Conservation Biology

New opportunities and challenges for conservation evidence synthesis from advances in natural language processing

Charlotte H. Chang[1] | **Susan C. Cook-Patton**[2] | **James T. Erbaugh**[3] | **Luci Lu**[4] |
Yuta J. Masuda[5] | **István Molnár**[6] | **Dávid Papp**[6,7] | **Brian E. Robinson**[8]

[1]Department of Biology and Environmental Analysis Program, Pomona College, Claremont, California, USA

[2]The Nature Conservancy, Arlington, Virginia, USA

[3]Department of Environmental Studies, Dartmouth College, Hanover, New Hampshire, USA

[4]Jornada Experimental Range, New Mexico State University, Las Cruces, New Mexico, USA

[5]Paul G. Allen Family Foundation, Seattle, Washington, USA

[6]Lexunit Zrt, Budapest, Hungary

[7]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

[8]Department of Geography, McGill University, Montreal, Quebec, Canada

Correspondence
Charlotte H. Chang, Department of Biology and Environmental Analysis Program, Pomona College, 175 W. Sixth Street, Claremont, CA 91711, USA.
Email: chchang@pomona.edu

Article impact statement: Machine learning can improve the speed and robustness of evidence synthesis across multidisciplinary conservation data sets.

Funding information
Nature Conservancy; Bezos Earth Fund

Abstract

Addressing global environmental conservation problems requires rapidly translating natural and conservation social science evidence to policy-relevant information. Yet, exponential increases in scientific production combined with disciplinary differences in reporting research make interdisciplinary evidence syntheses especially challenging. Ongoing developments in natural language processing (NLP), such as large language models, machine learning (ML), and data mining, hold the promise of accelerating cross-disciplinary evidence syntheses and primary research. The evolution of ML, NLP, and artificial intelligence (AI) systems in computational science research provides new approaches to accelerate all stages of evidence synthesis in conservation social science. To show how ML, language processing, and AI can help automate and scale evidence syntheses in conservation social science, we describe methods that can automate querying the literature, process large and unstructured bodies of textual evidence, and extract parameters of interest from scientific studies. Automation can translate to other research agendas in conservation social science by categorizing and labeling data at scale, yet there are major unanswered questions about how to use hybrid AI-expert systems ethically and effectively in conservation.

KEYWORDS
conservation social science, evidence synthesis, language models, machine learning, natural language processing

INTRODUCTION

The world is facing extreme challenges, including human-driven biodiversity loss (Díaz & Malhi, 2022), climate change (IPCC, 2018), and rising human needs for ecosystem services (Dasgupta, 2021; Pörtner et al., 2021). Addressing these issues will require integrating the natural and social sciences to find solutions to these multifaceted environmental and sociopolitical problems (Bennett et al., 2017; Dasgupta, 2021; Pörtner et al., 2023). Evidence synthesis is a crucial tool for bringing forth and compiling the best evidence across multiple fields, and several recent efforts have showcased how such syntheses can translate

research into evidence-based policy and practice (e.g., Cheng et al., 2019; McKinnon et al., 2016; Rosenstock et al., 2019).

Despite the importance of integrating scientific disciplines to identify and develop solutions to conservation challenges, several barriers prevent pooling diverse literatures for interdisciplinary understanding: the wide range of literature, inefficiencies related to text summarization, and errors or biases in data evaluation and coding. First, scientific evidence is dispersed across a range of peer-reviewed outlets and disciplines. To review this dispersed evidence base, traditional synthesis methods require enormous logistical, financial, and community-wide efforts. Second, numerous disciplines with diverse methodologies investigate relationships between people and nature. Ensuring consistency in coding evidence between people (interrater reliability) and over time (intrarater reliability), as well as addressing issues such as reviewer fatigue, can be a major challenge (DiMaggio, 2015; O'Mara-Eves et al., 2015). Text data—from manuscripts or digital platforms—may be difficult to evaluate efficiently and consistently, particularly at large scales (Edelmann et al., 2020; Gentzkow et al., 2019). Finally, as the number of relevant articles increases over time (Callaghan et al., 2021; Sietsma et al., 2024; Thomas et al., 2017), the time and effort needed to conduct an evidence synthesis also increase. For example, a global synthesis effort focused on natural forest regrowth took 3 years and hundreds of hours of manual labor (Cook-Patton et al., 2020). When published, the underlying database representing the most current synthesis was 3 years out of date and only covered priority areas for carbon accumulation.

Several recent advances in machine learning (ML) and natural language processing (NLP) provide new opportunities for synthesizing conservation social science and natural science. For example, conservation social scientists have applied NLP to news, social, and other media to examine public views of biodiversity or assess polarization in climate change discourse (Chang et al., 2022a, 2022b; Correia et al., 2021; Falkenberg et al., 2022; Giebink et al., 2024). These methods have utility in evidence synthesis (Farrell et al., 2022, 2024; Sietsma et al., 2024; Westgate et al., 2015), but potential users still struggle to understand the strengths and limitations of these tools (e.g., language bias, reliability). We sought to provide an overview of the emergence, spread, and advancement of ML and NLP in evidence synthesis. We also considered how such tools have become central to rapid and rigorous reviews.

OVERVIEW OF EVIDENCE SYNTHESIS

Systematic evidence syntheses often include developing search strings, screening and deduplicating search results, and coding for key variables of interest (Figure 1) (Grant & Booth, 2009; Khan et al., 2003; Lunny et al., 2017). Research teams focus on published research (e.g., Berrang-Ford, Siders, et al., 2021; Callaghan et al., 2021; McKinnon et al., 2016) or additionally include unpublished data (e.g., IPBES, 2018; Porciello et al., 2020). Systematic reviews of peer-reviewed publications are more common due to their prominence, replicability, and relatively complete indexing (Foo et al., 2021; Gusenbauer & Haddaway, 2020; van Driel et al., 2009).

ML or artificial intelligence (AI) tools can accelerate or automate every step in evidence synthesis (Figure 1; Tables 1 & 2), offering value when manual review and synthesis are insufficient and intractable. Even with teams of hundreds of experts collectively contributing thousands of person years of review effort, global efforts relying exclusively on manual review could only process 14,000 (IPCC, 2021) to 15,000 publications (IPBES, 2019). Arguably, the scale and state of global conservation problems now require much more rapid evidence synthesis. Aligning decision-making with comprehensive, up-to-date information requires leveraging ML to accelerate identifying the best available scientific evidence base. ML- or AI-assisted approaches could reduce bias and increase reproducibility relative to teams of human coders. Large language models (LLMs) hold particularly high promise. For instance, Callaghan et al. (2021) trained a relevance classifier on 2000 abstracts to predict whether >600,000 abstracts contained information on climate impacts. We outlined advances to date along each step of an evidence synthesis effort (Figure 1) that enable more rapid and robust data integration.

ACCELERATING THE IDENTIFICATION OF RELEVANT ARTICLES

Developing search strings for academic databases or search engines is a foundational step for evidence syntheses and can itself be an involved process (Cooke et al., 2012; Villanueva et al., 2001). Keyword searches are often designed to capture the universe of potentially relevant publications and so may return orders of magnitude more articles than are actually relevant. Researchers may adjust keywords to achieve a manageable number of results, which can exclude relevant literature from review. New methods leverage NLP and network science to automate this process, producing more refined search strings (Grames et al., 2019; Kwabena et al., 2023; Figure 1; Table 1). Researchers then compile and screen articles for relevance and duplication. ML algorithms have assisted with screening for nearly a decade (O'Mara et al., 2015). Platforms such as abstrackr (Gates et al., 2018; Wallace et al., 2012), metagear (Lajeunesse, 2016), and colandr (Cheng et al., 2018) use human coding of a subset of abstracts and keywords to probabilistically evaluate the relevance of additional abstracts. These platforms use NLP to identify sentences or word clusters common among articles deemed to be relevant and assess if unscreened articles contain similar text (Cheng et al., 2018). As more articles are screened, the algorithm's accuracy improves (Tsou et al., 2020). Such a human-in-the-loop process, where ML assists with article screening and coding that are checked by experts for validity, can balance cost efficiency and timeliness with consistency (O'Connor et al., 2019; Thomas et al., 2017). Such approaches have been used in global evidence synthesis projects.

For example, Berrang-Ford, Siders, et al. (2021) used NLP to screen 48,000 articles and an expert team of 126 researchers who collectively coded 1682 articles for evidence on climate adaptation. However, in Berrang-Ford, Siders, et al.'s (2021) review, attributes such as the regional focus of each included study ($n = 2032$) were manually extracted by teams of experts.
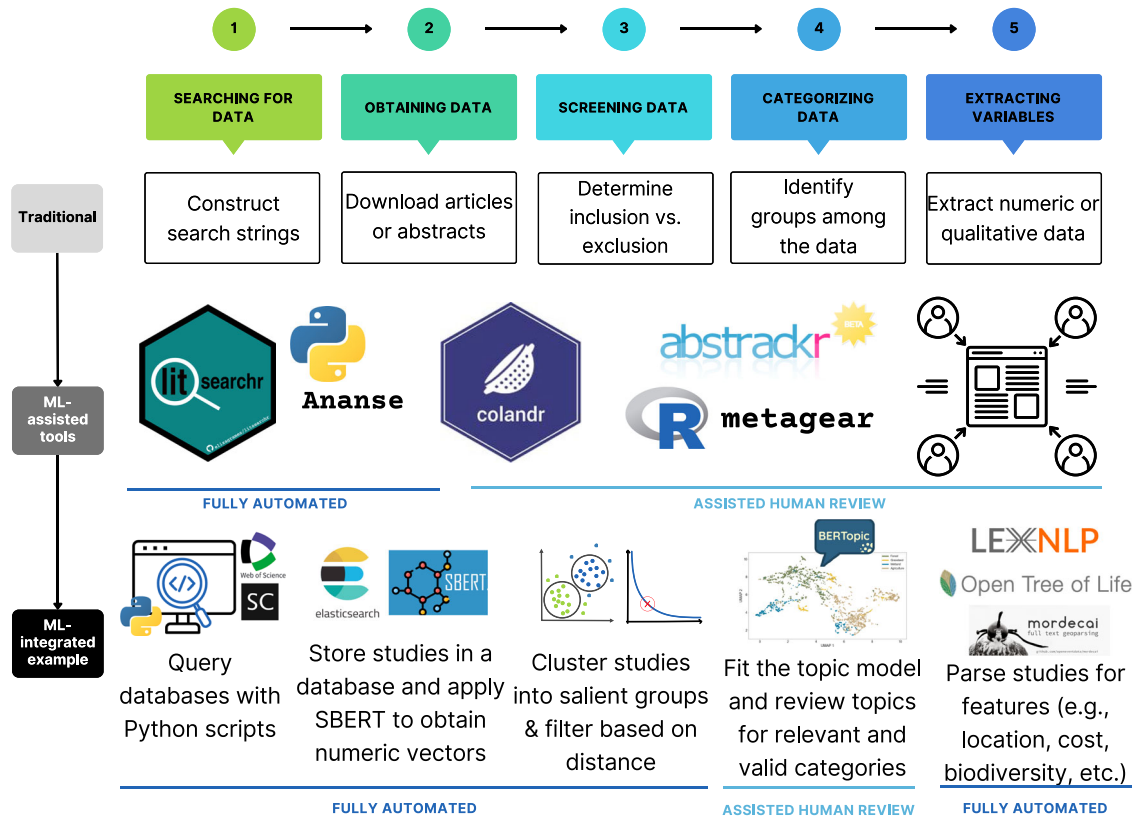
FIGURE 1 Traditional evidence synthesis tasks, examples of machine learning (ML)-assisted tools (Table 1), and how ML and natural language processing (NLP) can be more deeply integrated throughout an evidence synthesis process (Table 2).

TABLE 1 Descriptions of different software packages that can support automated evidence syntheses in conservation social science.

Category	Tool	Purpose	Citation
Machine-learning-assisted tools	litsearchR	Determine search terms based on text mining and keyword co-occurrence	Grames et al., 2019
	Ananse	Implementation of litsearchR in Python	Kwabena et al., 2023
	colandr	Semiautomated, human-in-the-loop platform to screen abstracts for relevance	Cheng et al., 2018
	abstrackr	Semiautomated platform to screen abstracts for relevance	Gates et al., 2018; Wallace et al., 2012
	metagear	Tools to help teams of reviewers screen and process abstracts for relevance and other objectives	Lajeunesse, 2016
Machine-learning-integrated example	SBERT	Access and use transformer language models	Devlin et al., 2019; Reimers & Guryevich, 2019
	BERTopic	Perform topic modeling with transformer model input (contextual embeddings)	Grootendorst, 2022
	LexNLP	Structured information extraction for legal and financial documents	Bommarito et al., 2021
	Open Tree of Life	Comprehensive taxonomic information across the tree of life	Rees & Cranston, 2017; OpenTreeofLife et al., 2021
	Mordecai	Detect locations in text data	Halterman, 2017

Evidence syntheses featuring greater automation have used ML and AI systems to go beyond predicting article relevance to predicting relevant categories or features in and across studies. For example, researchers have predicted labels for individual stud-ies focused on global development evidence for agriculture and food security (Edwards et al., 2024), climate change impacts (Callaghan et al., 2021), or natural climate solutions impacts to human and environmental well-being (Chang et al., 2025).

TABLE 2 Resources for learning more about text analysis methods or replication code accompanying machine learning integrated syntheses.

Source	Focus	Link	Type
Chang et al., 2025	Natural climate solutions well-being impacts	https://doi.org/10.5281/zenodo.14206056	Code repository
Jackson et al., 2022	Overview of text analysis for psychology	https://osf.io/hvcg3/	Code repository
Callaghan et al., 2021	Climate change impacts	https://doi.org/10.5281/ZENODO.5327409	Code repository
Berrang-Ford, Sietsma, et al., 2021	Connections between climate and health	https://zenodo.org/records/4322697	Code repository

AUTOMATING DATA QUERIES IN EVIDENCE SYNTHESIS

AI systems can automate additional steps in evidence synthesis workflows, such as querying and downloading studies (e.g., Berrang-Ford, Siders, et al., 2021; Callaghan et al., 2021; Edwards et al., 2024; Sietsma et al., 2024). For example, scripts and application programming interfaces (APIs) have been used to query and retrieve scientific abstracts and other relevant publication information from Scopus and the Web of Science in a more automated fashion (Callaghan et al., 2021; Chang et al., 2025). In addition to saving time, this approach is more transparent and replicable, which avoids errors that can emerge in manual data querying (e.g., data entry errors, deviations from protocol). Further, manual querying may not be feasible for some syntheses. Our process culminated in over 2 million abstracts—a volume that would be difficult, and perhaps impossible, to manually download.

There are some advantages of manual data curation that are not typically integrated into ML-assisted workflows. For instance, experts may be interested in expanding their evidence base by following a citation trail from initially selected studies. However, automated querying approaches can emulate these tasks through careful design. For example, one might obtain citation networks from scholarly databases and use supervised classification to predict which abstracts may correspond to areas that would benefit from additional samples (Ammar et al., 2018; Priem et al., 2022).

Moving beyond abstract-level data to automatically accessing the full text of scientific articles is at the cutting edge and has been facilitated by open science mandates and platforms, such as PubMed Central (Farrell et al., 2022). Open-source tools can extract text, tables, and figures from PDFs or parse the full text of articles delivered in structured formats such as HTML (Liu & McKie, 2024; Yang et al., 2019). To date, however, we are not aware of examples of extensive automation of such full-text features to accelerate evidence synthesis in conservation. The diversity of disciplines, methodologies, and results (e.g., tables and figures presenting statistical analyses, qualitative analyses and results, etc.) presents a unique challenge for adopting these methods in conservation and beyond.

EVOLUTION OF TEXT ANALYSIS TIME TO PROCESS BODIES OF EVIDENCE

Key to any evidence synthesis project is acquiring information and extracting data from relevant articles, such as cataloging themes or topics across studies. We considered how advances in text analysis approaches through time have enabled different types of analyses. With manual reviews as a starting point, data extraction followed a coding template to record information consistently (Grant & Booth, 2009; Khan et al., 2003; Lunny et al., 2017). Research in the social sciences and conservation social science shows how manual coding can be supplanted or augmented by NLP (Ash & Hansen, 2023; Grimmer et al., 2022). In some of the earliest advances in computational social science, relatively simple statistical NLP methods, such as frequency-based n-grams that identify a sequence of adjacent terms, were used. Although simple, these earliest NLP methods offered major new insights. For example, Michel et al. (2011) provided a watershed advance for quantitatively researching human culture by assessing n-gram patterns across time in the Google Books corpus. Anderson et al. (2021) used a similar n-gram approach to document shifts in ecological research across 3 decades.

Other researchers have used probabilistic, unsupervised language models, most notably topic models (Blei, 2012; Boyd-Graber et al., 2017; Grimmer et al. 2022). A topic model is an unsupervised clustering model that discovers themes across documents based on the assumptions that documents are a mixture of themes (or topics) and that different topics tend to be associated with different terms—for instance, zoonotic disease is described with terms distinct from those for habitat loss. In conservation social science, topic modeling has been used to track trends in climate disinformation (Farrell, 2016) or environmental discourse on social media (Chang et al., 2022a, 2022b). NLP in general and topic modeling in particular have been used relatively rarely in evidence syntheses. However, in a few literature reviews, topic models were used to document shifts in conservation science and common pool resource research (Lambert et al., 2021; Westgate et al., 2015). Topic models are appropriate when researchers do not have strong prior knowledge on which categories clearly exist in a data set. Berrang-Ford, Sietsma, et al. (2021) used topic modeling to induct categories in a global evidence review focused on climate hazards and public health.

The outputs of probabilistic language models or simpler keyword- or n-gram-based outputs can be further augmented. For example, researchers could use sentiment analysis to quantify valence (e.g., positive to negative labels), emotions (e.g., fear vs. disgust vs. hope), or opinions in text (Mohammad, 2016). Van Houtan et al. (2020) used a sentiment analysis of keywords in species reintroduction studies to examine drivers of successes versus failures. These earlier studies processed text data with relatively simple bag-of-words vectorization, where each span of

text would be converted to a list of distinct words. Such text vectorization, however, loses the rich meaning conveyed in the order of words and the semantic structure of human language.

NLP can extract additional features in scientific articles or text data, such as study locations, species studied, or financial measures (Farrell et al., 2022; Figure 1). Named entity recognition (NER) is an NLP task that evaluates how well models can detect and correctly classify entities. One example of an NER task is differentiating Lima, Peru, as a place name from lima bean as a food item. Detecting location from text data often relies on NER to identify place names, which are then resolved to geographic coordinates (Halterman, 2017).

Recent advances in NLP, such as transformer language models, are at the vanguard of evidence synthesis and are particularly relevant for conservation social science due to their ability to distill insights from much larger bodies of evidence while capturing more nuanced meaning and context than bag-of-words approaches (Ash & Hansen, 2023; Burnham, 2024). Transformer language models are deep neural network models designed to capture associations between spans of text and predict which portions of the text are important (Vaswani et al., 2017). As a result, transformer language models, such as Google's Bidirectional Encoder Representations from Transformers model (BERT) (Devlin et al., 2019), can better represent subtle variations in meaning in different pieces of text. The BERT model has been used as an input for supervised models to categorize evidence for climate impacts and adaptation (Callaghan et al., 2021) (Table 2), food systems (Porciello et al., 2020), and global development and food security (Edwards et al., 2024).

Transformer models, such as BERT, can also be used in unsupervised topic models. For example, Chang et al. (2025) aimed to discover what combinations of natural climate solutions (NCS) pathways and impacts to human and environmental well-being existed in a large body of scientific evidence (Table 2). NCS refers to a set of preexisting management practices grouped in new categories. For instance, extended timber rotation is one practice contained in the improved forest management NCS pathway (Griscom et al., 2017). Therefore, Chang et al. (2025) were uncertain which pathway and impact combinations would even exist as valid categories. A topic modeling procedure that included outputs from the BERT model provided results that were robustly sampled across a large set of abstracts. These results were then mapped to combinations of NCS pathways and impacts.

Review teams have consistently reported significant time savings from ML-assisted or automated evidence synthesis. Simple abstract screening tools have reduced workload by 35–99% at various stages (Gates et al., 2019). Edwards et al. (2024) found that their automated system cut human screening effort by 55%. Chang et al. (2025) reported that manually reviewing their initial sample of over 2 million abstracts would have been infeasible, even with a large expert team dedicating thousands of hours. Instead, by reviewing the outputs of the BERT topic model (Figure 1; Table 1), which included, for example, highly relevant abstracts for each cluster, their core team of 5 data coders managed the task in just tens of hours per person (Chang et al., 2025).

However, ML-supported or automated systematic reviews and meta-analyses also introduce challenges. These challenges include selecting ML models (Marshall & Wallace, 2019; Thomas et al., 2017) and ensuring that ML-assisted findings are valid (Qureshi et al., 2023). Additionally, many ML algorithms require practitioners to make choices about hyperparameters or model settings that are not directly estimated from data. For instance, in a categorical prediction model that always returns 5 categories, the number of categories output is itself a hyperparameter. As an example in the context of topic modeling, Farrell (2016) had to determine the number of topics for categorizing climate disinformation generated by fossil fuel companies. Similarly, to categorize environmental discourse with a topic model, Chang et al. (2022b) had to consider hyperparameters such as the numbers of topics and ways to pool social media posts.

In Chang et al. (2025), one critical hyperparameter was the number of clusters used to categorize NCS co-impacts. That study used Bayesian optimization to identify the hyperparameter combinations that maximized the coherence of the resulting topics (Chang et al., 2025; Snoek et al., 2012). Coherence has been a standard way to examine the parsimony and fit of topic model results (Mimno et al., 2011), and a more coherent model will have topics that are highly distinct from other topics (Grimmer et al., 2022). Chang et al. (2025) noted that Bayesian optimization was necessary because, given the 5.6 million possible hyperparameter combinations, one would otherwise require almost 320 years to perform a grid search over all possible combinations. Bayesian hyperparameter optimization methods balance feasibility with rigor. Hyperparameters and other model design choices are also relevant for supervised classification models. For example, practitioners must make choices ranging from the division between training and test data to the minimum number of examples that can constitute a valid category (Grimmer et al., 2022; Sietsma et al., 2024).

Free resources are available for practitioners seeking to learn more. Some literature reviews offer an overview of ML and AI integrated methods in social science (e.g., Jackson et al., 2022) (Table 2), and there are NLP courses and tutorials online (e.g., Climate Change AI, Summer Institutes in Computational Social Science, and the open-source platform Hugging Face). However, the field is changing rapidly, and data sources or code to implement certain analyses can become deprecated in unpredictable ways.

FUTURE DIRECTIONS FOR NLP, ML, AND AI IN CONSERVATION SOCIAL SCIENCE

Generative pretrained transformer (GPT) LLMs offer new possibilities for conservation social science evidence synthesis because such models can excel at identifying relatively nuanced, qualitative attributes from text corpora (Farrell et al., 2024). Such LLMs, also known as foundation models, because they have been trained on a broad range of data, have shown state-of-the-art performance across a wide range of downstream uses (Kojima et al., 2022). Henceforth, we use the term *LLM* to refer to GPT types of models. An LLM can automate tasks such as content analysis (Chew et al., 2023), a common task in

qualitative studies. Scheepens et al. (2024) demonstrate the use of OpenAI's GPT-4 LLM to automatically extract hierarchical and taxonomic information on pests and pest controllers, such as natural enemies or parasitoids, from the ecological literature.

Research in the social sciences provides new examples, seemingly by the week, of how LLMs can augment the efforts of human coders in data labeling. LLMs have shown high levels of performance for categorizing political opinions (Burnham, 2024); psychological or political science labels, such as disinformation (Ziems et al., 2024); and affective and cognitive attributes in text, such as implicit bias (Demszky et al., 2023). Moreover, some LLMs, particularly closed-source LLMs, can generalize across languages beyond English. Smaller, open-source LLMs are also increasingly being trained on multilingual data sets, offering exciting possibilities for examining non-English language evidence with open-source models (Grattafiori et al., 2024). Some LLMs may outperform human coders for NLP tasks, such as sentiment analysis or moral values coding, even for languages with little to no labeled training data (e.g., Turkish, Kinyarwanda, or Bahasa Indonesia) (Rathje et al., 2024). Most remarkable is that these LLMs can achieve best-in-class outcomes without any custom training data. In some cases, even with zero-shot predictions, LLMs have surpassed specialized deep learning models by a substantial margin. Taranukhin et al. (2024) documented that zero-shot LLM climate opinion predictions have an F_1 score (the harmonic mean of precision and recall) 28.4 points higher than previous, specialized models.

Use of LLMs to perform qualitative data coding in conservation science is still relatively limited (Farrell et al., 2024). Yet, the power and performance of these models across a variety of domains offer exciting possibilities for conservation social science research and synthesis. For example, LLMs could assist evidence synthesis teams in coding abstracts or primary data to answer questions, such as what services and values nature provides to people; determine the impacts of conservation management on local communities; and identify how much attention research topics or management approaches have received. Arguably, systems that automate data querying, processing, and categorization with ML and NLP could lead to living evidence syntheses that generate updates of new evidence on a regularly scheduled basis (Farrell et al., 2022; Sietsma et al., 2024; Thomas et al., 2017).

Large closed- and smaller open-sourced LLMs have exceeded nonexpert human coders in accuracy at a fraction of the cost (Alizadeh et al., 2025). These results suggest exciting horizons for hybrid LLM-expert systems for large-scale evidence reviews for which complete manual review is infeasible. Nevertheless, there are still substantive methodological questions without one-size-fits-all solutions, such as the most appropriate way to engineer prompts for zero- or few-shot LLM predictions (Farrell et al., 2024; Qureshi et al., 2023; Scheepens et al., 2024). In contrast to traditional manual data coding and evidence synthesis, research on bias in ML for environmental evidence synthesis remains limited (Farrell et al., 2024; Sietsma et al., 2024). We posit that interdisciplinary conservation social science teams could make exciting contributions in the domain of ethics and effective use. New guidance that builds on existing frameworks,

such as the CARE principles for Indigenous data governance (Carroll et al., 2020; Jennings et al., 2023) and FAIR data management principles (Wilkinson et al., 2016), for LLM-driven analyses is critically needed.

Transparency may help in this endeavor. Explainable AI (XAI) aims to make AI decisions more transparent through generated explanations (Adadi & Berrada, 2018; Phillips et al., 2020), a concept rooted in rule-based systems from nearly 50 years ago (Scott et al., 1977; Xu et al., 2019). Although XAI enhances transparency, curbs ethical risks, and accelerates discovery, rigorous validation across diverse data, tasks, and models is still crucial (Nauta et al., 2023). Careful scoping guided by interdisciplinary expertise, cocreating AI systems with conservation and community partners, and accounting for maintenance and deployment upfront will make AI systems more positively efficacious (Gebru & Torres, 2024; Moorosi et al., 2023).

In general, human review is essential because LLMs can fail at coding data in unpredictable ways, particularly for topics or attributes that are rare, are esoteric, involve complex logical reasoning, or require difficult trade-offs with no single correct decision (Kristensen-McLachlan et al., 2023; Zhu et al., 2023). Alizadeh et al. (2025) relied on data annotated by trained individuals using a structured codebook to ascertain LLM and crowdworker coding accuracy. Consider, for example, coding the statement, *The shift to electric vehicles was seen as a key climate action, though some pointed to the environmental costs of battery production*, for a positive, negative, or neutral opinion toward decarbonization via electrification. One could conceivably argue for any of those possible classes; thus, any AI system is unlikely to then come to a universally accepted decision. Reflecting on collaborations between social and computer scientists, DiMaggio (2015) wryly concluded that although computer scientists trust human judgment as a gold standard, social scientists, wary of human biases, hope algorithms can outperform flawed and inconsistent human reasoning, only to find that both can fall short in similar ways.

DISCUSSION

NLP and ML techniques are evolving rapidly. These new techniques promise to disrupt expectations related to the resources that evidence syntheses require and the scale and speed at which they can be conducted. For example, the latest advances in LLMs can accelerate automation of middle- and end-of-process steps in evidence collation and synthesis (e.g., filtering, data extraction, and analysis) (Farrell et al., 2024) (Figure 1; Marshall & Wallace, 2019; Thomas et al., 2017). For conservation scientists interested in topics that are not confined to a specific discipline or literature, such as climate change resilience, land tenure security, and environmental justice, existing and new NLP tools will continue to accelerate the synthesis of greater evidence volume (Farrell et al., 2022; Sietsma et al., 2024). This reflects a broader need for advancing sustainability science: the integration and synthesis of information with a focus on systems rather than disciplines.

Several areas require further research and debate for the rigorous application of NLP to evidence syntheses across conservation. ML and NLP do not provide a silver bullet for all evidence synthesis or primary research analysis needs (Marshall & Wallace, 2019; Thomas et al., 2017). At present, they are most useful when synthesis focuses on a broad topic with vast evidence. Assessing when and to what extent human review is necessary remains important for advancing ML synthesis methods. Given the emerging state of algorithms applicable at different review process steps (Figure 1; Table 1), expert guidance and review are essential (O'Connor et al., 2019; O'Mara-Eves et al., 2015; Qureshi et al., 2023).

Future advances in ML, NLP, and data extraction offer promise for enhancing the scope of automation. Increased ability for LLMs to take in large amounts of text at once (longer context windows), for example, makes it increasingly possible to process a published article in its entirety to more fully determine relevance for inclusion or to extract richer insights. Ongoing and future developments may make it possible to straightforwardly and robustly extract text versus data in tables and figures (Liu & McKie, 2024; Yang et al., 2019). However, setting and meeting uniform standards for data archiving and results presentation in papers would help synthesis science flourish with or without advanced ML tools (Fidler et al., 2017; Michener, 2015; Reichman et al., 2011; Sietsma et al., 2024).

Although NLP can integrate human checks of data, there are currently no standards or guidelines for determining acceptable thresholds and cutoffs for key decision points, which may lead to inconsistent use and results (O'Connor et al., 2019). In general, evidence syntheses must contend with uncertainty, whether workflows are automated or manual. NLP can introduce uncertainty when models use probabilistic determination, such as decisions on including versus excluding evidence. In contrast, manual syntheses include a comprehensive review of the entire sample set but may contain inconsistencies among human reviewers, uncertainties regarding inclusion criteria or coding, and human errors. Such mistakes can be mitigated but not eliminated by, for example, double coding. The assumption of human coding primacy has been challenged in interesting and difficult-to-anticipate ways (DiMaggio, 2015; Ziems et al., 2024). Thus, an intriguing use case for ML-assisted systems is correcting mislabeled data generated by human coders (O'Mara-Eves et al., 2015). As new tools leveraging cutting-edge NLP techniques emerge, future research should revisit assumptions in workflow design (O'Connor et al., 2019).

Although there are guidelines for traditional evidence syntheses (Haddaway et al., 2018; Page et al., 2021), thus far, there are no equivalent guidelines for AI-assisted syntheses in conservation and the environment. As a starting point, Beller et al. (2018) and O'Connor et al. (2019) discuss best practices for automated systematic reviews. Groups such as the Campbell Collaboration (2023) have also proposed future directions for standards of NLP and other ML techniques for evidence synthesis.

Conservation scientists eager to employ NLP for evidence synthesis must also grapple with doing so equitably. Existing NLP tools have an English language bias (Bang et al., 2023; Blevins & Zettlemoyer, 2022; Huang et al., 2023). Despite the surprising and strong performance of closed-source LLMs in multilingual text classification (Rathje et al., 2024), AI researchers caution that multilingual data sets, especially for the lowest-resource languages, may instead contain nonsense or machine-generated text rather than human writing (Hadgu et al., 2023). Overcoming these inequities requires a concerted effort to develop LLM training and instruction-tuning data for non-English languages.

There are substantive ethical questions about using AI for any social endeavor (Gebru & Torres, 2024), one of which revolves around the energy and environmental footprint of the largest models (Bender et al., 2021; Luccioni et al., 2024). Given that the energy and environmental costs of training LLMs have already been incurred, some would argue for applying existing models to conservation use cases. However, generative AI models demand much more energy and water than previous ML models; estimates indicate that one generative AI search request could demand 10 times as much energy as a conventional search (de Vries, 2023; Li et al., 2023; Luccioni et al., 2024). Moreover, although LLMs can provide positive benefits, there are also risks that these models reproduce social harms, such as biases in the training data that preclude marginalized groups or languages from inclusion (Bender et al., 2021; Hadgu et al., 2023) and privacy violations (Staab et al., 2023). Researchers are beginning to grapple with how to anticipate and combat potential harms in generative AI, calling for greater public accountability, fairness, and transparency (Gebru & Torres, 2024; Harrer, 2023).

Finally, as NLP methods advance, there must be an emphasis on transparency and replicability. Code and data should be publicly available (O'Connor et al., 2019). We should be able to call on time-stamped versions of LLMs so that results can be replicable; indeed, this is arguably an advantage for open-source LLMs (Abdurahman et al., 2024). Synthesis teams should fully document the tools they develop, flag where possible biases may emerge, and report uncertainty when appropriate (Sietsma et al., 2024; Thomas et al., 2017). This is especially important given that NLP analyses are often black-box processes and results are sensitive to choices, such as prompt engineering (Khattab et al., 2023; Kojima et al., 2022).

We outlined how emerging approaches to evidence synthesis can transform environmental, conservation, and sustainability science. Although we provided up-to-date examples, we emphasize that these methods are relatively new and poised to develop rapidly. Critical to the success of this burgeoning field will be developing a high-level process that can ensure the reliability and validity of ML-aided robust synthesis efforts. Equally important is the development of ethical frameworks to ensure that ML tools are applied in ways that reflect the diverse needs of all stakeholders. The potential for ML to enhance learning is vast, but it must be complemented by use-oriented design, stakeholder engagement, and multidisciplinary collaboration to guide ML-assisted efforts to the best global outcomes.

ACKNOWLEDGMENTS

ORCID

Charlotte H. Chang ⓘ https://orcid.org/0000-0002-1274-7721
Susan C. Cook-Patton ⓘ https://orcid.org/0000-0002-7194-4397
James T. Erbaugh ⓘ https://orcid.org/0000-0002-0602-9045
Luci Lu ⓘ https://orcid.org/0000-0003-2943-4032
Yuta J. Masuda ⓘ https://orcid.org/0000-0002-1698-4855
Dávid Papp ⓘ https://orcid.org/0000-0002-8814-2745
Brian E. Robinson ⓘ https://orcid.org/0000-0002-8972-8318

REFERENCES

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), Article 245.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2025). Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, *8*, Article 17. https://doi.org/10.1007/s42001-024-00345-9

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.-H., Peters, M., Power, J., Skjonsberg, S., Wang, L. L., … Etzioni, O. (2018). *Construction of the literature graph in semantic scholar*. arXiv. https://doi.org/10.48550/arXiv.1805.02262

Anderson, S. C., Elsen, P. R., Hughes, B. B., Tonietto, R. K., Bletz, M. C., Gill, D. A., Holgerson, M. A., Kuebbing, S. E., McDonough MacKenzie, C., Meek, M. H., & Veríssimo, D. (2021). Trends in ecology and conservation over eight decades. *Frontiers in Ecology and the Environment*, *19*(5), 274–282.

Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, *15*(1), 659–688.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*. arXiv. https://doi.org/10.48550/arXiv.2302.04023

Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T., Xia, J., Robinson, K., & Glasziou, P. (2018). Making progress with the automation of systematic reviews: Principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *7*, Article 77.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S., (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAccT '21)* (pp. 610–621). Association for Computing Machinery.

Bennett, N. J., Roth, R., Klain, S. C., Chan, K. M. A., Clark, D. A., Cullman, G., Epstein, G., Nelson, M. P., Stedman, R., Teel, T. L., Thomas, R. E. W., Wyborn, C., Curran, D., Greenberg, A., Sandlos, J., & Veríssimo, D. (2017). Mainstreaming the social sciences in conservation. *Conservation Biology*, *31*(1), 56–66.

Berrang-Ford, L., Siders, A. R., Lesnikowski, A., Fischer, A. P., Callaghan, M. W., Haddaway, N. R., Mach, K. J., Araos, M., Shah, M. A. R., Wannewitz, M., Doshi, D., Leiter, T., Matavel, C., Musah-Surugu, J. I., Wong-Parodi, G., Antwi-Agyei, P., Ajibade, I., Chauhan, N., Kakenmaster, W., … Abu, T. Z. (2021). A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change*, *11*(11), 989–1000.

Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F., Haddaway, N. R., Haines, A., & Dangour, A. D. (2021). Systematic mapping of global research on climate and health: A machine learning review. *The Lancet Planetary Health*, *5*(8), e514–e525.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blevins, T., & Zettlemoyer, L. (2022). Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3563–3574). Association for Computational Linguistics.

Bommarito, M. J., II, Katz, D. M., & Detterman, E. M. (2021). LexNLP: Natural language processing and information extraction for legal and regulatory texts. In R. Vogl (Ed.), *Research handbook on big data law* (pp. 216–227). Edward Elgar Publishing.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, *11*(2-3), 143–296.

Burnham, M. (2024). *Semantic SCALING: Bayesian ideal point estimates with large language models*. arXiv. https://doi.org/10.48550/arXiv.2405.02472

Callaghan, M., Schleussner, C.-F., Nath, S., Lejeune, Q., Knutson, T. R., Reichstein, M., Hansen, G., Theokritoff, E., Andrijevic, M., Brecha, R. J., Hegarty, M., Jones, C., Lee, K., Lucas, A., Van Maanen, N., Menke, I., Pfleiderer, P., Yesil, B., & Minx, J. C. (2021). Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change*, *11*(11), 966–972.

Campbell Collaboration. (2023, November). Stepping up evidence synthesis: faster, cheaper and more useful. https://www.campbellcollaboration.org/2023/11/stepping-up-evidence-synthesis/

Carroll, S., Garba, I., Figueroa-Rodríguez, O., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, *19*, Article 43.

Chang, C. H., Armsworth, P. R., & Masuda, Y. J. (2022a). Twitter data reveal six distinct environmental personas. *Frontiers in Ecology and the Environment*, *20*(8), 481–487.

Chang, C. H., Armsworth, P. R., & Masuda, Y. J. (2022b). Environmental discourse exhibits consistency and variation across spatial scales on Twitter. *Bioscience*, *72*(8), 789–797.

Chang, C. H., Erbaugh, J. T., Fajardo, P., Lu, L., Molnár, I., Papp, D., Robinson, B. E., Austin, K. G., Castro, M., Cheng, S. H., Cook-Patton, S., Ellis, P. W., Garg, T., Hochard, J. P., Kroeger, T., Mcdonald, R. I., Poor, E. E., Smart, L. S., Tilman, A. R., & Masuda, Y. J. (2025). A global evidence map of human well-being and biodiversity co-benefits and trade-offs of natural climate solutions. *Nature Sustainability*, *8*, 75–85. https://doi.org/10.1038/s41893-024-01454-z

Cheng, S. H., Augustin, C., Bethel, A., Gill, D., Anzaroot, S., Brun, J., DeWilde, B., Minnich, R., Garside, R., Masuda, Y., Miller, D. C., Wilkie, D., Wongbusarakum, S., & McKinnon, M. C. (2018). Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conservation Biology*, *32*, 762–764.

Cheng, S. H., MacLeod, K., Ahlroth, S., Onder, S., Perge, E., Shyamsundar, P., Rana, P., Garside, R., Kristjanson, P., McKinnon, M. C., & Miller, D. C. (2019). A systematic map of evidence on the contribution of forests to poverty alleviation. *Environmental Evidence*, *8*, Article 3.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). *LLM-assisted content analysis: Using large language models to support deductive coding*. arXiv. https://doi.org/10.48550/arXiv.2306.14924

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO : The SPIDER tool for qualitative evidence synthesis. *Qualitative Health Research*, *22*(10), 1435–1443.

Cook-Patton, S. C., Leavitt, S. M., Gibbs, D., Harris, N. L., Lister, K., Anderson-Teixeira, K. J., Briggs, R. D., Chazdon, R. L., Crowther, T. W., Ellis, P. W., Griscom, H. P., Herrmann, V., Holl, K. D., Houghton, R. A., Larrosa, C., Lomax, G., Lucas, R., Madsen, P., Malhi, Y., … Griscom, B. W. (2020). Mapping carbon accumulation potential from global natural forest regrowth. *Nature*, *585*(7826), 545–550.

Correia, R. A., Ladle, R., Jarić, I., Malhado, A. C. M., Mittermeier, J. C., Roll, U., Soriano-Redondo, A., Veríssimo, D., Fink, C., Hausmann, A., Guedes-Santos, J., Vardi, R., & Di Minin, E. (2021). Digital data sources and methods for conservation culturomics. *Conservation Biology*, *35*(2), 398–411.

Dasgupta, P. (2021). *The economics of biodiversity: The Dasgupta review*. HM Treasury.

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11), 688–701.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423

Díaz, S., & Malhi, Y. (2022). Biodiversity: Concepts, patterns, trends, and perspectives. *Annual Review of Environment and Resources*, 47, 31–63.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2), 2053951715602908.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, 46(1), 61–81.

Edwards, K., Song, B., Porciello, J., Engelbert, M., Huang, C., & Ahmed, F. (2024). ADVISE: Accelerating the creation of evidence syntheses for global development using natural language processing-supported human-artificial intelligence collaboration. *Journal of Mechanical Design*, 146(5), Article 051404.

Ellis, P. W., Page, A. M., Wood, S., Fargione, J., Masuda, Y. J., Carrasco Denney, V., Moore, C., Kroeger, T., Griscom, B., Sanderman, J., & Atleo, T. (2024). The principles of natural climate solutions. *Nature Communications*, 15(1), 547.

Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W., & Baronchelli, A. (2022). Growing polarization around climate change on social media. *Nature Climate Change*, 12(12), 1114–1121.

Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 113(1), 92–97.

Farrell, M. J., Brierley, L., Willoughby, A., Yates, A., & Mideo, N. (2022). Past and future uses of text mining in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, 289(1975), Article 20212721.

Farrell, M. J., Le Guillarme, N., Brierley, L., Hunter, B., Scheepens, D., Willoughby, A., Yates, A., & Mideo, N. (2024). The changing landscape of text mining: A review of approaches for ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, 291(2027), Article 20240423.

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., & Gordon, A. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. *Bioscience*, 67(3), 282–289.

Foo, Y. Z., O'dea, R. E., Koricheva, J., Nakagawa, S., & Lagisz, M. (2021). A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. *Methods in Ecology and Evolution*, 12(9), 1705–1720.

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews*, 8, Article 278.

Gates, A., Johnson, C., & Hartling, L. (2018). Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews*, 7(1), Article 45.

Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 29(4), Article 13636. https://doi.org/10.5210/fm.v29i4.13636

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.

Giebink, N., Gupta, A., Veríssimo, D., Chang, C. H., Chang, T., Brennan, A., Dickson, B., Bowmer, A., & Baillie, J. (2024). *Automating the analysis of public saliency and attitudes towards biodiversity from digital media*. arXiv. https://doi.org/10.48550/arXiv.2405.01610

Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, 10(10), 1645–1654.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2), 91–108.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 herd of models*. arXiv. https://doi.org/10.48550/arXiv.2407.21783

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Griscom, B. W., Adams, J., Ellis, P. W., Houghton, R. A., Lomax, G., Miteva, D. A., Schlesinger, W. H., Schoch, D., Siikamäki, J. V., Smith, P., Woodbury, P., Zganjar, C., Blackman, A., Campari, J., Conant, R. T., Delgado, C., Elias, P., Gopalakrishna, T., … Fargione, J. (2017). Natural climate solutions. *Proceedings of the National Academy of Sciences*, 114(44), 11645–11650.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://doi.org/10.48550/arXiv.2203.05794

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2), 181–217.

Haddaway, N. R., Macura, B., Whaley, P., & Pullin, A. S. (2018). ROSES RepOrting Standards for Systematic Evidence Syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environmental Evidence*, 7(1), Article 7. https://doi.org/10.1186/s13750-018-0121-7

Hadgu, A. T., Azunre, P., & Gebru, T. (2023). *Combating harmful hype in natural language processing*. Paper presented at the International Conference on Learning Representations (ICLR) 2023, Kigali, Rwanda. https://pml4dc.github.io/iclr2023/pdf/PML4DC_ICLR2023_39.pdf

Halterman, A. (2017). Mordecai: Full text geoparsing and event geocoding. *The Journal of Open Source Software*, 2(9), 91.

Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90, Article 104512.

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., & Wei, F. (2023). *Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting*. arXiv. https://doi.org/10.48550/arXiv.2305.07004

Intergovernmental Panel on Climate Change (IPCC). (2018). *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. Cambridge University Press.

Intergovernmental Panel on Climate Change (IPCC). (2021). *Climate Change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). (2018). *Summary for policymakers of the regional assessment report on biodiversity and ecosystem services for Africa of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES Secretariat.

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). (2019). *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES Secretariat. https://doi.org/10.5281/zenodo.3553579

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, 17(3), 805–826.

Jennings, L., Anderson, T., Martinez, A., Sterling, R., Chavez, D. D., Garba, I., Chavez, D. D., Garba, I., Hudson, M., Garrison, N. A., & Carroll, S. R. (2023). Applying the 'CARE Principles for Indigenous Data Governance' to ecology and biodiversity research. *Nature Ecology & Evolution*, 7(10), 1547–1551.

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, 96(3), 118–121.

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., & Miller, H. (2023). *Dspy: Compiling declarative language model calls into self-improving pipelines*. arXiv. https://doi.org/10.48550/arXiv.2310.03714

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Kristensen-McLachlan, R. D., Canavan, M., Kardos, M., Jacobsen, M., & Aarøe, L. (2023). *Chatbots are not reliable text annotators*. arXiv. https://doi.org/10.48550/arXiv.2311.05769

Kwabena, A. E., Wiafe, O. B., John, B. D., Bernard, A., & Boateng, F. A. (2023). An automated method for developing search strategies for systematic review using Natural Language Processing (NLP). *MethodsX*, *10*, Article 101935.

Lajeunesse, M. J. (2016). Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution*, *7*(3), 323–330.

Lambert, J., Epstein, G., Joel, J., & Baggio, J. (2021). Identifying topics and trends in the study of common-pool resources using natural language processing. *International Journal of the Commons*, *15*(1), 206–217.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models*. arXiv. https://doi.org/10.48550/arXiv.2304.03271

Liu, R., & McKie, J. X. (2024). *PyMuPDF*. http://pymupdf.readthedocs.io/en/latest/

Luccioni, S., Jernite, Y., & Strubell, E. (2024). Power hungry processing: Watts driving the cost of AI deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 85–99). Association for Computing Machinery.

Lunny, C., Brennan, S. E., McDonald, S., & McKenzie, J. E. (2017). Toward a comprehensive evidence map of overview of systematic review methods: Paper 1—Purpose, eligibility, search and data extraction. *Systematic Reviews*, *6*, Article 231.

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, *8*, Article 163.

McKinnon, M. C., Cheng, S. H., Dupre, S., Edmond, J., Garside, R., Glew, L., Holland, M. B., Levine, E., Masuda, Y. J., Miller, D. C., Oliveira, I., Revenaz, J., Roe, D., Shamer, S., Wilkie, D., Wongbusarakum, S., & Woodhouse, E. (2016). What are the effects of nature conservation on human well-being? A systematic map of empirical evidence from developing countries. *Environmental Evidence*, *5*(1), Article 8.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, *29*, 33–44.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Association for Computational Linguistics.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 201–237). Woodhead Publishing.

Moorosi, N., Sefala, R., & Luccioni, S. (2023). AI for whom? Shedding critical light on AI for social good. In *NeurIPS 2023 Computational Sustainability: Promises and Pitfalls from Theory to Deployment*. Neural Information Processing Systems.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42.

O'Connor, A. M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S. B., & Hutton, B. (2019). A question of trust: Can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews*, *8*, Article 143.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, *4*(1), Article 5.

Cranston, K. A., Redelings, B., Reyes, L. L. S., Allman, J., McTavish, E. J., & Holder, M. T., OpenTreeofLife. (2021). *Open Tree of Life taxonomy (v. 13.4)*. Zenodo. https://doi.org/10.5281/zenodo.3937750

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, Article n71.

Phillips, P. J., Hahn, A. C., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). *Four principles of explainable artificial intelligence* (Internal report 8312). US Department of Commerce, National Institute of Standards and Technology Interagency.

Porciello, J., Ivanina, M., Islam, M., Einarson, S., & Hirsh, H. (2020). Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning. *Nature Machine Intelligence*, *2*(10), 559–565.

Pörtner, H. O., Scholes, R. J., Agard, J., Archer, E., Arneth, A., Bai, X., Barnes, D., Burrows, M., Chan, L., Cheung, W. L., Diamond, S., Donatti, C., Duarte, C., Eisenhauer, N., Foden, W., Gasalla, M. A., Handa, C., Hickler, T., Hoegh-Guldberg, … Ngo, H. T. (2021). *Scientific outcome of the IPBES-IPCC co-sponsored workshop on biodiversity and climate change*. IPBES Secretariat.

Pörtner, H. O., Scholes, R. J., Arneth, A., Barnes, D. K. A., Burrows, M. T., Diamond, S. E., Duarte, C. M., Kiessling, W., Leadley, P., Managi, S., McElwee, P., Midgley, G., Ngo, H. T., Obura, D., Pascual, U., Sankaran, M., Shin, Y. J., & Val, A. L. (2023). Overcoming the coupled climate and biodiversity crises and their societal impacts. *Science*, *380*(6642), eabl4881.

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv. https://doi.org/10.48550/arXiv.2205.01833

Qureshi, R., Shaughnessy, D., Gill, K. A., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews*, *12*(1), Article 72.

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(34), Article e2308950121.

Rees, J. A., & Cranston, K. (2017). Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*, *5*, Article e12581.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703–705.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)* (pp. 3982–3992). Association for Computational Linguistics.

Rosenstock, T. S., Dawson, I. K., Aynekulu, E., Chomba, S., Degrande, A., Fornace, K., Jamnadass, R., Kimaro, A., Kindt, R., Lamanna, C., Malesu, M., Mausch, K., Mcmullin, S., Murage, P., Namoi, N., Njenga, M., Nyoka, I., Paez Valencia, A. M., Sola, P., … Steward, P. (2019). A planetary health perspective on agroforestry in Sub-Saharan Africa. *One Earth*, *1*(3), 330–344.

Scheepens, D., Millard, J., Farrell, M., & Newbold, T. (2024). Large language models help facilitate the automated synthesis of information on potential pest controllers. *Methods in Ecology and Evolution*, *15*, 1261–1273.

Scott, A. C., Clancey, W. J., Davis, R., & Shortliffe, E. H. (1977). Explanation capabilities of production-based consultation systems. *American Journal of Computational Linguistics*, *62*, 1–50.

Sietsma, A. J., Ford, J. D., & Minx, J. C. (2024). The next generation of machine learning for tracking adaptation texts. *Nature Climate Change*, *14*(1), 31–39.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*, 2960–2968. https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf

Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). *Beyond memorization: Violating privacy via inference with large language models*. arXiv. https://doi.org/10.48550/arXiv.2310.07298

Taranukhin, M., Shwartz, V., & Milios, E. (2024). Stance Reasoner: Zero-Shot Stance Detection on Social Media with Explicit Reasoning. arXiv. https://doi.org/10.48550/arXiv.2403.14895

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., Mcdonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., Elliott, J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., Pestridge, C., Albrecht, L., Horsley,

T., … Pearson, L. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology*, *91*, 31–37.

Tsou, A. Y., Treadwell, J. R., Erinoff, E., & Schoelles, K. (2020). Machine learning for screening prioritization in systematic reviews: Comparative performance of Abstrackr and EPPI-Reviewer. *Systematic Reviews*, *9*(1), Article 73. https://doi.org/10.1186/s13643-020-01324-7

Van Driel, M. L., De Sutter, A., De Maeseneer, J., & Christiaens, T. (2009). Searching for unpublished trials in Cochrane reviews may not be worth the effort. *Journal of Clinical Epidemiology*, *62*(8), 838–844.

Van Houtan, K. S., Gagne, T., Jenkins, C. N., & Joppa, L. (2020). Sentiment analysis of conservation studies captures successes of species reintroductions. *Patterns*, *1*(1), Article 100005.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 6000–6010). Neural Information Processing Systems.

Villanueva, E. V., Burrows, E. A., Fennessy, P. A., Rajendran, M., & Anderson, J. N. (2001). Improving question formulation for use in evidence appraisal in a tertiary care setting: A randomised controlled trial [ISRCTN66375463]. *BMC Medical Informatics and Decision Making*, *1*(1), Article 4. https://doi.org/10.1186/1472-6947-1-4

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 819–824). Association for Computing Machinery.

Westgate, M. J., Barton, P. S., Pierson, J. C., & Lindenmayer, D. B. (2015). Text analysis tools for identification of emerging topics and research gaps in conservation science. *Conservation Biology*, *29*(6), 1606–1614.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 160018.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference* (pp. 563–574). Springer International Publishing.

Yang, H., Aguirre, C. A., Maria, F., Christensen, D., Bobadilla, L., Davich, E., Roth, J., Luo, L., Theis, Y., Lam, A., & Han, T. Y. J. (2019). Pipelines for procedural information extraction from scientific literature: Towards recipes using machine learning and data science. In *2019 International conference on document analysis and recognition workshops (ICDARW)* (Vol. 2, pp. 41–46). Institute of Electrical and Electronics Engineers.

Zhu, Y., Zhang, P., Haq, E. U., Hui, P., & Tyson, G. (2023). *Can ChatGPT reproduce human-generated labels? A study of social computing tasks*. arXiv. https://doi.org/10.48550/arXiv.2304.10145

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291.