Databases (LEDs) with an AI-human hybrid system. By putting a human-AI collaborative framework at the heart of our pipeline, we avoid creating a 'black box' system that replaces human experts. By shifting the effort of screening vast amounts of literature to AI models, researchers' time can be better harnessed by focusing on higher-level tasks such as verifying edge cases, as well as interpreting and communicating the outputs of LEDs for decision-makers. Our finding that using model self-consistency (multiple runs of parts of the pipeline) can significantly improve precision has key practical implications, offering a tunable mechanism to manage the trade-off between automation and human verification workloads. Synthesis teams could decide on a confidence threshold (e.g., whether to only review articles included in more than 4 or 5 runs) to prioritise effort. Our work also shows that an end-to-end AI pipeline for evidence synthesis can be traceable and reproducible, whilst being self-hosted means that issues of data sovereignty and privacy can be managed more easily. Such considerations are extremely important to the future of large-scale synthesis that leverages AI models as these issues can often be barriers to adoption in different research and policy environments.

Our work also highlights the critical importance of effective integrations with publishers, without which this pipeline could not feasibly function. By providing automated access to the full-text of articles via APIs or bulk download facilities, the system can prioritise recall at the abstract screening stage. This enhances the system's comprehensiveness, as it results in significantly more full-text screening than would be feasible in a manual review. Crucially, this process uncovers articles that would have been erroneously excluded based on their abstract in a conventional review.

Future directions

Our proof-of-concept pipeline may have achieved impressive levels of performance, particularly at the screening stages, but still requires further refinement. Currently, the primary limitation is that there is typically low initial precision, which would require significant human verification efforts to screen (despite reducing the size of the corpus needing to be screened substantially, in this case by 97%). More sophisticated, multi-stage filtering models and classifiers could help to reduce this problem.

The pipeline also struggled at complex data extraction, for example for study design classification. Future work could explore how to fine-tune smaller, specialised models to deal with these specific tasks, whilst more advanced prompting techniques and extraction schema might also improve pipeline performance.

We must also acknowledge that our current evaluation was based only on a single synopsis of conservation intervention studies and so the results are potentially not generalisable to other subject areas. Further research is needed to assess this generalisability across diverse topic areas, although the design of the pipeline is deliberately interoperable to enable such work.

Future work will also aim to expand on the pipeline's capabilities by integrating automated critical appraisal and risk-of-bias assessments, as well as processing non-English language studies and grey literature.

6. Conclusion

This paper details the architecture and successful evaluation of a self-hosted, end-to-end AI pipeline designed to power LEDs. Our results demonstrate that this approach can achieve very high recall in terms of finding relevant papers - even finding additional relevant papers due to being able to screen a larger number of papers at full-text stage. While challenges in precision and complex data extraction remain, we show that practical solutions, such as leveraging model self-consistency and human-AI collaborative workflows, can mitigate these issues. Whilst further refinements are needed, our work provides a viable technical foundation for the transition from static, single-use reviews to dynamic, sustainable, and transparent living evidence ecosystems, fundamentally changing how evidence is synthesized and used to inform policy and practice in conservation.

References

[1] S. J. Cooke *et al.*, 'Environmental evidence in action: on the science and practice of evidence synthesis and evidence-based decision-making', *Environ. Evid.*, vol. 12, no. 1, p. 10, May 2023, doi: 10.1186/s13750-023-00302-5.

[2] D. C. Rose *et al.*, 'Policy windows for the environment: Tips for improving the uptake of scientific knowledge', *Environ. Sci. Policy*, vol. 113, pp. 47–54, Nov. 2020, doi: 10.1016/j.envsci.2017.07.013.

[3] C. Knill and Y. Steinebach, 'What has happened and what has not happened due to the coronavirus disease pandemic: a systemic perspective on policy change', *Policy Soc.*, vol. 41, no. 1, pp. 25–39, Mar. 2022, doi: 10.1093/polsoc/puab008.

[4] L. Beresford, R. Walker, and L. Stewart, 'Extent and nature of duplication in PROSPERO using COVID-19-related registrations: a retrospective investigation and survey', *BMJ Open*, vol. 12, no. 12, p. e061862, Dec. 2022, doi: 10.1136/bmjopen-2022-061862.

[5] K. G. Shojania, M. Sampson, M. T. Ansari, J. Ji, S. Doucette, and D. Moher, 'How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis', *Ann. Intern. Med.*, vol. 147, no. 4, pp. 224–233, Aug. 2007, doi: 10.7326/0003-4819-147-4-200708210-00179.

[6] M. A. Hanson, P. G. Barreiro, P. Crosetto, and D. Brockington, 'The strain on scientific publishing', *Quant. Sci. Stud.*, vol. 5, no. 4, pp. 823–843, Nov. 2024, doi: 10.1162/qss_a_00327.

[7] S. A. Reynolds *et al.*, 'Will AI speed up literature reviews or derail them entirely?', *Nature*, vol. 643, no. 8071, pp. 329–331, July 2025, doi: 10.1038/d41586-025-02069-w.

[8] K. C. Siontis and J. P. A. Ioannidis, 'Replication, Duplication, and Waste in a Quarter Million Systematic Reviews and Meta-Analyses', *Circ. Cardiovasc. Qual. Outcomes*, vol. 11, no. 12, p. e005212, Dec. 2018, doi: 10.1161/CIRCOUTCOMES.118.005212.

[9] J. H. Elliott *et al.*, 'Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap', *PLOS Med.*, vol. 11, no. 2, p. e1001603, Feb. 2014, doi: 10.1371/journal.pmed.1001603.

[10] Q. Zheng *et al.*, 'Past, present and future of living systematic review: a bibliometrics analysis', *BMJ Glob. Health*, vol. 7, no. 10, Oct. 2022, doi: 10.1136/bmjgh-2022-009378.

[11] W. J. Sutherland and C. F. R. Wordley, ‘A fresh approach to evidence synthesis’, *Nature*, vol. 558, no. 7710, pp. 364–366, June 2018, doi: 10.1038/d41586-018-05472-8.
[12] W. J. Sutherland *et al.*, ‘Building a tool to overcome barriers in research-implementation spaces: The Conservation Evidence database’, *Biol. Conserv.*, vol. 238, p. 108199, Oct. 2019, doi: 10.1016/j.biocon.2019.108199.
[13] H. Bastian, P. Glasziou, and I. Chalmers, ‘Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up?’, *PLOS Med.*, vol. 7, no. 9, p. e1000326, Sept. 2010, doi: 10.1371/journal.pmed.1000326.
[14] E. M. Grames, A. N. Stillman, M. W. Tingley, and C. S. Elphick, ‘An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks’, *Methods Ecol. Evol.*, vol. 10, no. 10, pp. 1645–1654, 2019, doi: 10.1111/2041-210X.13268.
[15] A. D. Cassai *et al.*, ‘Evaluating the utility of large language models in generating search strings for systematic reviews in anesthesiology: a comparative analysis of top-ranked journals’, *Reg. Anesth. Pain Med.*, Jan. 2025, doi: 10.1136/rapm-2024-106231.
[16] K. E. K. Chai, R. L. J. Lines, D. F. Gucciardi, and L. Ng, ‘Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews’, *Syst. Rev.*, vol. 10, no. 1, p. 93, Apr. 2021, doi: 10.1186/s13643-021-01635-3.
[17] R. van de Schoot *et al.*, ‘An open source machine learning framework for efficient and transparent systematic reviews’, *Nat. Mach. Intell.*, vol. 3, no. 2, pp. 125–133, Feb. 2021, doi: 10.1038/s42256-020-00287-7.
[18] ‘abstrackr’. Accessed: Sept. 15, 2025. [Online]. Available: https://abstrackr.com/
[19] ‘EPPI-Reviewer: introduction’. Accessed: Sept. 15, 2025. [Online]. Available: https://eppi.ioe.ac.uk/cms/default.aspx?tabid=1913
[20] S. R. Jonnalagadda, P. Goyal, and M. D. Huffman, ‘Automating data extraction in systematic reviews: a systematic review’, *Syst. Rev.*, vol. 4, no. 1, p. 78, June 2015, doi: 10.1186/s13643-015-0066-7.
[21] I. J. Marshall, J. Kuiper, and B. C. Wallace, ‘RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials’, *J. Am. Med. Inform. Assoc.*, vol. 23, no. 1, pp. 193–201, Jan. 2016, doi: 10.1093/jamia/ocv044.
[22] A. Konet *et al.*, ‘Performance of two large language models for data extraction in evidence synthesis’, *Res. Synth. Methods*, vol. 15, no. 5, pp. 818–824, 2024, doi: 10.1002/jrsm.1732.
[23] M. A. Khan *et al.*, ‘Collaborative large language models for automated data extraction in living systematic reviews’, *J. Am. Med. Inform. Assoc.*, vol. 32, no. 4, pp. 638–647, Apr. 2025, doi: 10.1093/jamia/ocae325.
[24] ‘doi Foundation’, doi Foundation. Accessed: Sept. 29, 2025. [Online]. Available: https://www.doi.org
[25] J. Priem, H. Piwowar, and R. Orr, ‘OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts’, June 17, 2022, *arXiv*: arXiv:2205.01833. doi: 10.48550/arXiv.2205.01833.
[26] R. Kinney *et al.*, ‘The Semantic Scholar Open Data Platform’, Apr. 25, 2025, *arXiv*: arXiv:2301.10140. doi: 10.48550/arXiv.2301.10140.
[27] P. Lopez, *kermitt2/grobid*. (Sept. 16, 2025). Java. Accessed: Sept. 16, 2025. [Online]. Available: https://github.com/kermitt2/grobid
[28] P. Lopez, *kermitt2/Pub2TEI*. (July 28, 2025). XSLT. Accessed: Sept. 16, 2025. [Online]. Available: https://github.com/kermitt2/Pub2TEI
[29] A. Kusupati *et al.*, ‘Matryoshka Representation Learning’, in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds, Curran Associates, Inc., 2022, pp. 30233–30249. [Online]. Available:

https://proceedings.neurips.cc/paper_files/paper/2022/file/c32319f4868da7613d78af9993100e42-Paper-Conference.pdf
[30] 'nomic-ai/modernbert-embed-base · Hugging Face'. Accessed: Sept. 16, 2025. [Online]. Available: https://huggingface.co/nomic-ai/modernbert-embed-base
[31] DeepSeek-AI *et al.*, 'DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning', Jan. 22, 2025, *arXiv*: arXiv:2501.12948. doi: 10.48550/arXiv.2501.12948.
[32] X. Wang *et al.*, 'Self-Consistency Improves Chain of Thought Reasoning in Language Models', presented at the The Eleventh International Conference on Learning Representations, Sept. 2022. Accessed: Sept. 15, 2025. [Online]. Available: https://openreview.net/forum?id=1PL1NIMMrw
[33] A. Sneyd and M. Stevenson, 'Modelling Stopping Criteria for Search Results using Poisson Processes', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3484–3489. doi: 10.18653/v1/D19-1351.
[34] M. Stevenson and R. Bin-Hezam, 'Stopping Methods for Technology-assisted Reviews Based on Point Processes', *ACM Trans Inf Syst*, vol. 42, no. 3, p. 73:1-73:37, Dec. 2023, doi: 10.1145/3631990.
[35] M. W. Callaghan and F. Müller-Hansen, 'Statistical stopping criteria for automated screening in systematic reviews', *Syst. Rev.*, vol. 9, no. 1, p. 273, Dec. 2020, doi: 10.1186/s13643-020-01521-4.
[36] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, 'Reducing Workload in Systematic Review Preparation Using Automated Citation Classification', *J. Am. Med. Inform. Assoc.*, vol. 13, no. 2, pp. 206–219, Mar. 2006, doi: 10.1197/jamia.M1929.
[37] 'Butterfly and Moth Conservation: Global evidence for the effects of interventions for butterflies and moths'. Accessed: Sept. 15, 2025. [Online]. Available: https://www.repository.cam.ac.uk/items/74c1bc51-9072-4f09-8d37-bc9a3869f458
[38] 'Conservation Evidence - Site', Conservation Evidence. Accessed: Sept. 15, 2025. [Online]. Available: https://conservationevidence.com/
[39] A. Bladon, E. Bladon, R. Smith, and W. Sutherland, *Butterfly and Moth Conservation: Global evidence for the effects of interventions for butterflies and moths*. Conservation Evidence, 2022. doi: 10.17863/CAM.105954.
[40] 'Estimated Marginal Means, aka Least-Squares Means • emmeans'. Accessed: Sept. 15, 2025. [Online]. Available: https://rvlenth.github.io/emmeans/
[41] 'The IUCN Red List of Threatened Species', IUCN Red List of Threatened Species. Accessed: Sept. 15, 2025. [Online]. Available: https://www.iucnredlist.org/en
[42] M. Kastner, S. E. Straus, K. A. McKibbon, and C. H. Goldsmith, 'The capture–mark–recapture technique can be used as a stopping rule when searching in systematic reviews', *J. Clin. Epidemiol.*, vol. 62, no. 2, pp. 149–157, Feb. 2009, doi: 10.1016/j.jclinepi.2008.06.001.
[43] J. Boetje and R. van de Schoot, 'The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses', *Syst. Rev.*, vol. 13, no. 1, p. 81, Mar. 2024, doi: 10.1186/s13643-024-02502-7.
[44] G. V. Cormack and M. R. Grossman, 'Engineering Quality and Reliability in Technology-Assisted Review', in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, in SIGIR '16. New York, NY, USA: Association for Computing Machinery, July 2016, pp. 75–84. doi: 10.1145/2911451.2911510.
[45] R. Ros, E. Bjarnason, and P. Runeson, 'A Machine Learning Approach for Semi-Automated Search and Selection in Literature Studies', in *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, in EASE '17. New York, NY, USA: Association for Computing Machinery, June 2017, pp. 118–127. doi: 10.1145/3084226.3084243.