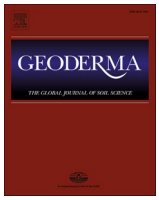
Geoderma 471 (2026) 117890

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



Rapid evidence mapping of soil fauna responses to agricultural management assisted by large language models

Ricardo Leitao [a,*], Vid Podpečan [b], Luis Cunha [a], Carmen Vazquez [c], Felix David [c,e], Camille Imbert [d], Sarah Symanczik [e], José Paulo Sousa [a], Else K. Bünemann [e], Marko Debeljak [b], Rachel Creamer [c], Martina Lori [e]

[a] Centre for Functional Ecology, Associate Laboratory TERRA, Department of Life Sciences, University of Coimbra, Coimbra, Portugal
[b] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[c] Soil Biology Group, Department of Environmental Science, Wageningen University and Research, Wageningen, the Netherlands
[d] Research Area 3 "Agricultural Landscape Systems", Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder Straße 84, 15374 Müncheberg, Germany
[e] Department of Soil Sciences, Research Institute of Organic Agriculture (FiBL), Ackerstrasse 113, 5070 Frick, Switzerland

ARTICLE INFO

Handling Editor: Yvan Capowiez

Keywords:
Text mining
Knowledge extraction
Artificial intelligence
Soil biota
Biochar
Crop residues

ABSTRACT

The exponential growth of scientific literature challenges traditional synthesis methods, limiting our ability to derive broad insights into complex environmental processes. Large language models (LLMs) may help address this challenge by supporting the extraction of structured knowledge from unstructured text. To evaluate and demonstrate this potential in soil science, we developed a multi-module, LLM-assisted knowledge-extraction workflow built on a recently published *meta*-data-analysis of management–biota interactions as a contextual framework, using an iteratively refined prompt chain to extract directional relationships between agricultural management practices and soil fauna from scientific abstracts. Benchmarking against manually curated datasets showed high precision and recall, while expert-guided iterative development indicated that the information content of abstracts was likely a major practical constraint on extraction performance. To assess interpretative soundness, we applied the workflow in two use cases: an illustrative comparison with the well-established literature on reduced and no-tillage effects on soil fauna, and a knowledge-gap application on biochar and crop-residue retention. The workflow indicated predominantly beneficial reported patterns for crop residue retention, particularly for earthworms and nematodes, whereas biochar showed a more heterogeneous and context-dependent pattern. Overall, results show that LLM-assisted workflows can support rapid, large-scale evidence mapping of soil fauna responses to management practices when formal quantitative syntheses are unavailable. The proposed framework is best understood as a complementary tool for organising and screening dispersed ecological evidence, and not as a substitute for full-text synthesis, effect-size-based *meta*-analysis, or decision-grade inference. The complete workflow is publicly available and broadly transferable across environmental research domains.

1. Introduction

Understanding how human interventions alter ecosystems has become a scientific challenge, especially in an era of exponentially growing scientific literature and accelerating technological change (Keck et al., 2025; McFadden et al., 2023). The surge of artificial intelligence (AI) adds a new dimension to this challenge by accelerating ongoing transformations in how human–environment interactions are studied, reshaping not only science but society as a whole (Gao and Wang, 2024; Gohr et al., 2025).

In this rapidly evolving AI landscape, recent methodological advances have been particularly transformative for how scientific knowledge can be processed and mapped. Among these advances, large language models (LLMs) represent a consequential shift, transforming AI from a specialist domain into a broadly accessible technology, enabling researchers across disciplines to tackle complex analytical tasks (Birhane et al., 2023; Zhang et al., 2025) and extend its application beyond conventional AI fields (e.g., computer vision, robotics) to

* Corresponding author.
E-mail address: ricvl@ci.uc.pt (R. Leitao).

https://doi.org/10.1016/j.geoderma.2026.117890
Received 13 January 2026; Received in revised form 5 April 2026; Accepted 1 June 2026
Available online 10 June 2026

methodological tasks such as *meta*-analysis (Galli et al., 2025).

These developments are supported by a set of technical capabilities that distinguish LLMs from earlier text mining and rule-based approaches (Pai et al., 2024). Key strengths include the ability to parse complex grammatical structures and understand specialized, domain-specific terminology (especially when fine-tuned on domain-specific data) (Idan and Einav, 2025), identify semantic relationships between entities mentioned far apart in the text, and process immense volumes of unstructured text, far exceeding human capacity (Li et al., 2025). However, challenges remain. LLMs can be prone to *hallucinations*, generating plausible but factually incorrect information, and their performance is susceptible to the specific phrasing of the input prompt (Chen et al., 2025). Extracting precise quantitative data or navigating highly nuanced or implicitly stated knowledge can also be difficult without sophisticated prompting strategies or model fine-tuning. Furthermore, potential biases embedded in the training data can also influence outputs (Navigli et al., 2023).

Despite these limitations, an expanding body of research shows that LLMs can already enable automated knowledge extraction across diverse scientific and applied domains. In environmental sciences, LLMs are currently used for a range of tasks, from identifying future research opportunities (Ji et al., 2025) to informing environmental decision-making strategies (Nie and Liu, 2025). Soil science offers vast application potential, as many key soil processes and management–biota interactions remain poorly resolved across scales (Wadoux, 2025; Wu, 2025), with LLM-assisted data-mining approaches pioneered by Blanchy et al. (2023).

Among environmental disciplines, soil science is particularly compelling because of its ecological importance and the complexity of its management-driven biological interactions. In this context, a healthy soil, supported by its high biodiversity, forms the foundation of resilient terrestrial ecosystems and sustainable agriculture, ultimately ensuring global food and nutritional security (Bardgett and van der Putten, 2014; Lehmann et al., 2020). The complex subterranean world hosts diverse biological actors, including soil fauna, (e.g., earthworms, springtails, nematodes) and vast microbial communities, that collectively drive ecosystem functions (Delgado-Baquerizo et al., 2020; Wall et al., 2015) such as nutrient cycling, carbon and climate regulation, water regulation and purification, and disease and pest regulation (Creamer et al., 2022). The composition, abundance, and activity of these vital soil communities can be strongly shaped by agricultural management practices that alter soil physical disturbance and organic matter inputs, including contrasting tillage regimes and carbon-based amendments such as biochar and crop-residue retention (Lori et al., 2025, 2017; Tsiafouli et al., 2015). Consequently, gaining a comprehensive understanding of the intricate relationships between specific agronomic interventions and the responses of key soil fauna groups is essential to ensuring healthy soils for future generations (FAO, 2024; Zwetsloot et al., 2022). Given the volume and heterogeneity of the literature addressing soil biota responses to agricultural management, scalable approaches for rapidly identifying and structuring reported evidence are increasingly valuable.

Abstracts occupy a central role in scientific communication because they provide open access, concise, structured summaries of a study's objectives, methods, and principal findings, and they are widely used for relevance screening in systematic reviews and evidence (Haddaway et al., 2018). Their high information density makes them a practical resource for rapid, large-scale literature analyses. Additionally, full texts are often access- or license-restricted, which can impose practical and legal constraints on large-scale AI-assisted analysis (Esteve, 2024; Margoni and Kretschmer, 2022). At the same time, abstract-based approaches have recognised limitations that must be considered when interpreting the resulting patterns. Because abstracts are selective summaries, they may preferentially emphasise positive or statistically significant findings, while negative, null, or more nuanced results may be underrepresented relative to the full text (Duyx et al., 2019). More broadly, abstracts tend to foreground high-level claims and often capture less methodological detail, conditional responses, and finer-grained biological or ecological information than the main article body (Cohen et al., 2010; Penning de Vries et al., 2020; Schuemie et al., 2004). Consistent with this, full-text approaches achieve more comprehensive information coverage than analyses based on abstracts alone (Westergaard et al., 2018). Nevertheless, for rapid evidence mapping, abstracts remain a valuable and efficient entry point for identifying broad study topics and reported response patterns (Glasziou et al., 2008). Accordingly, abstract-based extraction is most appropriate for rapid evidence mapping of reported outcomes, rather than for estimating effect sizes, weighting evidential strength, or supporting quantitative causal inference.

Drawing on these considerations, LLM-assisted approaches offer a promising way to extract structured ecological knowledge directly from large volumes of abstracts. Building on this potential, the present study examines the effectiveness of state-of-the-art LLMs for targeted and complex structured knowledge extraction from a curated collection of agricultural soil science abstracts. More specifically, our workflow is designed to generate a rapid evidence map of reported relationships between soil management practices and soil fauna by extracting and organising directional study outcomes from abstracts. The resulting summaries are intended to identify broad reported tendencies, knowledge gaps, and areas of inconsistency across fragmented literatures. Rather than replicating the inferential framework of formal *meta*-analysis, our approach provides a complementary tool for quickly mapping the existing evidence landscape, identifying knowledge gaps, and prioritising topics for future in-depth quantitative reviews, and therefore does not estimate effect sizes, uncertainty, or variance-weighted pooled responses. We present a practical workflow that can be adopted across various domains to support the extraction of structured knowledge from abstracts.

Building on this rationale, the specific objectives of this paper are:

1. To create and demonstrate an LLM workflow for automated extraction of structured knowledge on land management practices, soil fauna, and their reported interactions, from scientific abstracts.
2. To quantitatively and qualitatively evaluate the performance of the LLM-assisted knowledge-extraction benchmarked against expert-annotated sets of abstracts.
3. To demonstrate the workflow's applicability through two complementary use cases: (i) benchmarking its outputs against existing *meta*-analyses on reduced and no-tillage effects on soil fauna, and (ii) applying it to the evidence base on biochar and crop-residue retention, where comprehensive *meta*-analyses are lacking.
4. To critically discuss the interpretive scope, key challenges, opportunities for methodological refinement, and future research directions for using LLMs to accelerate knowledge mapping within soil science and related environmental fields, as well as their potential applications.

2. Methods

2.1. Workflow description

We developed a multi-module, LLM-assisted workflow to extract domain-specific information on soil management practices, soil biological actors, and their interactions from scientific abstracts of peer-reviewed, published scientific articles (Fig. 1). Technically, this workflow can be considered a rapid evidence mapping framework that integrates elements of systematic maps and rapid evidence assessments with the directional summarisation approach of vote counting (Dicks et al., 2017). It is structured around a carefully designed prompt chain (Supplementary Data 1) and a set of complementary Python and R programs (SI programs 1 to 4) that ensure contextual relevance, consistency, and reproducibility throughout the knowledge-extraction workflow. The workflow was grounded in the *meta*-data-analysis

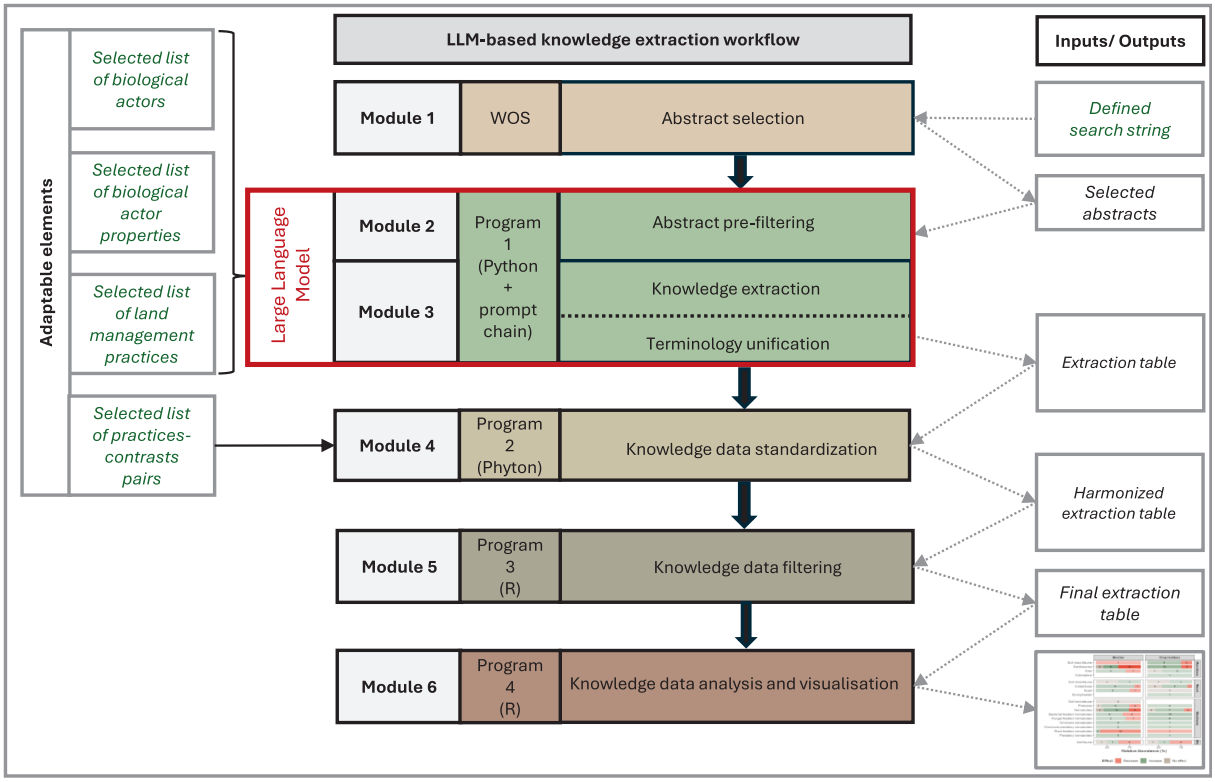


Fig. 1. Schematic overview of the LLM-based knowledge extraction. This figure presents the architecture, programs, and output files of the multi-module large language model (LLM)-assisted knowledge extraction workflow. Module 1 (Abstract selection) is run manually on Web of Science (WOS). Modules 2 (Abstract pre-filtering) and Module 3 (Knowledge extraction and Terminology unification) use the prompt chain template (Program 1, implemented in Python). Module 4 (Knowledge data standardization) is implemented in Python (Program 2). Module 5 (Knowledge data filtering) and Module 6 (Knowledge data analysis and visualization) are based on R (Programs 3 and 4, respectively).

framework of Lori et al. (2025), a systematic review of *meta*-analysis, which provided the context and a template for the knowledge structure and analytical approach. Although the workflow is LLM-assisted, the final outputs reflect both LLM performance and human-designed analytical structure. The LLM performs the core extraction, classification, and terminology-unification tasks, whereas the workflow defines the conceptual elements, prompt logic, controlled categories, and curation rules used to organise the results. Accordingly, the resulting evidence maps are shaped both by the language-processing capacity and by design choices embedded in the analytical framework.

The workflow comprises six sequential functional modules (Fig. 1). Module 1, *Abstract selection*, begins with the selection of abstracts, which is performed through the design of a tailored search string to target the Web of Science (WOS) Core Collection. In Module 2, *Abstract pre-filtering*, the workflow continues with a prompt chain embedded action that automatically pre-filters the selected abstracts for topical relevance consistent with the Module 3 configuration. This pre-filtering step ensures that only relevant abstracts are passed to Module 3, which constitutes the central knowledge-extraction step of the workflow and is described in detail below.

Module 3, *Knowledge extraction*, is implemented as a double-staged process that applies a prompt chain to identify and extract knowledge on management practices, biological actors, and their interactions during the extraction stage. In parallel, a Terminology unification stage, through an additional prompt chain action, employs predefined selected lists (adaptable elements) to map multiple extracted items onto a standardized set of terminology.

As the central and most technically demanding component of the workflow, Module 3 was developed through the creation of instructions for the LLM to perform knowledge extraction (Fig. 1). Built around the LLM as its core tool, this module was developed through successive prompt chain templates, each progressively refining the extraction and structuring of domain-specific information (see Supplementary Data 2, Section 1 for more details). During its development, the model's performance was benchmarked across several abstract collections (Supplementary Data 3) using ChatGPT versions 3.5-4o. Considering the authors' knowledge background in soil science, certain choices and applications within the workflow reflect the agricultural soil sciences context, as well as the use cases adopted in this study (see Section 2.3). Nonetheless, the workflow was conceived as a flexible framework, readily adaptable to diverse research domains and analytical objectives beyond the soil context (Supplementary Data 2, Table S1).

In detail, the prompt chain in Module 3 (named LEPAMTIC template) instructed the LLM to extract knowledge using a structured format composed of ten interconnected *extraction elements*:

Land management practice (L)* – agricultural intervention applied (e.g., biochar, tillage);
Effect (E)* – direction or outcome of the intervention, limited to the options: increase, decrease, neutral, or not available;
Property (P)* – measurable characteristic evaluated (e.g., biomass, richness);
Soil biota actor (A)* – biological entity affected (e.g., earthworms, nematodes);
Method or measurement (M) – approach used to assess the property (e.g., qPCR, microscopy);
Temporal scope (T) – time frame of the assessment (e.g., growing season, months after management operation);
Locational scope (I) – location of the sampling (e.g., soil layers or sampling position);

Contrasting land management practice (C)* – control management scenario used for comparison (e.g., unfertilized, conventional tillage).
Location country – country where the study was conducted
Study type – type of study (e.g., pot, greenhouse)

By using this prompt chain template, the model was instructed to identify and extract specific, compact knowledge patterns representing complete and interpretable units of information on the effects of agronomic interventions on biological actors. The extraction elements were specifically defined to capture these effects in the context of soil biological actors and their properties. Nonetheless, the workflow is readily adaptable, allowing different extraction elements to be defined for other research contexts, as described earlier. The result of the Module 3 extraction stage is the *extraction table*, a structured flat file in which each row corresponds to an extraction pattern and each column corresponds to an *extraction element*.

Alongside the key *extraction elements* (highlighted with *) of an *extraction pattern* used for the directional evidence map, additional extraction elements enabled the workflow to capture contextual metadata from abstracts (such as method or measurement, temporal scope, locational scope, country, and study type) when explicitly reported. In the present study, these variables were retained as complementary contextual descriptors, although only study type was used operationally for filtering, because the remaining qualifiers were reported too inconsistently across abstracts to support robust, context-resolved analysis at scale. An additional scoring component was retained in the prompt structure as an exploratory output to support ongoing testing of whether abstract reporting quality is associated with extraction performance; however, because preliminary analyses did not reveal a sufficiently informative relationship, this output was not used in the main analyses of the present study.

To support downstream evidence mapping and enable structured comparisons, we included in Module 3 an additional Terminology unification stage for the key *extraction elements*, resulting in additional columns in the *extraction table*. In this process, embedded in the prompt chain, the LLM was instructed to match the content of each column to a predefined term from a list of *adaptable elements* (Supplementary Data 4). These lists, whether provided as external files or included directly in the model's prompt, were selected for this study's context and can be tailored to other research contexts. They were designed to unify terminology and thereby minimize variability arising from inconsistent terminology across abstracts. The *effect* column, due to its simplicity, was embedded in the prompt chain instructions and not included in this stage. The prompt chain was executed via the OpenAI API using a custom Python program (SI Program 1), and all programs are publicly available on GitHub (https://github.com/vpodpecan/LEPAMTIC).

Module 4, *Knowledge data standardization*, was implemented using an eight-step post-processing data approach using a Python program (for more details, see Supplementary Data 2, Fig. S2; SI program 2). The overarching goal was to establish single, unambiguous levels for each key *extraction element*, validate the practice–contrast logic aligned with the *adaptable elements* of Module 3 (Supplementary Data 4), and eliminate redundancies, resulting in a *harmonization table* that ensured comparability of practice–actor combinations across studies.

Module 5*, Knowledge data filtering*, implemented in R (SI program 3), tailors the harmonized data to address the workflow use cases to produce the *final extraction table*.

The final Module 6, *Knowledge data analysis and visualization*, also implemented in R (SI program 4), generates ready-to-publish overview visualizations adapted to the selected workflow use cases. All analyses were conducted in R version 4.4.3 within the RStudio environment version 2024.12.1, with plots generated using the tidyverse (Wickham et al., 2019) and ggh4x (van den Brand, 2024) packages.

2.2. Module 3 evaluations: Knowledge extraction and terminology unification

Module 3 is the only workflow component that relies on probabilistic LLM outputs and subjective semantic interpretation; therefore, it requires explicit evaluation to assess performance and suitability for structured knowledge extraction. Accordingly, we evaluated the prompt chain in the knowledge extraction and terminology unification stages of Module 3 using expert annotations as the reference standard (SI program 6). To assess the quality of LLM knowledge extraction, a collection of 40 abstracts from peer-reviewed agriculture and soil science journals was assembled (see the final *abstract set* in Supplementary Data 3). This curated set was used as input for LLM Module 3 to generate an *extraction table*. Using the same abstract collection, experts independently manually extracted knowledge (Supplementary Data 5) following detailed extraction guidelines (Supplementary Data 6), enabling a standardized and directly comparable evaluation of the LLM's extraction performance using the extraction evaluation framework (Fig. 2).

In parallel with the iterative workflow development and evaluation carried out over the study period, we benchmarked newly released LLMs against the model initially used for development (GPT-4o) to confirm the most suitable option for structured knowledge extraction in Module 3. A dual LLM- and expert-based assessment was implemented to compare models' performance (Module 3, SI program 5); however, this complementary evaluation was used only for model screening, and did not replace the expert-based benchmark evaluation reported in the manuscript. In this evaluation step, additional structural metrics, including the number of extracted patterns and the number of levels per pattern, were first analysed, after which reports generated by an LLM (Gemini 2.0 Flash, Google) were reviewed by human evaluators (see Supplementary Data 2, Figs. S2–S9 and *LLMs_evaluation final report*, respectively). GPT-4o remained the most consistent and human-aligned, and was therefore retained for this study (see Supplementary Data 2, Figs. S2–S9).

2.2.1. Evaluation of the knowledge extraction stage

To quantitatively evaluate LLM-assisted knowledge extraction performance, we adopted standard classification metrics: precision (P), recall (R), and F1 score (F1) (Fig. 2). Performance was assessed by requiring exact matches between model outputs and expert annotations, at both the individual extraction element level and the extraction pattern level.

Precision is the proportion of correctly extracted elements among all predicted elements, showing the model's resistance to false positives. Recall measures the proportion of relevant elements correctly extracted, capturing retrieval completeness. The F1 score, the harmonic mean of P and R, balances completeness and correctness. P, R, and F1 were evaluated separately for each individual *extraction element* and for each *extraction pattern*. Because Module 3 relies on probabilistic LLM outputs, robustness and reproducibility were addressed primarily through prompt-chain design. The LEPAMTIC prompt chain was iteratively refined to make instructions highly explicit, reduce ambiguity, and account for recurrent special cases, thereby promoting consistent and reproducible structured outputs. Repeated runs during development suggested only minor practical variability, although this was not formally quantified. Where supported by the model, the extraction pipeline also allows the specification of seed and temperature parameters to further enhance reproducibility.

For the pattern-level evaluation, *extraction patterns* were defined by combining entries from selected *extraction elements* in the output, with each unique combination treated as a distinct pattern. Three out of ten *extraction elements* (method, temporal, and location scopes) were excluded here due to irrelevance for downstream analysis. Performance was assessed by requiring exact matches between the selected combined *extraction elements* and the corresponding expert annotations. This means that even a single mismatch in one of the selected elements

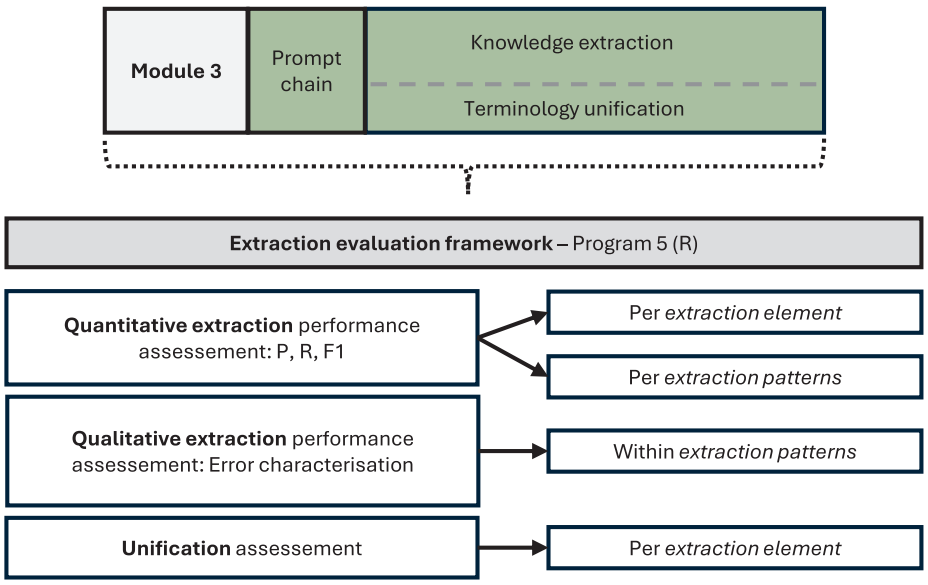


Fig. 2. Schematic overview of the extraction evaluation framework. This figure illustrates the extraction evaluation framework used to evaluate the performance of the *Knowledge extraction* and *Terminology unification* stages within Module 3, as well as their various components and associated analyses. Precision (P), recall (R) and harmonic mean of P and R (F1) are used as evaluation metrics.

results in the entire pattern being classified as incorrect. Such an approach is deliberately conservative, as it magnifies potential discrepancies and thus provides a very stringent baseline for estimating overall performance across patterns.

For each abstract and expert, we counted true positive elements/patterns (TP: exact matches between model and expert extractions), false positive elements/patterns (FP: model extractions not present in the expert annotation), and false negative elements/patterns (FN: human extractions missed by the model). From these, we derived, per abstract:

$$\text{Precision (P)} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall (R)} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{F1 score (F1)} = 2 \times (\text{P} \times \text{R})/(\text{P} + \text{R})$$

To remove bias from varying pattern counts, we applied per-abstract normalization. Inter-annotator agreement was assessed using the same method as the expert-LLM comparison, but limited to expert-expert comparisons per pattern. To complement the quantitative evaluation, the extraction evaluation framework included a qualitative characterization of model extraction errors. This approach enabled us to move beyond binary (correct/incorrect) performance metrics and explore how model errors were distributed across semantic categories, providing a more nuanced understanding of model limitations.

2.2.2. *Evaluation of the terminology unification stage*

To further evaluate the reliability of the *Terminology unification* stage (Module 3; Fig. 1) of the LLM's key *extraction elements*, four expert annotators independently assessed the unified columns as part of the extraction evaluation framework (Fig. 2). Thus, for each of the 40 abstracts in the *abstract set final* (Supplementary Data 3), annotators assessed the model's output for each unified key *extraction element* and classified the results (Supplementary Data 5).

2.3. *Workflow application*

The LLM-assisted knowledge extraction workflow was applied to two complementary use cases, grounded in the *meta*-data-analysis used as guidance (Lori et al., 2025). The first use case consists of an illustrative comparison with the well-established literature on the effects of reduced and no-tillage on soil fauna (macro-, meso-, and microfauna). It is intended to benchmark the workflow against a reference *meta*-data-analysis and to clarify its inferential boundaries, rather than to provide a strong independent validation across heterogeneous or contested topics.

The second use case addressed an area with limited and fragmented evidence, as identified in the reference *meta*-data-analysis of Lori et al. (2025), concerning the impacts of biochar and crop residue retention on soil fauna, and was intended to generate new evidence-mapping insights through the application of the developed workflow.

In both use cases, the directional outcomes extracted by the LLM-assisted workflow were summarised through frequency-based aggregation (vote counting), such that each extraction pattern contributed equally to the resulting overview (Dicks et al., 2017; Friedman, 2001). Accordingly, these summaries were intended to characterise the distribution of reported directions across the literature, rather than to estimate effect sizes, uncertainty, or variance-weighted pooled responses.

To facilitate comparative analysis, taxonomic handling followed the same general strategy adopted in the reference *meta*-data-analysis of Lori et al. (2025). When abstracts provided sufficient detail, responses were retained at the most specific biological actor (e.g., earthworms, collembola) or trophic-group level (e.g., fungal feeding nematodes) extracted by the workflow. When abstracts reported only broader categories (e.g., soil fauna, macrofauna, nematodes), these were preserved as distinct aggregated categories rather than redistributed across more specific groups. Thus, broader and finer-resolution categories were treated as complementary, non-overlapping evidence units, reflecting the level of taxonomic detail actually available in the source abstracts. Likewise, where response metrics were not reported consistently enough to support stable interpretation at finer resolution, results were aggregated to higher biological-actor levels. This approach necessarily sacrifices some biological resolution, but avoids imposing unsupported taxonomic specificity during harmonisation.

To further facilitate comparative analysis and the interpretation of results, data visualization methods were also aligned with the graphics and systematic framework established by Lori et al. (2025).

For both use cases, a comprehensive search was conducted in the Web of Science (Module 1), encompassing all available databases. Retrieval was strictly limited to original research articles (document type: article), explicitly excluding review papers, *meta*-analyses, and records from the Preprint Citation Index to ensure the acquisition of primary empirical data. The search for the biochar and crop residue retention component was conducted on August 26th, 2025. The no-

tillage and reduced tillage search, later incorporated as the illustrative comparison with the reference *meta*-data-analysis, was conducted on December 16th, 2025.

Following the finalization of the search string and the prompt chain–model configuration, the complete workflow across all modules required approximately two working days to run for each component. All procedural aspects of the workflow remained consistent between the two use cases, with the exception of Module 1 search string (Supplementary Data 2, Section 6), Module 5, which was tailored for filtering each land management practice (no-tillage/reduced tillage vs. biochar/crop residues), and Module 6, which incorporated minor adjustments aligned with the particularities of each knowledge domain and results.

3. Results and discussion

3.1. Module 3 evaluation

To benchmark the LLM-assisted workflow's core module knowledge extraction (Module 3, Stage 1) performance, the results of expert annotators were compared with those of the LLM-assisted knowledge extraction. The highest P, R and F1 values were observed for the elements *land management practice* and *actor,* with values across annotators ranging from 94 to 97% (P), 83 to 89% (R), and 87 to 92% (F1), and from 94 to 96% (P), 84 to 91% (R) and 88 to 93% (F1), respectively (Fig. 3). The *contrasting land management practice* and *property* elements consistently showed the lowest values across metrics, ranging from 80 to 85% (P), 74 to 78% (R), and 76 to 80% (F1), and 85 to 95% (P), 78 to 84% (R), and 81 to 88% (F1) respectively (Fig. 3), suggesting these extraction tasks were particularly complex for the model. Despite element-specific limitations, the overall high values observed for P, R, and F1 provide high confidence in the model's reliability when assessed from an extraction-element perspective.

When evaluated at the level of complete *extraction patterns* rather than individual elements, the performance across annotators remained consistently high, yielding mean P, R, and F1 across experts of 62–70%, 57–64%, and 59–65%, respectively (Fig. 3). P and R values were comparable, indicating balanced performance of the model. This balance is important, as improvements in one metric come at the expense of the other. The similar P and R levels indicate a good compromise, reinforcing the robustness of the extraction process. These findings demonstrate that, despite the conservative criteria applied, the model maintained a relatively strong performance when evaluated at the level of complete *extraction patterns* (Wan et al., 2024; Zhao and Goto, 2025). In the qualitative evaluation of *extraction-pattern* error characterization, the *all-correct* category was dominant, representing the majority of the model's extractions across annotators, followed by the *partially correct* category, whereas only a small number of extractions were classified as *incorrect* (for a more detailed error description, see Supplementary Data 2, Section 7).

Interestingly, performance varied considerably among annotators and *extraction elements*. Inter-annotator agreement (Supplementary Data 2, Table S2) revealed substantial differences, despite the use of detailed *expert extraction guidelines* (Supplementary Data 6), indicating that extraction variability was not limited to the model and that discrepancies in P, R, and F1 were often greater among experts than between the LLM and the experts. These differences can be understood as a consequence of less clear abstracts, where subjectivity inevitably plays a role. When information is incompletely reported, vague, or ambiguously presented, even experts applying standardized criteria may interpret or prioritize elements differently. Thus, lack of clarity and completeness in abstracts not only hinders automated knowledge mining but also affects the consistency of human extractions. In turn, this argues for employing LLMs to support such tasks, as they can reduce subjectivity and enhance consistency.

To benchmark the T*erminology unification* stage of LLM-assisted knowledge extraction in Module 3 (Fig. 1), the correctness of the unified outputs was evaluated. Correctness was defined as agreement between the model-assigned unified category and the corresponding expert annotation for each key extraction element. The LLM-assisted *Terminology unification* process demonstrated high accuracy across annotators and all respective key *extraction elements* (Fig. 4). The *correct* unification category was consistently above 75%, ranging from 76% for *contrasting land management practices* to 96% for the *actor* element. The *incorrect* category occurred from 0.5% for *contrasting land management practices* to 21% for *property*. For the multiplexed categories (more than one element extracted), *one-correct* cases (one element correctly unified) ranged from 0.5%, for the *actor* category, to 18.5%, for the *contrasting land management practice* category. *Multiple-correct* (all elements correctly unified) had a residual occurrence. For more detailed information on unification evaluation and prompt chain versions, see Supplementary Data 2, Section 9.

These results provided confidence that the *Terminology unification* stage, performed exclusively by the LLM and relying only on pretrained knowledge and relevant sentences, can be trusted for downstream analyses. Some inter-annotator variability was observed, reflecting differences in human judgment when aligning terms.

3.2. Workflow application outputs and data aggregation

To assess the overall effects of management practices on soil fauna, we aggregated LLM-extracted and harmonised patterns at the biological actor level in both application use cases (Fig. 5). While aggregation to higher levels increases analytical robustness, the workflow captures knowledge at multiple hierarchical levels, demonstrating the level of resolution achievable through LLM-assisted extraction. However, at the property–actor level (abundance, activity, diversity, and ecological index), the number of extracted patterns was insufficient for meaningful inference; the corresponding plots are shown in the Supplementary Material (Supplementary Data 2, Figs. S14 and S17) and are not discussed further. If needed, actors or trophic groups can be readily re-aggregated to support alternative groupings. In our use cases, we included microfauna, mesofauna, and macrofauna, and focused our discussion of results on those groups that yielded the highest number of extracted patterns. In addition to the detailed faunal groups, some categories in our datasets are intrinsically aggregated by their semantics, such as *Soil fauna*, *Soil macrofauna*, or *Nematodes*. These broader categories resulted either from the unification process during data harmonization or from the lack of detail reported in the primary abstract. Importantly, they do not overlap with the more specific categories but instead contribute complementary information that expands the scope of our synthesis. For nematode actors, we chose to retain information at the trophic group level whenever available, while also keeping a naturally aggregated category for the undetailed data; however, as an exception, the LLM generated a separate category called *pest-parasitic nematodes*, which we merged with *plant-feeding nematodes* after revising the corresponding abstracts, since both represent the same functional group. Consistent with the reference systematic framework established by Lori et al. (2025), studies investigating no-tillage and reduced tillage were consolidated under a single *Reduced tillage* category. Also, to ensure a conservative approach, multiple identical patterns within a single study were normalized into a single pattern, thereby potentially skewing the results toward the less dominant effect, which, for the practices included in both framework application use cases, was the decreasing effect.

The tailored Web of Science (WOS) searches (Module 1) yielded distinct datasets for each case-study component (see Supplementary Data 2, Section 6 for search string information). For the illustrative comparison component (reduced tillage), the initial search retrieved 998 records, of which 10 had empty abstracts; 374 records were subsequently deemed relevant. In parallel, the search addressing the second use case (biochar and crop residue retention) returned 2,816 records.

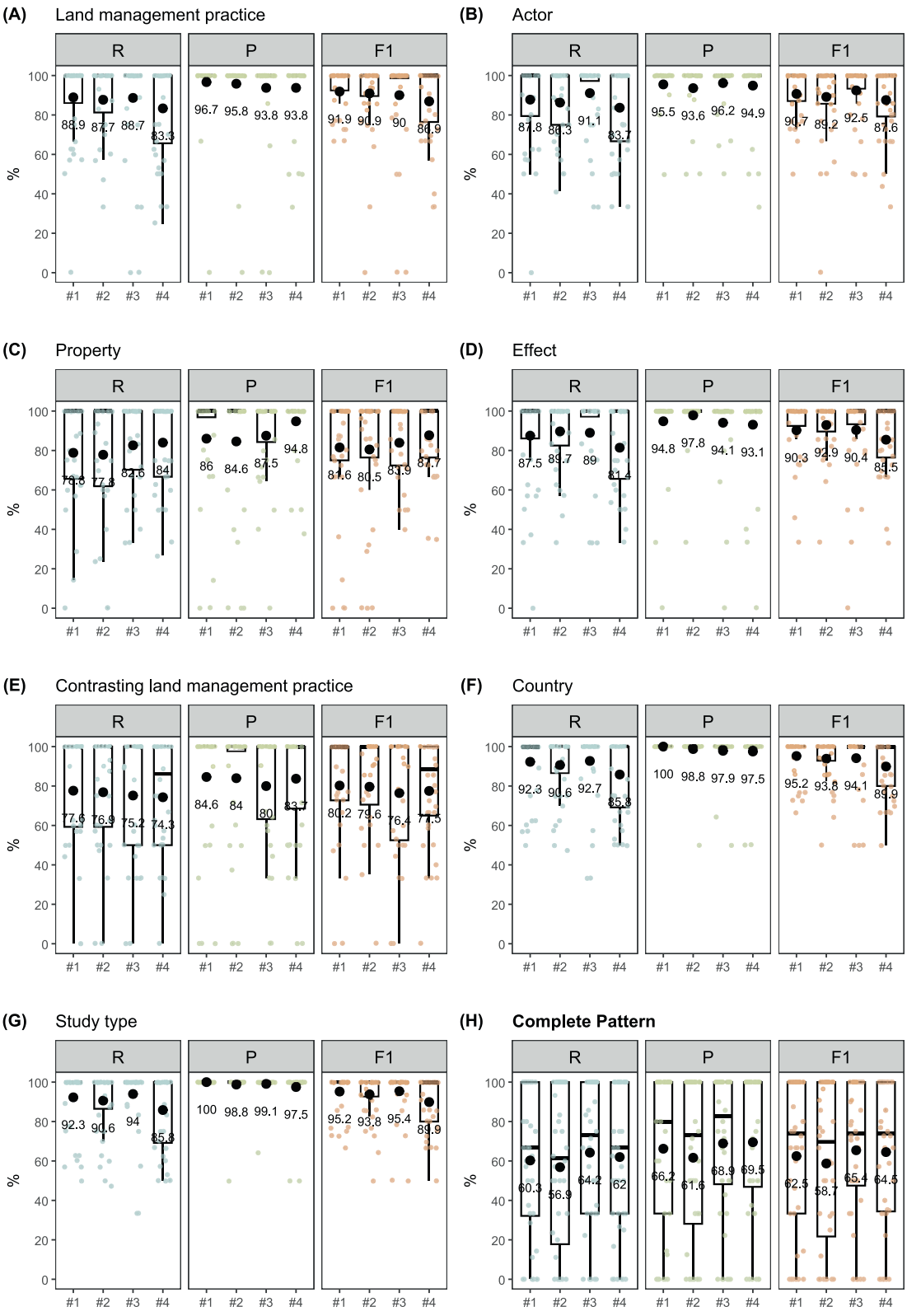


Fig. 3. Quantitative evaluation of large language model (LLM) performance across extraction elements and complete extraction patterns. Performance was assessed against expert annotations using Precision (P), Recall (R), and F1 score (F1) evaluation metrics. Columns correspond to the three evaluation metrics, and rows to the seven extraction elements and the complete extraction pattern: Practice (Land management practice), Actor, Property, Effect, Contrast (Contrasting land management practice), Country, Study type, and Complete pattern (Extraction pattern). The complete extraction pattern represents the exact combination of the seven extraction elements. For each panel, four box plots are shown along the x-axis, corresponding to the four independent expert annotators (#1–#4). Individual abstract-level scores are overlaid as points, and mean values are indicated within each box plot. The y-axis shows metric values as percentages, enabling comparison of extraction performance across annotators, metrics, and knowledge-extraction levels.

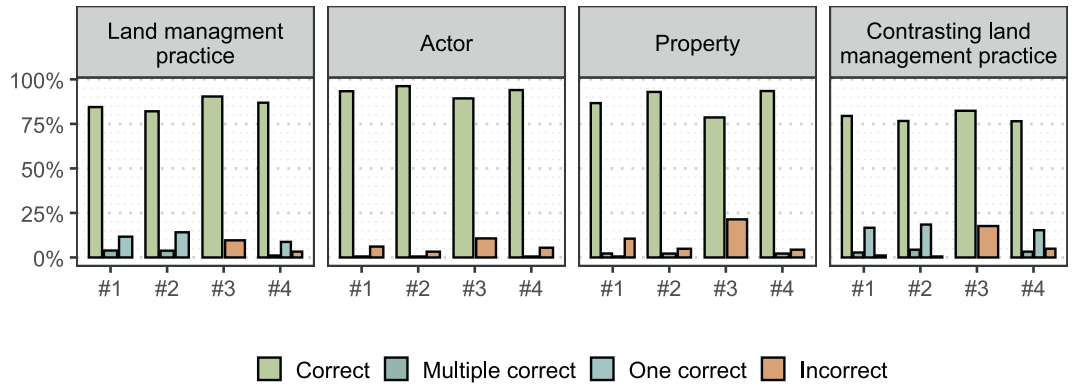


Fig. 4. Bar plot with relative frequencies showing the evaluation of the large language model (LLM) unification process by expert annotators. Each panel corresponds to a key *extraction element*: Land management practice, Actor, Property, and Contrasting land management practice. Bars represent the relative frequency (0–100%) of unification outcomes for each annotator. Outcomes are classified into four categories: *Correct*, *Multiple correct*, *One correct*, and *Incorrect*. The four expert annotators are labelled #1 to #4.

Among these, 2,773 abstracts examined the effects of these amendments on soil fauna, and 773 were ultimately considered relevant. Following the application of the successive workflow modules (Fig. 1), the final output tables generated robust datasets for analysis (Supplementary Data 7). For reduced tillage, the workflow yielded 374 extraction patterns derived from 173 abstracts (Table 1; Supplementary Data 2, Section 10). For the biochar and crop residue component, the process extracted 202 patterns from 118 abstracts (Table 1; Supplementary Data 2, Section 13). This resulted in an average extraction density of 1.91 and 1.71 patterns per abstract for the reduced tillage and biochar/crop residue datasets, respectively, indicating a diverse source of information across the selected literature. Within the latter dataset, biochar-related studies provided a greater volume of evidence (n = 115 abstracts) than crop residue retention (n = 87 patterns), reflecting broader research coverage (see Supplementary Data 2, Section 14). To ensure methodological comparability with the *meta*-data-analysis established by Lori et al. (2025) (Fig. 5A; original figure provided in Supplementary Data 2, Fig. S18), the synthesis for reduced tillage was strictly limited to field experiments, by filtering greenhouse and pot studies (N = 7). In contrast, to maximize evidence retrieval despite varying study availabilities, the synthesis for biochar and crop residue practices encompassed experiments conducted in field, greenhouse, and controlled pot environments. Further details regarding these environmental parameters and the complete dataset distribution are provided in Supplementary Data 2, Sections 11 and 14.

3.3. Illustrative comparison with a reference meta-data-analysis

Regarding the illustrative comparison component, the extracted results for reduced tillage indicated an overall predominance of reported increases in soil fauna responses compared to conventional tillage, consistent with Lori et al. (2025) (Fig. 5A and B). Overall, the general pattern converged across approaches; however, some distinct deviations were apparent, likely reflecting differences in their inferential properties. Conventional *meta*-analysis and LLM-assisted evidence mapping differ fundamentally in what they synthesise and how they aggregate evidence (Dicks et al., 2017). Meta-analysis uses effect sizes as the unit of analysis and typically reports pooled estimates rather than individual primary-study outcomes, which can bias directional heterogeneity (Gurevitch et al., 2018), for example, by indicating an overall positive effect even when many primary studies report decreasing responses. By contrast, the LLM framework relies on pattern-level vote counting, in which each extraction pattern contributes equally regardless of sample size or variance (Friedman, 2001). Thus, while the LLM-assisted evidence mapping can capture the diversity and frequency of reported responses, the *meta*-analytic framework integrates effect sizes using variance-based weighting and heterogeneity modelling, thereby attenuating extreme estimates (either positive or negative), and potentially converging toward a non-significant overall effect size and thus a neutral outcome, as reported by Lori et al. (2025). In addition, because the workflow operates exclusively on abstracts, it is inherently more sensitive to the selective reporting of scientific abstracts, which tend to emphasise salient, statistically significant, or directional findings while compressing null, conditional, or context-dependent results. Accordingly, the LLM-assisted evidence map should be interpreted as a synthesis of reported abstract-level response patterns rather than as a direct representation of underlying ecological effect distributions.

For nematodes, vote counting indicated a predominance of increasing effects (48 increases vs. 25 decreases and 4 neutral outcomes), whereas the *meta*-data-analysis yielded predominantly no-effect estimates, with a dominance of increasing over decreasing effects (656 increases vs. 121 decreases and 2,137 neutral effects). For nematodes' functional groups, a similar pattern was observed in the vote-counting, except for Omnivores and pest-parasitic nematodes (and Protozoa), which showed only increasing patterns.

For Collembola and Acari, vote counting based on LLM-extracted primary studies suggested predominantly increasing responses with a limited number of decreasing responses (Collembola: 18 increases vs. 2 decreases; Acari: 9 increases, 4 decreases, and 1 neutral effect). The *meta*-data-analysis similarly identified an overall increasing trend for both groups; however, decreasing effects observed in the LLM-assisted evidence map were largely replaced by neutral outcomes rather than retained as opposing directional signals. This shift indicates that decreasing responses reported in primary studies may not be sufficiently consistent or precise to influence the pooled effect when *meta*-analytic between-study heterogeneity and variance-based weighting are taken into account. Consequently, the *meta*-data-analytic framework tends to retain the increasing central tendency while attenuating context-specific decreasing responses as neutral. At the same time, because the present workflow is based exclusively on abstracts, the apparent predominance of increasing responses may also partly reflect the preferential reporting of directional or statistically salient outcomes in abstracts. However, the clear dominance of increasing effects seen in the *meta*-data-analysis was correctly captured by our LLM-assisted framework.

This trajectory of increasing responses under reduced tillage practices was equally evident among Oligochaeta; Enchytraeids showed only increasing responses (2 patterns), while Earthworms displayed a robust increasing response, with 92 patterns reporting increases, clearly outnumbering neutral outcomes (3 patterns) and decreasing responses (6 patterns). In the *meta*-data-analysis, only increasing effects were identified; however, because *meta*-analyses report pooled effect sizes, the underlying dataset may still include individual studies showing

Fig. 5. Contextualization of established evidence and practices with limited and fragmented evidence in the global literature on soil management effects on soil fauna. Panel (A) is adapted and reduced from Lori et al. (2025), a systematic review of *meta*-analyses (i.e. a *meta*-data-analysis), and summarizes the overall effect directions reported across available *meta*-analyses for different management practice × contrast combinations. In Lori et al. (2025), direction and significance of effect sizes extracted from individual *meta*-analyses constitute the unit of analysis; the number within each bar indicates the number of observations contributing to the overall effect direction, and the x-axis shows the percentage distribution. Panels (B) and (C) present LLM-assisted evidence mapping based on vote counting from primary studies' abstracts, in which each LLM-extracted pattern from each study contributes one vote to the reported effect direction. In these panels, the numeric value within each bar represents the absolute number of harmonized LLM-extracted patterns supporting that effect direction, whereas the x-axis shows the percentage distribution. Panel (B), highlighted by the green frame, represents the well-established evidence base on the effects of reduced tillage × tillage on soil fauna (macro-, meso-, and microfauna), where no-tillage and reduced tillage were combined in the present study into a single category termed reduced tillage and used as the benchmark component. Panel (C), highlighted by the red frame, represents the practices with limited and fragmented evidence addressed here, namely, biochar × no biochar and crop residue retention × no crop residue retention. Although panel (A) is based on *meta*-analytic effect sizes and panels (B) and (C) on primary-study vote counting, the three panels remain structurally comparable because they are organized using management practice × contrast combinations as columns and biological actor groups as rows. For all panels, the direction of effect is indicated by color (green = increase, red = decrease, grey = neutral/no effect). Patterns are grouped by fauna size class (macrofauna, mesofauna, microfauna, and microbiome). In panel (A), Red-till = reduced tillage and Microf. = microfauna. In panels (B) and (C), opacity reflects evidence robustness, with greater transparency indicating fewer underlying observations for a given practice–actor combination. Mesof. = mesofauna; Mix = soil fauna not assigned to a specific size class.

Table 1
Quantitative summary of the data flow across the knowledge extraction workflow modules (Fig. 1). Columns display the total number of extraction patterns and the number of unique abstracts retained at each stage, stratified by the two workflow application use cases: Reduced tillage (illustrative comparison) and Biochar/Crop residues (application to limited and fragmented evidence).

Knowledge extraction workflow modules	Reduced tillage		Biochar/Crop residues	
	Patterns count	Abstracts count	Patterns count	Abstracts count
Web of Science search (Module 1)	–	1,375	–	2,773
Abstract screening (Module 2)	–	374	–	773
LLM extraction and terminology unification (Module 3)	1,375	374	2,929	773
Knowledge data harmonization (Module 4)	593	255	1,095	518
Final curation (Module 5)	334	173	202	118

decreases that are outweighed through variance- and sample-size-based weighting, yielding an increasing pooled estimate. Likewise, within the present abstract-based workflow, strongly directional increasing responses may be more readily captured when they are preferentially emphasised in abstracts. The observed concordance nonetheless suggests limited between-study heterogeneity and a relatively consistent response direction across studies, resulting in only a minor divergence between the LLM-assisted evidence map and the pooled *meta*-analytic inference.

Broad aggregate groups, specifically *Soil fauna*, *soil microfauna*, *soil mesofauna*, and *soil macrofauna*, were not individually resolved in the study by Lori et al. (2025), and the other macrofaunal groups also revealed predominance of reported increasing responses. Because these broad categories were not directly resolved in the reference *meta*-data-analysis, they should be interpreted here primarily as evidence-mapping outputs that extend the descriptive coverage of the literature, rather than as directly comparable *meta*-analytic groupings. Moreover, because these categories are derived from abstract-level reporting, they may reflect how responses are framed and aggregated in the scientific literature rather than discrete underlying ecological processes.

Taken together, these results demonstrate that an LLM-assisted, vote-counting knowledge-extraction framework can recover broad directional patterns of abstract-reported responses identified by *meta*-analytic references, particularly for biological groups exhibiting consistent responses. However, these extracted patterns should be interpreted as signals of reported scientific discourse that can inform ecological interpretation, but should not be treated uncritically as direct representations of underlying ecological effect distributions. Even so, the illustrative convergence observed despite methodological differences, together with consistent precision and recall performance, provides a reasonable basis for applying the workflow to the biochar and crop-residue knowledge-gap component, within the interpretative constraints of abstract-based evidence extraction.

3.4. Workflow application and ecological evaluation

For the second use case focusing on biochar and crop-residue retention as practices with limited and fragmented evidence, analysis of the results revealed overall biologically meaningful patterns (Fig. 5C), consistent with expected responses to practices that enhance soil organic matter and improve soil physicochemical properties. Due to the limited number of primary studies, these results should be interpreted cautiously. The relatively low number of patterns and abstracts is consistent with previous findings on the effects of other management practices on macro-, meso-, and microfauna (Lori et al., 2025).

To keep focused on the methodological scope of the workflow and its interpretative boundaries, we discuss one illustrative biological actor in detail for each management practice: nematodes for biochar and earthworms for crop-residue retention. The corresponding group-level ecological interpretations for the remaining faunal categories are provided in Supplementary Data 2, Section 17, where they can be consulted as complementary descriptive outputs of the workflow.

Biochar, a carbon-rich material produced through the pyrolysis of organic biomass, can modify soil physicochemical properties and shift microbial and faunal communities (Bolan et al., 2024; Lehmann et al., 2011). Owing to its high porosity and recalcitrant aromatic structure, biochar persists in the soil matrix for extended periods, serving as a stable soil conditioner that creates distinct microhabitats for soil biota (Bolan et al., 2024). However, evidence specific to soil fauna remains limited and fragmented (Lori et al., 2025).

Examining nematodes at the actor level, biochar application showed an overall predominance of increasing responses, with nine abstracts reporting increases, three reporting decreases, and one neutral response (Fig. 5C). This broad actor-level signal is consistent with the comparatively strong performance of the workflow for the actor element, for which recall ranged from 83.7 to 91.1%, precision from 93.6 to 96.2%, and F1 from 87.8 to 92.5% across annotators (Fig. 3B), while actor terminology unification reached 96% correctness (Fig. 4). Applied illustratively to the 13 extracted nematode patterns, these values suggest that the actor identity would likely be correct in nearly all cases, with at most about one extracted pattern affected by actor-level misclassification and perhaps one to two additional nematode patterns remaining unrecovered. However, when evaluated under the deliberately conservative exact-match pattern criterion, recall, precision, and F1 were lower, ranging from 56.9 to 64.2%, 61.6 to 69.5%, and 58.7 to 65.4%, respectively (Fig. 3H). If similar performance held for this subset, roughly 8 to 9 of the 13 extracted complete patterns would be expected to fully match expert annotation, about 4 to 5 might differ in at least one element, and approximately 4 to 7 additional complete patterns could remain unrecovered. These estimates are likely conservative, as some discrepancies reflect over-extraction or partially correct extractions rather than entirely incorrect patterns. Operationally, this means that the dominant actor-level direction is unlikely to reverse, whereas finer trophic-group contrasts remain more sensitive to individual missed or mismatched patterns. These nematode patterns should also be interpreted in light of the methodological differences between vote-counting evidence mapping and quantitative synthesis, which may attenuate directional asymmetries toward weaker or neutral pooled estimates. Disaggregation by trophic group nonetheless revealed a differentiated response structure. The nematode functional groups omnivorous (5 patterns), predatory (5 patterns), and omnivore-predatory (2 patterns) showed exclusively increasing responses. Both bacterial and fungal feeders exhibited a predominance of increasing responses (3 and 2 patterns, respectively), accompanied by decreasing responses (2 and 1 patterns, respectively). This suggests that, in specific contexts, biochar may enhance microbial-driven food web pathways (Yuan et al., 2025; Zhuang et al., 2024). However, the impact of biochar is likely influenced by additional factors, such as resource availability, soil chemical properties, and competitive interactions, which may vary depending on the type of biochar, application rate, and environmental context (Gul et al., 2015). In contrast, plant-feeding nematodes exhibited an overall decrease (10 decreases, 1 increase) (Fig. 5C). This reduction could be attributed to enhanced plant health resulting from biochar application, which can reduce root damage, lower the attractiveness or accessibility of roots to plant-parasitic nematodes (Bolan et al., 2024; Murtaza et al., 2023; Poveda et al., 2021). At the same time, because trophic-level contrasts are based on smaller numbers of abstract-derived patterns, and because agronomically desirable or otherwise salient outcomes may be preferentially highlighted in abstracts, these finer response differences should be interpreted cautiously.

Retaining crop residues, consisting of the above- and belowground remains of cultivated plants, is among the most common and

increasingly adopted organic inputs in agricultural soils (Brichi et al., 2023; Chen et al., 2014). Unlike biochar, which is more chemically stable, residues decompose relatively quickly and act as an immediate source of carbon and nutrients (Brichi et al., 2023; Ruis and Blanco-Canqui, 2017). Their incorporation is generally associated with improvements in soil organic matter, nutrient availability, and structure, although the magnitude and direction of these effects can vary depending on residue type, decomposition dynamics, and management practices such as tillage or mulching (Blanco-Canqui and Lal, 2009; dos Santos Soares et al., 2019). Similar to biochar, retaining crop residues can alter soil physicochemical conditions and thereby influence microbial communities as well as broader soil faunal assemblages (Chen et al., 2014; Ranaivoson et al., 2017; Turmel et al., 2015).

For soil macrofauna overall, residue retention showed a predominance of increasing responses, with 9 patterns reporting increases and 3 reporting decreases, with an increase in abundance reported across multiple groups. Within this category, earthworms were the most frequently extracted actor and likewise showed a clear predominance of increasing responses, with 16 patterns indicating increases and 5 decreases (Fig. 5C). This actor-level signal is consistent with the comparatively strong performance of the workflow for the actor element, for which recall ranged from 83.7 to 91.1%, precision from 93.6 to 96.2%, and F1 from 87.8 to 92.5% across annotators, while actor terminology unification reached 96% correctness. Applied illustratively to the 21 extracted earthworm patterns, these values suggest that actor identification would likely be correct in nearly all cases, with only a small number of earthworm patterns potentially remaining unrecovered. By contrast, under the deliberately conservative exact-match pattern evaluation, recall, precision, and F1 were lower, ranging from 56.9 to 64.2%, 61.6 to 69.5%, and 58.7 to 65.4%, respectively, indicating that some complete patterns may still remain unrecovered or differ from expert annotation in at least one element. Operationally, this suggests that the dominant earthworm-level direction is likely robust, even if the exact balance between increasing and decreasing patterns may be somewhat sensitive to residual extraction error. However, extracted patterns should be interpreted cautiously, as vote-counting based exclusively on abstracts may amplify reported directional asymmetries relative to *meta*-analytic synthesis. Within these constraints, the observed patterns can be explained by the direct contribution of residues to food availability, both as an organic substrate and through the stimulation of microbial populations that form the base of the soil food web (Turmel et al., 2015). In the case of earthworms, residues can serve as a direct food source, supporting increased abundance and activity while simultaneously improving soil structure, organic matter content, and moisture retention, reinforcing their role as ecosystem engineers (Turmel et al., 2015). At the same time, the persistence of some decreasing patterns indicates that these responses are not uniformly positive across contexts and should therefore be interpreted as dominant reported tendencies rather than invariant ecological outcomes.

3.5. Further limitations and challenges

Overall, the benchmarking evaluation showed that the LLM-assisted pattern extraction achieved high performance (Figs. 2 and 3). Nevertheless, extraction evaluations were influenced by a small subset of abstracts exhibiting very low performance, which disproportionately affected the overall evaluation results. These cases highlight that knowledge-extraction performance is sensitive to the clarity and completeness of abstracts, because information that is incompletely reported or presented with linguistic or statistical ambiguity can compromise extraction accuracy. Similarly, abstracts describing highly complex experimental designs remain more challenging to interpret, as essential details are often not explicitly reported, are conveyed in vague terms, or are restricted to selected practice–contrast combinations. Based on repeated expert observations made during the iterative prompt development process and testing across multiple abstract sets, abstract reporting quality was therefore considered a major practical constraint on knowledge mining. However, because this attribution was not isolated through a formal empirical comparison against all model- and prompt-related sources of variation, it should be interpreted as an expert-informed assessment rather than as a directly demonstrated benchmark result; related prospects for improving abstract design to better support LLM-assisted extraction are discussed in Section 3.6.

A further limitation is that, although the workflow can recover contextual information when it is explicitly reported in abstracts, such information is often incomplete and unevenly distributed across the abstract set. As a result, important qualifiers for soil-fauna responses, such as local environmental conditions, sampling design, or temporal context, may be insufficiently represented to support robust context-resolved interpretation at scale. Accordingly, the resulting patterns are best understood as broad evidence-mapping outputs, rather than context-complete ecological synthesis.

A related limitation concerns the balance between the management and biological dimensions of the extraction framework. Although the workflow includes explicit biological elements, particularly actor, property, and method or measurement, the biological component depends more directly on the level of detail reported in abstracts than the management component, which was more tightly structured to preserve practice definitions and contrasts. As a result, when abstracts omit taxonomic specificity, ecological guilds, or response metrics, biologically meaningful heterogeneity may be compressed during harmonisation and aggregation. Accordingly, downstream interpretation is generally more robust for broad biological actors and reported directional tendencies than for finer taxonomic or metric-resolved ecological inference.

From a methodological perspective, it is also important to note that workflow performance remains partly dependent on the specific LLM selected for knowledge extraction and terminology unification. Although recent OpenAI models tested during workflow development showed broadly similar behaviour, the present study did not include a formal robustness assessment across different high-capacity LLM architectures or alignment strategies. Thus, while the highly explicit prompt design likely reduces model-related variability, we cannot rule out that different models may resolve abstracts with limited reporting clarity and completeness differently, thereby altering extracted patterns. This issue should be addressed in future cross-model benchmarking studies.

Also from a methodological perspective, prompt robustness and reproducibility were not evaluated through a formal sensitivity framework. Specifically, we did not quantify stability across repeated runs, sensitivity to prompt paraphrasing, or the effects of modifying the predefined unification lists, and these aspects should therefore be addressed in future methodological benchmarking.

An important boundary of the present workflow is that it is not designed for formal detection–attribution analysis (Gonzalez et al., 2023; Schrodt et al., 2025) and should not be used as a standalone basis in assessment or policy-oriented contexts. Although the extraction structure records directional relationships between management practices and soil-fauna responses, these outputs represent abstract-reported patterns rather than formally detected trends or attributed causal effects. The framework does not include key elements required for detection–attribution, such as explicit causal modelling, temporal baselines, time-series structure, or counterfactual comparisons. Instead, it provides a structured evidence map of reported directional outcomes that can support hypothesis generation and the screening of broad tendencies in societally relevant management contexts, but not formal attribution claims or decision-grade environmental assessment.

3.6. Applications and future directions

A promising direction for further development is the selective incorporation of article sections beyond abstracts to improve contextual resolution while preserving the scalability that makes the present

workflow useful. The exclusive use of abstracts is not only a limitation but also a methodological advantage: abstracts are short, relatively standardised, and consistently available, making them especially suitable for rapid, large-scale screening. At the same time, adding selected sections could improve the capture of qualifiers, contrasts, and boundary conditions that are often compressed in abstracts (Schuemie et al., 2004; Westergaard et al., 2018). Rather than moving directly toward unrestricted full-text extraction, a more practical path may be to augment abstracts selectively with highly informative sections or metadata fields. In this regard, empirical sections appear particularly valuable, especially the Results section, because it contains substantial study-specific information, while the Methods and Results together may better anchor extraction in reported findings rather than broader interpretative discourse (Penning de Vries et al., 2020; Schuemie et al., 2004). Conclusions may also represent a useful intermediate target because they often summarise findings in a relatively condensed form (Cohen et al., 2010). In parallel, enriched metadata fields, especially expanded keywords, could further improve evidence detectability at low additional cost (Duyx et al., 2019). Accordingly, future versions of the workflow could test targeted hybrid strategies that preserve the efficiency of abstract-based screening while improving contextual resolution. Where contextual descriptors are reported with sufficient consistency, such strategies could also support stratified evidence mapping or subset analyses, for example, by study type, taxonomic group, temporal scope, or other abstract-derived qualifiers, thereby improving ecological specificity and reducing overgeneralisation.

Given the limitations associated with abstract-only extraction, a promising future direction is to retain conventional abstracts while complementing them with more standardised, machine-readable reporting formats. One practical option would be to adopt structured abstracts or parallel metadata repositories in which authors deposit harmonisable descriptors alongside the narrative abstract, including taxa, response metrics, management contrasts, study type, and other key contextual variables. Such formats could remain compatible with conventional scientific communication while providing higher-quality inputs for rapid synthesis tools, reducing ambiguity, improving biological specificity, and preserving scalability. LLMs could further support the generation, validation, and refinement of these structured reporting layers. In the longer term, such enriched reporting structures could also allow LLM-assisted workflows to move incrementally toward detection-oriented applications by retaining the publication year and study period for each extracted pattern, linking patterns to explicit study-design and environmental-context metadata, and selectively integrating comparator or control information when reported, thereby enabling more structured temporal and counterfactual interpretation while preserving scalability.

The presented workflow was developed within the context of agriculture and soil sciences, with a primary focus on management practices and their interaction with soil fauna, in line with Lori et al. (2025). The accompanying workflow application focused on three commonly used management practices and three major groups of soil biological actors, and served as a targeted demonstration within this context. While the use cases presented in this study served as a demonstration, the workflow is readily extensible to address other areas with limited and fragmented evidence, such as the effects of pest management on soil biota or the links between soil vegetation management and fauna. For that purpose, only Modules 5 and 6, as well as the initial search string (Module 1), were tailored to the specifics of the presented use cases. In contrast, Modules 2 to 4 were designed to be methodologically generalizable across research questions within this disciplinary context. At the same time, transferability across contexts depends not only on the LLM itself, but also on the adaptation of the conceptual elements, controlled categories, and curation logic embedded in the workflow to the structure of the target knowledge domain. Moreover, the abstract-based, LLM-assisted data-mining strategy presented here is well-suited for broader application across ecology and environmental sciences, as illustrated in Supplementary Data 2, Table S1, particularly in fields where large and fragmented bodies of literature pose challenges for synthesis, such as pollination ecology, invasive species dynamics, biodiversity–ecosystem functioning research, or studies on the impacts of climate change on ecological communities. Given the workflow's fully transparent, open-access design, applications beyond the soil and agricultural domain can be supported through adaptations ranging from straightforward refinements to more moderate adjustments of the prompt-chain text and code, depending on the structure of the target extraction patterns and the specificity of the research context.

4. Conclusions

Overall, this study shows that an LLM-assisted workflow can efficiently extract, structure, and map reported evidence on soil-fauna responses to agricultural management from large corpora of scientific abstracts with limited manual input. Rather than providing quantitative evidence synthesis in the *meta*-analytic sense, the workflow generates a rapid evidence map of abstract-reported directional outcomes that is most informative for identifying broad reported tendencies, inconsistencies, and knowledge gaps across fragmented literatures. Within this scope, crop-residue retention emerged as a predominantly beneficial signal in the reported literature, particularly for earthworms and nematodes, whereas biochar was characterised by a more heterogeneous, context-dependent pattern. At the same time, repeated expert-guided development and benchmarking indicated that the interpretative value of the outputs is strongly shaped by the scope, quality, and completeness of abstract reporting, even though this influence was not identified as a formal benchmark result. Accordingly, the proposed framework is best understood as a complementary tool for organising, screening, and prioritising dispersed ecological evidence, and for supporting hypothesis generation and subsequent in-depth review, rather than as a substitute for full-text synthesis, effect-size-based *meta*-analysis, or decision-grade ecological inference.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and Gemini (Google) to develop and iteratively refine the LLM-assisted prompt chain and workflow for knowledge extraction, and to assist with language editing (grammar and scientific style). After using these tools, the author(s) reviewed, edited, and verified the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the European Union's Horizon Europe Research and Innovation Program, under Grant Agreement: 101091010, Project BENCHMARKS. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract no. 22.00619. Ricardo Leitão was supported by a PhD scholarship from the Portuguese Foundation for Science and Technology (Fundação para a Ciência e a Tecnologia, FCT), reference 2022.11630. BD. We thank Alessio Corti for his assistance in evaluating the LLM's subjective performance. Research was also supported by the Slovenian Research and Innovation Agency (ARIS), (grant P2-0103).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2026.117890. Supplementary material includes the lists of scientific papers used as LLM inputs across the different stages of framework development, extended methodological details, expert evaluation guidelines, supporting tables and figures, additional results, raw evaluation outputs, LLM-generated extractions, and the data frames underlying the workflow assessment and complementary use cases analyses.

Data availability

All data is available as Supplementary files. All analysis programs are available on GitHub: https://github.com/vpodpecan/LEPAMTIC

References

Bardgett, R.D., van der Putten, W.H., 2014. Belowground biodiversity and ecosystem functioning. Nature 515, 505–511. https://doi.org/10.1038/nature13855.

Birhane, A., Kasirzadeh, A., Leslie, D., Wachter, S., 2023. Science in the age of large language models. Nat. Rev. Phys. 5, 277–280. https://doi.org/10.1038/s42254-023-00581-4.

Blanchy, G., Albrecht, L., Koestel, J., Garré, S., 2023. Potential of natural language processing for metadata extraction from environmental scientific publications. Soil 9, 155–168. https://doi.org/10.5194/soil-9-155-2023.

Blanco-Canqui, H., Lal, R., 2009. Crop Residue Removal Impacts on Soil Productivity and Environmental Quality. Crit. Rev. Plant Sci. 28, 139–163. https://doi.org/10.1080/07352680902776507.

Bolan, S., Sharma, S., Mukherjee, S., Kumar, M., Rao, C.S., Nataraj, K.C., Singh, G., Vinu, A., Bhowmik, A., Sharma, H., El-Naggar, A., Chang, S.X., Hou, D., Rinklebe, J., Wang, H., Siddique, K.H.M., Abbott, L.K., Kirkham, M.B., Bolan, N., 2024. Biochar modulating soil biological health: a review. Sci. Total Environ. 914, 169585. https://doi.org/10.1016/j.scitotenv.2023.169585.

Brichi, L., Fernandes, J.V.M., Silva, B.M., Vizú, J.D.F, Junior, J.N, Cherubin, M.R., 2023. Organic residues and their impact on soil health, crop production and sustainable agriculture: a review including bibliographic analysis. Soil Use and Management 39, 686–706. https://doi.org/10.1111/sum.12892.

Chen, B., Liu, E., Tian, Q., Yan, C., Zhang, Y., 2014. Soil nitrogen dynamics and crop residues. A Review. Agron. Sustain. Dev. 34, 429–442. https://doi.org/10.1007/s13593-014-0207-8.

Chen, B., Zhang, Z., Langrené, N., Zhu, S., 2025. Unleashing the potential of prompt engineering for large language models. PATTER 6. https://doi.org/10.1016/j.patter.2025.101260.

Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C., Hunter, L.E., 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinf. 11, 492. https://doi.org/10.1186/1471-2105-11-492.

Creamer, R.E., Barel, J.M., Bongiorno, G., Zwetsloot, M.J., 2022. The life of soils: Integrating the who and how of multifunctionality. Soil Biol. Biochem. 166, 108561. https://doi.org/10.1016/j.soilbio.2022.108561.

Delgado-Baquerizo, M., Reich, P.B., Trivedi, C., Eldridge, D.J., Abades, S., Alfaro, F.D., Bastida, F., Berhe, A.A., Cutler, N.A., Gallardo, A., García-Velázquez, L., Hart, S.C., Hayes, P.E., He, J.-Z., Hseu, Z.-Y., Hu, H.-W., Kirchmair, M., Neuhauser, S., Pérez, C. A., Reed, S.C., Santos, F., Sullivan, B.W., Trivedi, P., Wang, J.-T., Weber-Grullon, L., Williams, M.A., Singh, B.K., 2020. Multiple elements of soil biodiversity drive ecosystem functions across biomes. Nat. Ecol. Evol. 4, 210–220. https://doi.org/10.1038/s41559-019-1084-y.

Dicks, L.V., Haddaway, N., Hernández-Morcillo, M., Mattsson, B., Randall, N., Failler, P., Ferretti, J., Livoreil, B., Saarikoski, H., Santamaria, L., Rodela, R., Velizarova, E., Wittmer, H., 2017. Knowledge synthesis for environmental decisions: an evaluation of existing methods, and guidance for their selection, use and development – a report from the EKLIPSE project (Project report). EKLIPSE.

dos Santos Soares, D., Ramos, M.L.G., Marchão, R.L., Maciel, G.A., de Oliveira, A.D., Malaquias, J.V., de Carvalho, A.M., 2019. How diversity of crop residues in long-term no-tillage systems affect chemical and microbiological soil properties. Soil Tillage Res. 194, 104316. https://doi.org/10.1016/j.still.2019.104316.

Duyx, B., Swaen, G.M.H., Urlings, M.J.E., Bouter, L.M., Zeegers, M.P., 2019. The strong focus on positive results in abstracts may cause bias in systematic reviews: a case study on abstract reporting bias. Syst. Rev. 8, 174. https://doi.org/10.1186/s13643-019-1082-9.

Esteve, A., 2024. Copyright and Open Access to Scientific Publishing. IIC 55, 901–926. https://doi.org/10.1007/s40319-024-01479-z.

Food and Agriculture Organization of the United Nations (FAO), 2024. Time to address global soil monitoring? ITPS Soil Letters (Fact sheet). FAO, Rome. Available at: https://openknowledge.fao.org/handle/20.500.14283/cd1591en (accessed 20 Oct 2025).

Friedman, L., 2001. Why vote-count reviews don't count. Biol. Psychiatry 49, 161–162. https://doi.org/10.1016/S0006-3223(00)01075-1.

Galli, C., Gavrilova, A.V., Calciolari, E., 2025. Large Language Models in Systematic Review Screening: Opportunities, challenges, and Methodological Considerations. Information 16, 378. https://doi.org/10.3390/info16050378.

Gao, J., Wang, D., 2024. Quantifying the use and potential benefits of artificial intelligence in scientific research. Nat. Hum. Behav. 8, 2281–2292. https://doi.org/10.1038/s41562-024-02020-5.

Glasziou, P., Meats, E., Heneghan, C., Shepperd, S., 2008. What is missing from descriptions of treatment in trials and reviews? BMJ 336, 1472–1474. https://doi.org/10.1136/bmj.39590.732037.47.

Gohr, C., Rodríguez, G., Belomestnykh, S., Berg-Moelleken, D., Chauhan, N., Engler, J.-O., Heydebreck, L.V., Hintz, M.J., Kretschmer, M., Krügermeier, C., Meinberg, J., Rau, A.-L., Schwenck, C., Aoulkadi, I., Poll, S., Frank, E., Creutzig, F., Lemke, O., Maushart, M., Pfendtner-Heise, J., Rathgens, J., von Wehrden, H., 2025. Artificial intelligence in sustainable development research. Nat Sustain 8, 970–978. https://doi.org/10.1038/s41893-025-01598-6.

Gonzalez, A., Chase, J.M., O'Connor, M.I., 2023. A framework for the detection and attribution of biodiversity change. Philos. Trans. R. Soc. Lond. B Biol. Sci. 378, 20220182. https://doi.org/10.1098/rstb.2022.0182.

Gul, S., Whalen, J.K., Thomas, B.W., Sachdeva, V., Deng, H., 2015. Physico-chemical properties and microbial responses in biochar-amended soils: Mechanisms and future directions. Agr Ecosyst Environ 206, 46–59. https://doi.org/10.1016/j.agee.2015.03.015.

Gurevitch, J., Koricheva, J., Nakagawa, S., Stewart, G., 2018. Meta-analysis and the science of research synthesis. Nature 555, 175–182. https://doi.org/10.1038/nature25753.

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, 7. https://doi.org/10.1186/s13750-018-0121-7.

Idan, D., Einav, S., 2025. Primer on large language models: an educational overview for intensivists. Crit. Care 29, 238. https://doi.org/10.1186/s13054-025-05479-4.

Ji, X., Wu, X., Deng, R., Yang, Y., Wang, A., Zhu, Y., 2025. Utilizing large language models for identifying future research opportunities in environmental science. J. Environ. Manage. 373, 123667. https://doi.org/10.1016/j.jenvman.2024.123667.

Keck, F., Peller, T., Alther, R., Barouillet, C., Blackman, R., Capo, E., Chonova, T., Couton, M., Fehlinger, L., Kirschner, D., Knüsel, M., Muneret, L., Oester, R., Tapolczai, K., Zhang, H., Altermatt, F., 2025. The global human impact on biodiversity. Nature 641, 395–400. https://doi.org/10.1038/s41586-025-08752-2.

Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nat. Rev. Earth Environ. 1, 544–553. https://doi.org/10.1038/s43017-020-0080-8.

Lehmann, J., Rillig, M.C., Thies, J., Masiello, C.A., Hockaday, W.C., Crowley, D., 2011. Biochar effects on soil biota – A review. Soil Biology and Biochemistry, 19th International Symposium on Environmental Biogeochemistry 43, 1812–1836. https://doi.org/10.1016/j.soilbio.2011.04.022.

Li, J., Gao, Y., Yang, Y., Bai, Y., Zhou, X., Li, Y., Sun, H., Liu, Y., Si, X., Ye, Y., Wu, Y., Lin, Y., Xu, B., Ren, B., Feng, C., Huang, H., 2025. Fundamental Capabilities and applications of Large Language Models: a Survey. ACM Comput. Surv. 58, Article 38. https://doi.org/10.1145/3735632.

Lori, M., Leitao, R., Felix, D., Imbert, C., Corti, A., Cunha, L., Symanczik, S., Bünemann, E., Creamer, R., Vazquez, C., 2025. Response of soil biota to agricultural management practices: a systematic quantitative meta-data-analysis and method selection framework. Soil Biology and Biochemistry 207, 109815. https://doi.org/10.1016/j.soilbio.2025.109815.

Lori, M., Symnaczik, S., Mäder, P., De Deyn, G., Gattinger, A., 2017. Organic farming enhances soil microbial abundance and activity—A meta-analysis and meta-regression. PLOS ONE 12, e0180442. https://doi.org/10.1371/journal.pone.0180442.

Margoni, T., Kretschmer, M., 2022. A deeper look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. GRUR Int 71, 685–701. https://doi.org/10.1093/grurint/ikac054.

McFadden, I.R., Sendek, A., Brosse, M., Bach, P.M., Baity-Jesi, M., Bolliger, J., Bollmann, K., Brockerhoff, E.G., Donati, G., Gebert, F., Ghosh, S., Ho, H.-C., Khaliq, I., Lever, J.J., Logar, I., Moor, H., Odermatt, D., Pellissier, L., de Queiroz, L.J., Rixen, C., Schuwirth, N., Shipley, J.R., Twining, C.W., Vitasse, Y., Vorburger, C., Wong, M.K.L., Zimmermann, N.E., Seehausen, O., Gossner, M.M., Matthews, B., Graham, C.H., Altermatt, F., Narwani, A., 2023. Linking human impacts to community processes in terrestrial and freshwater ecosystems. Ecology Letters 26, 203–218. https://doi.org/10.1111/ele.14153.

Murtaza, G., Ahmed, Z., Eldin, S.M., Ali, B., Bawazeer, S., Usman, M., Iqbal, R., Neupane, D., Ullah, A., Khan, A., Hassan, M.U., Ali, I., Tariq, A., 2023. Biochar-Soil-Plant interactions: a cross talk for sustainable agriculture under changing climate. Front. Environ. Sci. 11. https://doi.org/10.3389/fenvs.2023.1059449.

Navigli, R., Conia, S., Ross, B., 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality 15, 10:1–10:21. https://doi.org/10.1145/3597307.

Nie, Q., Liu, T., 2025. Large language models: Tools for new environmental decision-making. Journal of Environmental Management 375, 124373. https://doi.org/10.1016/j.jenvman.2025.124373.

Pai, L., Gao, W., Dong, W., Ai, L., Gong, Z., Huang, S., Li, Z., Hoque, E., Hirschberg, J., Zhang, Y., 2024. A survey on open information extraction from rule-based model to large language model. Findings of the Association for Computational Linguistics: EMNLP 2024, 9586–9608. https://doi.org/10.18653/v1/2024.findings-emnlp.560.

Penning de Vries, B.B.L., van Smeden, M., Rosendaal, F.R., Groenwold, R.H.H., 2020. Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. Journal of Clinical Epidemiology 121, 55–61. https://doi.org/10.1016/j.jclinepi.2020.01.009.

Poveda, J., Martínez-Gómez, Á., Fenoll, C., Escobar, C., 2021. The use of Biochar for Plant Pathogen Control. Phytopathology® 111, 1490–1499. https://doi.org/10.1094/PHYTO-06-20-0248-RVW.

Ranaivoson, L., Naudin, K., Ripoche, A., Affholder, F., Rabeharisoa, L., Corbeels, M., 2017. Agro-ecological functions of crop residues under conservation agriculture. A Review. Agron. Sustain. Dev. 37, 26. https://doi.org/10.1007/s13593-017-0432-z.

Ruis, S.J., Blanco-Canqui, H., 2017. Cover Crops could Offset Crop Residue Removal Effects on Soil Carbon and Other Properties: a Review. Agronomy Journal 109, 1785–1805. https://doi.org/10.2134/agronj2016.12.0735.

Schrodt, F., Beck, M., Estopinan, J., Bowler, D.E., Fontaine, C., Gaüzère, P., Goury, R., Grenié, M., Martins, I.S., Morueta-Holme, N., Santini, L., Hedde, M., Martin, G., Porcher, E., Si-Moussi, S., Tzivanopoulos, M., Vernham, G., Violle, C., Thuiller, W., 2025. Advancing causal inference in ecology: Pathways for biodiversity change detection and attribution. Methods in Ecology and Evolution 16, 2276–2304. https://doi.org/10.1111/2041-210X.70131.

Schuemie, M.J., Weeber, M., Schijvenaars, B.J.A., van Mulligen, E.M., van der Eijk, C.C., Jelier, R., Mons, B., Kors, J.A., 2004. Distribution of information in biomedical abstracts and full-text publications. Bioinformatics 20, 2597–2604. https://doi.org/10.1093/bioinformatics/bth291.

Tsiafouli, M.A., Thébault, E., Sgardelis, S.P., de Ruiter, P.C., van der Putten, W.H., Birkhofer, K., Hemerik, L., de Vries, F.T., Bardgett, R.D., Brady, M.V., Bjornlund, L., Jørgensen, H.B., Christensen, S., Hertefeldt, T.D., Hotes, S., Gera Hol, W.H., Frouz, J., Liiri, M., Mortimer, S.R., Setälä, H., Tzanopoulos, J., Uteseny, K., Pižl, V., Stary, J., Wolters, V., Hedlund, K., 2015. Intensive agriculture reduces soil biodiversity across Europe. Global Change Biology 21, 973–985. https://doi.org/10.1111/gcb.12752.

Turmel, M.-S., Speratti, A., Baudron, F., Verhulst, N., Govaerts, B., 2015. Crop residue management and soil health: a systems analysis. Agricultural Systems, Biomass Use Trade-Offs in Cereal Cropping Systems: Lessons and Implications from the Developing World 134, 6–16. https://doi.org/10.1016/j.agsy.2014.05.009.

Wadoux, A.-M.-J.-C., 2025. Artificial intelligence in soil science. European Journal of Soil Science 76, e70080. https://doi.org/10.1111/ejss.70080.

Wall, D.H., Nielsen, U.N., Six, J., 2015. Soil biodiversity and human health. Nature 528, 69–76. https://doi.org/10.1038/nature15744.

van den Brand, T., 2024. ggh4x: Hacks for 'ggplot2'. R package version 0.3.0. https://CRAN.R-project.org/package=ggh4x.

Wan, M., Safavi, T., Jauhar, S.K., Kim, Y., Counts, S., Neville, J., Suri, S., Shah, C., White, R.W., Yang, L., Andersen, R., Buscher, G., Joshi, D., Rangan, N., 2024. TnT-LLM: Text Mining at Scale with Large Language Models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24. Association for Computing Machinery, New York, NY, USA, pp. 5836–5847. https://doi.org/10.1145/3637528.3671647.

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L.J., Brunak, S., 2018. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLOS Computational Biology 14, e1005962. https://doi.org/10.1371/journal.pcbi.1005962.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. Journal of Open Source Software 4, 1686. https://doi.org/10.21105/joss.01686.

Wu, Y., 2025. Large language models and the future of soil health: Bridging knowledge gaps through scalable semantic intelligence. Soil Advances 4, 100065. https://doi.org/10.1016/j.soilad.2025.100065.

Yuan, F., Hu, X., Shiping, W., Ma, L., Zhang, Z., Song, L., Li, H., Yao, B., Zhou, H., 2025. Biochar addition affects the microfood web structure of soil animals: a meta-analysis. CATENA 249, 108647. https://doi.org/10.1016/j.catena.2024.108647.

Zhang, Y., Khan, S.A., Mahmud, A., Yang, H., Lavin, A., Levin, M., Frey, J., Dunnmon, J., Evans, J., Bundy, A., Dzeroski, S., Tegner, J., Zenil, H., 2025. Exploring the role of large language models in the scientific method: from hypothesis to discovery. npj. Artif. Intell. 1, 14. https://doi.org/10.1038/s44387-025-00019-5.

Zhao, Y., Goto, S., 2025. Can Frontier LLMs Replace Annotators in Biomedical Text Mining? Analyzing Challenges and Exploring Solutions. https://doi.org/10.48550/arXiv.2503.03261.

Zhuang, W., Zhao, C., Zhang, Y., Yang, Z., Li, G., Su, L., Zhang, S., 2024. Synergistic application of biochar with organic fertilizer positively impacts the soil micro-food web in sandy loam soils. European Journal of Soil Biology 123, 103680. https://doi.org/10.1016/j.ejsobi.2024.103680.

Zwetsloot, M.J., Bongiorno, G., Barel, J.M., di Lonardo, D.P., Creamer, R.E., 2022. A flexible selection tool for the inclusion of soil biology methods in the assessment of soil multifunctionality. Soil Biology and Biochemistry 166, 108514. https://doi.org/10.1016/j.soilbio.2021.108514.