

Review

Open Access

Large language models for environmental research: systematic literature screening, relational knowledge mining, and quantitative data extraction

Yunlong Li^{1,2}, Jing Guo^{1,2*} and Wei Shi^{1,2}

Citation: Li Y; Guo J; Shi W. Large language models for environmental research: systematic literature screening, relational knowledge mining, and quantitative data extraction. *AI Environ.* 2026, 1(3): 139-150.

Received: 31 May 2026

Revised: 26 July 2026

Accepted: 20 August 2026

Published online: 10 September 2026

Abstract

Environmental research is facing a growing challenge of data abundance coupled with knowledge fragmentation. Evidence on pollutant exposure, toxicological effects, and pollution control is rapidly expanding, but it remains scattered across heterogeneous literature sources and reported in inconsistent formats. This fragmentation limits the throughput and efficiency of traditional manual literature screening and data extraction. Large language models (LLMs) have demonstrated methodological potential in natural language understanding, contextual integration, and structured output generation, offering new opportunities for automated environmental evidence extraction and integration. This review examines the capabilities of LLMs in three interconnected tasks: systematic literature screening, relational knowledge mining, and quantitative data extraction. Specifically, this review summarizes progress in how LLMs can support the identification of relevant studies, the extraction of relationships among multi-level entities, and the conversion of unstructured texts into structured quantitative datasets. Furthermore, applications across environmental exposure, toxicological effects, and pollution control are discussed, and key limitations regarding evaluation frameworks, result traceability, and expert oversight are highlighted. Current evidence suggests that while LLMs can substantially improve data extraction efficiency and structured data generation, their reliable deployment still requires clear task definition, structured constraints, rule-based validation, and human review. In conclusion, this review highlights that LLMs serve as an important methodological bridge between rapidly expanding environmental literature and reusable environmental knowledge systems.

Keywords: Large language models, Environmental research, Systematic literature screening, Relational knowledge mining, Quantitative data extraction

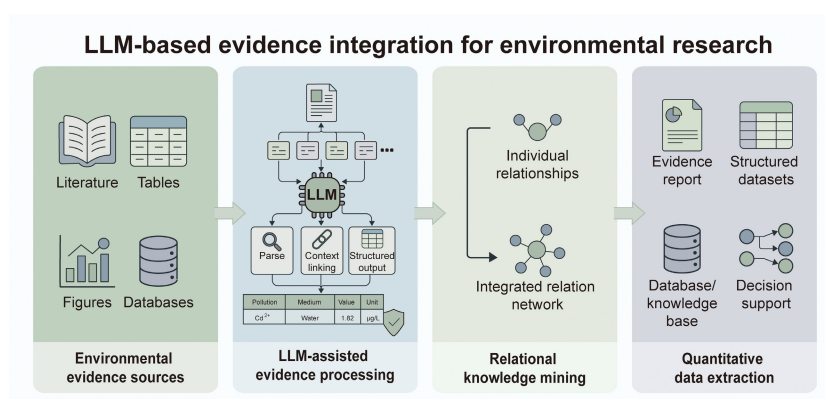
Highlights

- LLMs aid environmental evidence synthesis via literature screening, relational knowledge mining, and quantitative data extraction.
- LLMs should augment, not replace, experts in constrained, traceable, and validated environmental workflows.
- Task-specific benchmarks are needed to evaluate LLMs and enable cross-study comparison in environmental research.

* Correspondence: Jing Guo (jingguo@nju.edu.cn)

Full list of author information is available at the end of the article.

Graphical abstract



Introduction

Environmental research is rapidly expanding, yet it is increasingly challenged by knowledge fragmentation^[1]. With advances in high-throughput detection and assessment technologies, evidence on contaminant occurrence, distribution, and potential risks has grown rapidly^[2,3]. Systematically integrating these fragmented lines of evidence is essential for understanding contaminant fate, assessing environmental and health risks, and developing mitigation strategies^[1,4–7]. However, much of this evidence remains embedded in heterogeneous and unstructured literature sources, including narrative text, tables, figures, supplementary materials, and inconsistent reporting formats^[8,9]. Moreover, environmental data are also inherently multidimensional, simultaneously involving chemical identity, environmental media, exposure scenarios, test systems, effect endpoints, treatment processes, and degradation parameters^[10]. This complexity creates substantial challenges for extracting, linking, and organizing meaningful relationships across multi-level environmental entities. As a result, environmental research increasingly faces a paradox in which data are abundant, but reusable and structured evidence remains difficult to obtain^[11].

Traditional approaches for evidence integration in environmental research rely heavily on manual literature screening, manual data extraction, and expert annotation^[12,13]. These approaches remain indispensable for ensuring scientific accuracy, but they are often labor-intensive, time-consuming, prone to subjective bias, and difficult to scale with the rapid expansion of environmental literature^[13–15]. The challenge is particularly pronounced in environmental studies because relevant information is often multidimensional and context-dependent. For example, a single study may report pollutant concentrations across different environmental matrices, sampling locations, exposure durations, biological species, toxicological endpoints, treatment processes, and experimental conditions. Quantitative values such as environmental concentrations, effect thresholds, bioaccumulation factors, removal efficiencies, degradation rates, and risk quotients may be expressed using different units, formats, and contextual descriptions^[16]. These characteristics make environmental literature difficult to screen, compare, and integrate using conventional manual and rule-based text mining approaches alone^[12,17].

In recent years, the development of LLMs has provided new methodological pathways for processing unstructured data in environmental research. Owing to their capabilities in natural language understanding, contextual reasoning, semantic integration, and structured output generation, LLMs can support the transformation

of unstructured scientific text into organized and reusable information. Compared with conventional keyword-based retrieval or rule-based text mining, LLMs are better suited to handling complex inclusion criteria, integrating long-text context, extracting cross-sentence relationships, and generating structured outputs^[18]. These capabilities are highly relevant to environmental research, where critical information is frequently scattered across diverse textual expressions and requires interpretation within specific experimental or environmental contexts. Recent reviews have shown that LLMs are increasingly being incorporated into workflows for retrieval assistance, literature screening, data extraction, and evidence integration^[19–21]. Meanwhile, growing attention has been given to evaluation consistency, reporting transparency, and human review mechanisms, which are essential for ensuring the reliability and reproducibility of LLM-assisted evidence integration^[22]. However, critical challenges remain in adapting general-purpose LLM capabilities to evidence integration workflows in environmental research. Unlike generic information extraction, environmental evidence synthesis requires the coordinated identification of relevant studies, organization of cross-entity and process-oriented relationships, and reconstruction of structured quantitative records from heterogeneous literature sources.

This review provides a framework for understanding how LLMs can support environmental evidence integration across three interconnected tasks: systematic literature screening, relational knowledge mining, and quantitative data extraction. Specifically, it first examines how LLMs support systematic literature screening through the identification of relevant studies. It then evaluates relational knowledge mining through the organization of cross-entity and process-oriented relationships. Finally, it assesses quantitative data extraction through the reconstruction of structured records from heterogeneous literature sources. Across these tasks, we synthesize applications in environmental exposure, toxicological effects, and pollution control while assessing methodological strengths, limitations, validation requirements, and applicable boundaries. By integrating these perspectives, this review aims to clarify how LLMs can support traceable, validated, and reusable environmental knowledge systems and identify priorities for future research.

LLM-based evidence integration for environmental research

Building on the challenges outlined above, this review conceptualizes LLM-assisted environmental evidence integration as three

interconnected methodological steps: systematic literature screening, relational knowledge mining, and quantitative data extraction (Fig. 1). These steps represent interrelated stages in transforming fragmented environmental literature into reusable evidence. Systematic literature screening defines the scope and quality of the evidence base by identifying studies that meet complex inclusion criteria related to pollutants, environmental matrices, organisms, endpoints, treatment technologies, and exposure scenarios. Relational knowledge mining then organizes cross-entity associations and process-oriented mechanisms, including source–pathway–receptor chains, toxicity pathways, pollutant transformation routes, and process–condition–performance relationships. Quantitative data extraction further reconstructs record-level numerical information from texts, tables, figures, and semi-structured sources while preserving the correspondence among pollutants, environmental or experimental conditions, indicators, values, units, and evidence provenance.

Importantly, these three steps should be understood as an integrated workflow rather than as isolated applications. Literature screening determines which studies enter the evidence base; relational knowledge mining provides the contextual, mechanistic, and process-level structure needed to interpret environmental evidence; and quantitative data extraction converts selected evidence into comparable records for database construction, statistical analysis, modeling, risk assessment, and decision support^[23]. This workflow can be implemented across several major areas of environmental research, particularly environmental exposure assessment, toxicological effect evaluation, and pollution control. When guided by predefined schemas, domain knowledge constraints, validation rules, traceable evidence sources, and expert review, LLM-enabled workflows can help transform unstructured environmental literature into reliable and reusable knowledge systems for environmental research^[24].

LLM-assisted precise literature screening

Systematic literature screening serves as the first step in evidence organization by accurately identifying highly relevant studies. A typical LLM-assisted literature screening process can be summarized as five

steps: candidate literature retrieval, deduplication and title–abstract organization, relevance pre-screening and priority ranking based on inclusion/exclusion criteria, manual full-text eligibility assessment, and final inclusion of studies into an evidence base (Fig. 2). The resulting evidence base then provides the input for downstream data extraction and synthesis. LLMs primarily contribute to the third step—relevance judgment and priority assignment at the title–abstract level—whereas final inclusion decisions remain the responsibility of human reviewers^[21]. Although this workflow can improve transparency and traceability, its application in environmental research still faces challenges related to broad retrieval scope, complex inclusion criteria, and the cross-disciplinary distribution of relevant evidence. Factors such as pollutant type, environmental medium, exposure or reaction conditions, endpoint category, and study design often simultaneously constitute the basis for inclusion judgments. In such cases, conventional keyword and Boolean searches rely primarily on explicit terms and predefined logical combinations, which may retrieve low-specificity results when keywords occur outside the intended context and may overlook relevant studies that use synonyms, abbreviations, or implicit expressions^[17]. By contrast, LLMs can interpret titles and abstracts contextually, recognize synonymous or implicit expressions, and apply multiple inclusion and exclusion criteria simultaneously. These capabilities can improve specificity by excluding contextually irrelevant records and improve sensitivity by identifying relevant studies that do not contain the predefined keywords. LLMs can also process and prioritize records in batches, reducing manual screening workload while leaving final eligibility decisions to human reviewers^[22].

Recent studies have demonstrated the feasibility of LLM-assisted screening in environmental evidence synthesis. Zuo et al. combined ChatGPT with domain knowledge to screen 920 records on land use/land cover and fecal coliform contamination in rivers^[25]. Their study showed improved screening efficiency under complex inclusion and exclusion criteria spanning hydrology, public health, and landscape research, suggesting that LLMs can support front-end pre-screening in cross-disciplinary environmental evidence synthesis. Nykvist et al. further found that GPT-4 achieved 100% recall in title–abstract screening of nearly 12,000 records in an

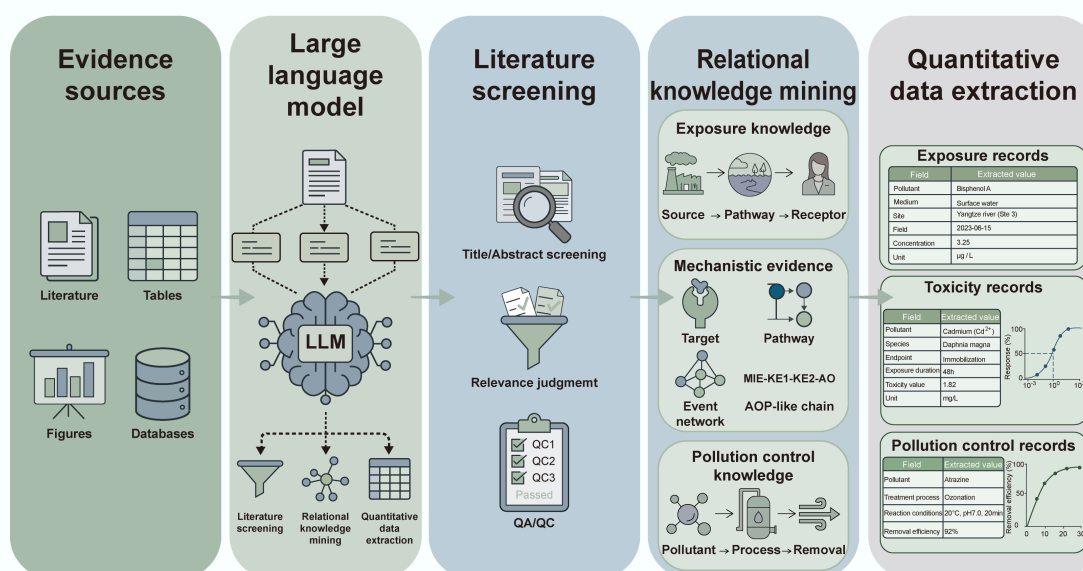


Fig. 1 LLM-assisted evidence integration framework in environmental research.

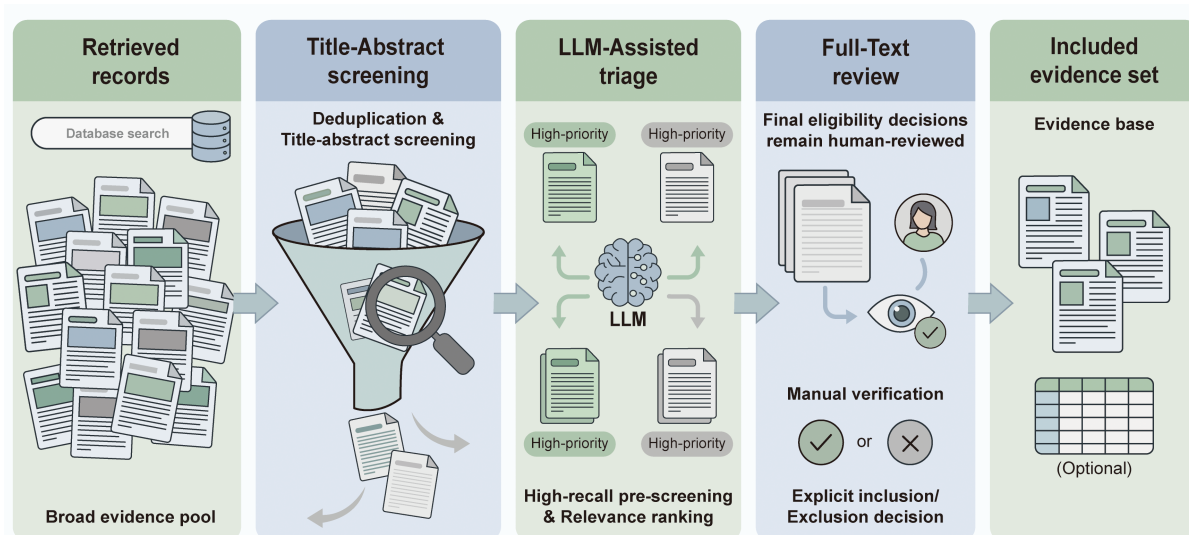


Fig. 2 LLM-assisted literature screening in environmental research.

environmental systematic evidence synthesis study, saving approximately 50% of manual screening time at the corresponding threshold setting, demonstrating the methodological feasibility of LLMs for large-scale record pre-screening in environmental research^[26]. Similarly, Nicholson Thomas et al. found that LLMs could complete relevant literature selection under complex concepts and multidimensional inclusion criteria in a systematic review of ecosystem condition indicators, suggesting that the applicable scope of such methods is not limited to pollution effects or toxicological questions but can extend to screening tasks across the broader environmental domain^[27]. These studies collectively demonstrate that the methodological feasibility of LLM-assisted screening has been validated across different environmental research contexts, with applicability that is cross-thematic and cross-disciplinary.

Because the erroneous exclusion of relevant studies can substantially bias environmental evidence synthesis, evaluations of LLM-assisted literature screening should prioritize recall and the traceability of screening errors, while also reporting precision, specificity, F1 score, and changes in manual workload^[21]. This evaluation principle has cross-domain applicability in data synthesis methodology; the automated framework proposed by Marshall et al. in biomedical evidence synthesis automation similarly emphasizes that automated screening tools should prioritize high recall, with precision optimization as a secondary objective^[28]. Accordingly, LLMs should be positioned as front-end pre-screening and priority-ranking tools that reduce manual screening burden and support full-text eligibility assessment and downstream data extraction, whereas final inclusion decisions and the exclusion of potentially relevant studies should remain under human review.

LLM-assisted environmental relational knowledge mining

Definition and characteristics of relational knowledge

Extracting and organizing the relationships embedded in selected literature is essential for synthesizing environmental evidence. Evidence related to environmental exposure, toxicological effects, and pollution control all involves extensive relational knowledge. In

exposure assessment, the core relationship is the "source–pathway–receptor" chain linking pollutant sources, environmental occurrence, transport pathways, and receptor exposure scenarios. In toxicological studies, relational knowledge is reflected in the hierarchical causal pathway connecting pollutants, molecular initiating events, key events, and adverse outcomes^[29,30]. In pollution control studies, the core relationship is the process–condition–performance correspondence among pollutant types, treatment processes, reaction conditions, transformation products, and removal efficiencies. These relationships differ in types, hierarchical structures, and organizational purposes, posing considerable challenges for the systematic organization of multidimensional relational knowledge. Therefore, relational knowledge mining goes beyond simple entity recognition. Its key objective is to extract directionality, hierarchy, contextual constraints, and process logic from dispersed literature and organize them into verifiable, retrievable, and reusable knowledge structures.

Existing reviews in the environmental domain indicate that the practical value of LLMs in this direction is primarily concentrated in relationship integration, knowledge organization, and structured expression^[19]. As shown in Fig. 3, LLM-based relational knowledge mining typically involves four steps: acquiring raw text from relevant literature; identifying and extracting key qualitative elements such as pollutants, environmental scenarios, hazard characteristics, action processes, treatment measures, and outcomes^[31]; on this basis, establishing associations among elements according to preset knowledge patterns and organizing them into relationship pairs, event chains, or multi-tuple structures; and further representing them as knowledge graphs, relationship networks, or data tables.

From entity recognition to relationship extraction

Currently, relational knowledge mining in environmental research is still primarily built on named entity recognition, entity co-occurrence analysis, and rule-driven relationship extraction. Such methods can support the identification of elements including pollutants, environmental media, hazard characteristics, treatment processes, and outcomes, but their determination of relationships typically relies on local text windows, preset trigger words, or fixed syntactic patterns,

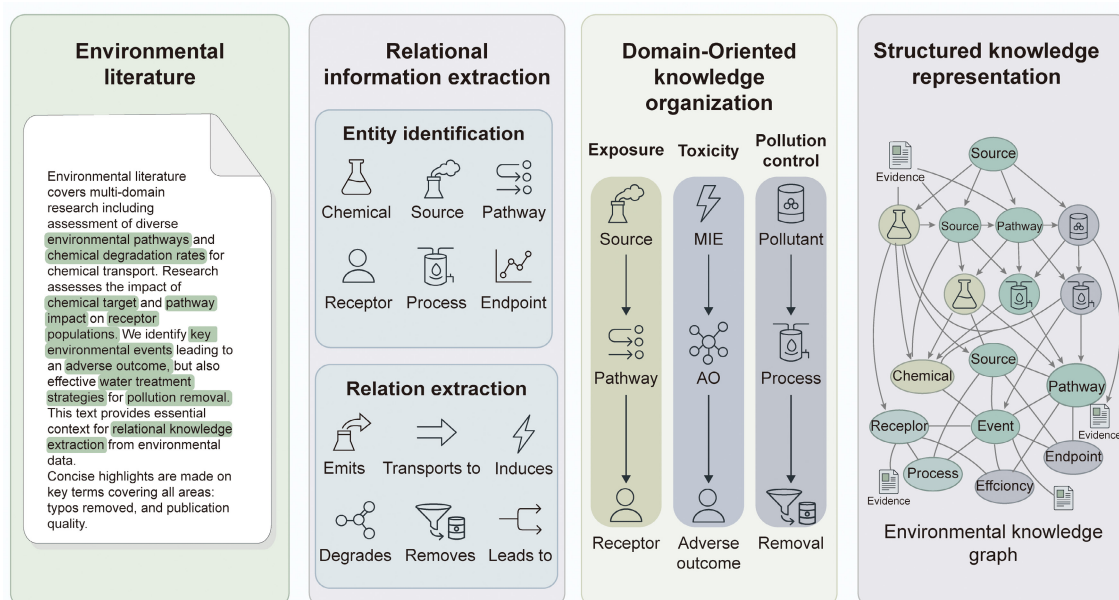


Fig. 3 LLM-assisted relational knowledge mining in environmental research.

resulting in clear limitations in cross-sentence coreference resolution, conditional context differentiation, relationship direction determination, and cross-literature data merging. Compared with traditional methods, the primary methodological advantage of LLMs lies in long-context integration capability and the ability to extract and merge multiple types of environmental relationships under pattern constraints^[18]. Relational knowledge in environmental research spans multiple levels including pollutant source–transport–fate, degradation and transformation processes, exposure–hazard associations, and toxicity mechanisms; representative applications of LLMs have emerged across all these directions. Regarding pollutant source and transport relationships, work by Chen et al. on oil and gas industry emission assessment demonstrated that LLMs can systematically extract associations among emission sources, activity levels, and concentration responses from unstructured industry reports, providing a data foundation for constructing source–exposure relationship networks for atmospheric pollutants^[32]. Regarding pollution treatment and control, Xu et al. in water and wastewater management research showed that LLMs can extract multidimensional relationships among pollutant types, treatment conditions, and removal efficiencies from process literature, supporting knowledge organization of pollutant degradation and transformation processes^[33]. Regarding exposure–hazard associations, Li et al. in aquatic risk assessment research noted that LLMs can not only identify local relationships but also integrate dispersed data on pollution sources, exposure backgrounds, and hazard clues, providing relational knowledge support for prioritizing emerging contaminants^[34]. Regarding toxicity mechanism relationships, Liang et al. established a GPT-based text mining workflow for chemicals modulating oxidative stress, compiling 17,780 records covering 2,558 compounds from 7,166 publications, demonstrating that LLMs can be used for systematic compilation of mechanistic relationship information and construction of specialized resources^[35]. It is thus evident that relational knowledge mining in environmental research has moved beyond local relationship identification and begun to enter the level of relationship integration and knowledge organization; however, such results are fundamentally still candidate qualitative data, and relationship direction, hierarchical

classification, and evidence strength determination still require integration with knowledge frameworks and human verification.

Network-based representation of relational knowledge

Among the various types of relational knowledge described above, some have advanced further toward network-based representation, with objectives no longer limited to identifying a set of relationship pairs but oriented toward knowledge objects that are computable, accessible, and reusable. Among the various types of environmental relational knowledge, the Adverse Outcome Pathway (AOP) framework, with its well-defined hierarchical structure and clear relationship types, represents one of the most mature scenarios for network-based representation research; the molecular initiating events, key events, and their associations involved are well suited as representative cases for organizing local relationship extraction into networked knowledge resources^[29]. Shi et al. systematically discussed the potential and limitations of LLMs in AOP construction, noting that they can already assist to some extent in identifying molecular initiating events, key events, and adverse outcomes, and generate preliminary structures approximating expert organizational logic^[36]. Jaylet et al. developed AOP-helpFinder 3.0, achieving automated identification of stressor–event and event–event relationships and network visualization, demonstrating that under explicit knowledge pattern constraints, qualitative relationships can be further transformed into queryable graph-structured resources^[37]. Carvillio et al. combined text mining with systems biology to identify associations between bisphenol S (BPS) and AOP-related events, further organizing them into adipogenesis- and obesity-related pathways^[38]. These studies demonstrate that under explicit knowledge pattern constraints, AOP-related mechanistic relationships can be further transformed from candidate relationships into queryable graph-structured resources^[39].

Relational knowledge mining is expanding toward more constrained and computable implementation pathways. James et al. attempted one-shot Biological Expression Language (BEL) relationship extraction using LLMs, directly transcribing causal relationships

from literature into computable mechanistic networks for comparison with expert-constructed biomedical knowledge graphs^[40]. El Khettari et al. proposed a two-stage information extraction framework based on ontological constraints, first identifying candidate processes and then determining process–process relationships at the document level, thereby organizing implicit mechanistic processes into structured triples^[41]. The implementation of mechanistic knowledge organization has extended from direct relationship extraction to relationship determination, knowledge representation, and conceptual expansion under ontological constraints; however, these methods are still largely at the exploratory stage, and their results likewise require interpretation and use under explicit knowledge frameworks and human review conditions.

Current application value and limitations

The value of LLMs in environmental relational knowledge mining is primarily reflected in three aspects: improving the ability to identify candidate relationships in complex text; facilitating the transformation of dispersed narratives into structured knowledge objects; and providing foundational interfaces for knowledge graph construction, relationship network analysis, and natural language retrieval^[42]. However, the boundaries of this class of applications also require clarification. First, model-generated relationships are not equivalent to verified causal conclusions; textual co-occurrence, semantic relevance, and environmental process associations must still be strictly distinguished^[43]. Second, relationship types, hierarchical constraints, and evidence strength expression vary across different tasks, and model outputs show substantially higher dependence on knowledge patterns than general text tasks. Third, model outputs may still exhibit systematic biases in relationship direction determination, cross-literature fusion, source tracking, and knowledge deduplication^[44,45]. Current evidence suggests that LLMs are most feasible when used to generate and organize candidate relationships within predefined ontologies, graph patterns, or task-specific rules. Final validation, archiving, and interpretation should still rely on rule-based checks and expert review.

LLM-assisted structured quantitative data extraction

Definition and extraction challenges for environmental quantitative data

Structured quantitative data are the basis for making environmental evidence reusable, comparable, and suitable for database construction. In environmental research, quantitative data refer not only to numerical values themselves, but also to complete record-level information that links a pollutant with its corresponding scenario or condition, indicator, value, unit, and evidence source. Therefore, the core task of quantitative data extraction is to reconstruct structured records from dispersed literature while preserving the correspondence among these fields. All three directions—pollutant

exposure, toxicological effects, and pollution control—involve large volumes of quantitative data, and the data forms in each direction are considerably complex. Exposure records require simultaneous annotation of pollutant identity, environmental medium, sampling conditions, and concentration values, with monitoring data across media and regions highly heterogeneous in units and nomenclature^[2]. Toxicity records require maintaining complete correspondence among chemical identity, test species, exposure conditions, and endpoint values, with the diversity of test systems and standards further increasing the difficulty of cross-study comparison^[46]. Degradation efficiency records must simultaneously cover target pollutants, treatment systems, reaction conditions, and performance indicators (removal rate, rate constants, etc.), with the high context-dependence of process types and operating parameters making record-level correspondence especially critical. The core field composition of the three types of quantitative data is shown in Table 1.

Despite originating from different domains, environmental quantitative data share several common extraction challenges. First, record completeness is essential. Unlike qualitative conclusions, quantitative data render a record unusable for subsequent comparative analysis or database entry if any field is missing (e.g., exposure concentration lacking sampling medium, toxicity data lacking test species, degradation data lacking reaction conditions). Olker et al. noted in ECOTOX database construction that record-level toxicity data must maintain correspondence among chemical, species, study design, test conditions, toxicity results, and derived endpoints^[10]. Schilling-Wilhelmi et al. similarly emphasized in a review of chemical data extraction that the key to chemical data extraction lies in maintaining record-level correspondence among reaction conditions, substance characteristics, and performance indicators^[47]. Second, terminology and units are highly heterogeneous. Concentration units, pollutant nomenclature, medium descriptions, and process types vary widely across publications, requiring external knowledge resources to support standardization. Third, numerical values are highly context-dependent. The correct interpretation of numerical values depends on background information such as sampling methods, detection limits, and reaction conditions; extracting values while ignoring context may lead to misuse.

Given these requirements, LLM-assisted quantitative data extraction should not be treated as simple identification of numerical values, but as a record-level reconstruction process guided by predefined schemas and quality-control rules^[47]. As shown in Fig. 4, LLM-based quantitative data extraction typically involves four steps: data source identification, quantitative information extraction, standardization and quality control, and structured output. The core objective is to extract information on chemicals, test systems, exposure conditions, toxicity endpoints, and effect values from unstructured materials, and to form structured records that are comparable, verifiable, and database-ready after standardization.

From entity recognition to structured extraction

The core technical challenge of quantitative data extraction lies in binding multiple fields dispersed across unstructured text into

Table 1 Core field composition of record-level quantitative data in environmental research

Data type	Chemical identifier	Scenario/system	Condition	Core indicator	Value & unit	Quality control
Exposure record	Pollutant name, CAS number, chemical category	Environmental medium, sampling location	Sampling time, season, detection method, detection limit	Concentration level	μg/L, ng/m ³ , mg/kg, etc.	Sample size, QC method, data source
Toxicity record	CAS, InChIKey, SMILES, or database ID	Test species, life stage, tissue/organ	Exposure concentration, duration, route	EC50, LC50, NOEC, LOEC	mg/L, μg/kg, etc.	Test standard, laboratory, literature source
Degradation record	Target pollutant name, initial concentration	Treatment system, catalyst type	pH, temperature, catalyst loading, light intensity	Removal rate, rate constant k, half-life t _{1/2} , mineralization rate	%, min ⁻¹ , h, etc.	Reaction scale, matrix type, literature source

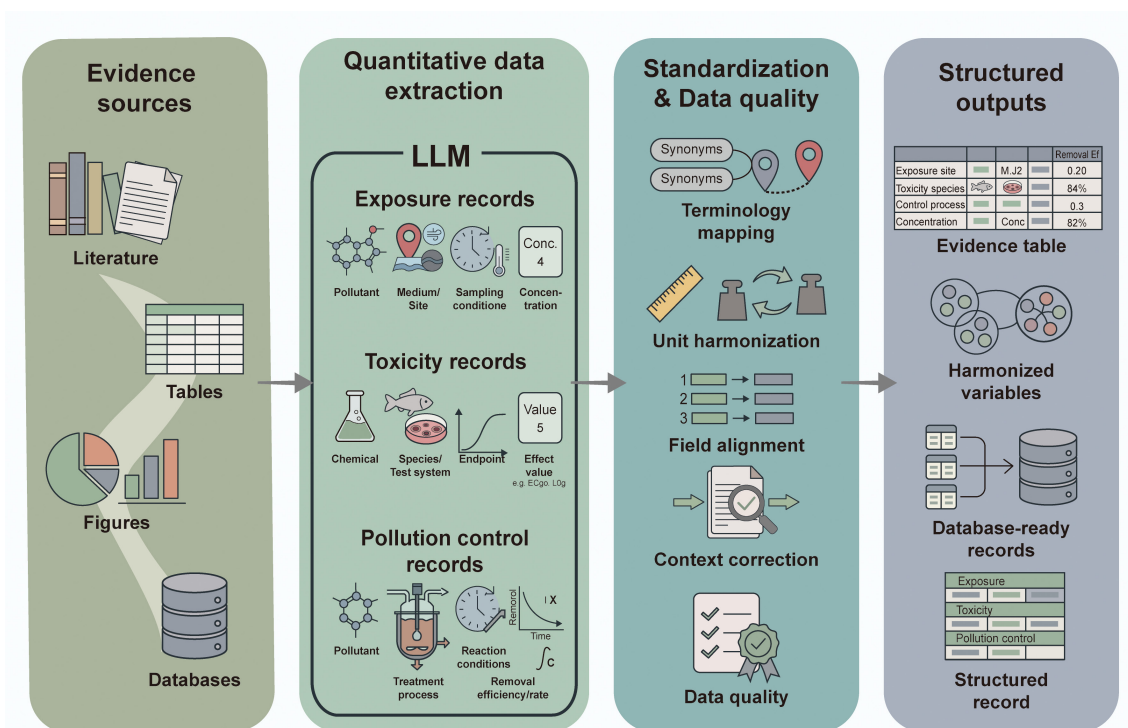


Fig. 4 LLM-assisted quantitative data extraction in environmental research.

complete records, rather than isolating individual terms or values. Existing research has begun advancing quantitative information extraction from field-level identification to record-level structured output^[18]. In exposure data extraction, the Literature-to-Dataset (L2D) pipeline developed by Cheng et al. specifically addresses complex, user-customized data extraction needs in environmental research, demonstrating LLM advantages over traditional NLP methods in multi-field joint extraction and user-defined data structures^[48]. Shah-Mohammadi et al. constructed an LLM-based exposure data extraction pipeline for extracting and normalizing metadata from full-text PDFs in exposure-health literature, including pollutant type, exposure medium, measurement method, and concentration range^[49]. In toxicity data extraction, LLMs have been used to reconstruct structured toxicity records from complex literature sources. Sonnenburg et al. applied LLMs to information extraction for bisphenol A bioactivity, demonstrating that extraction performance is highly dependent on task definition, output constraints, and evaluation design^[50]. Moreira-Filho et al. integrated document parsing and LLMs into the KNIME workflow platform, achieving end-to-end automation from PDF documents to structured toxicity records, demonstrating that toxicity data extraction has extended from text field extraction to structured reconstruction in complex document environments^[9]. Paula et al. established a large-scale meta-analysis pipeline for nanotoxicology literature using NLP, demonstrating LLM capabilities in handling nanotoxicology triplet extraction across nanomaterials, biological systems, and toxicity endpoints^[51]. Ha et al. further used LLM-extracted structured toxicity data directly for machine learning model training, achieving a complete pipeline from literature mining to toxicity prediction^[52]. In degradation efficiency data extraction, Wu et al. applied LLMs to large-scale data extraction from single-atom catalyst literature in the advanced oxidation processes (AOPs) domain, effectively identifying complex correspondences among catalyst types, target pollutants, reaction conditions, and removal efficiencies^[53].

Together, these studies indicate that LLM-assisted quantitative extraction should be understood as structured record reconstruction rather than the collection of independent fields. Its reliability depends on maintaining correct field pairing, contextual completeness, and standardized output formats.

Multimodal LLMs for figure-based data extraction

Quantitative evidence in environmental literature is not limited to text and tables. Degradation kinetics, dose-response relationships, temporal trends, and toxicological endpoints are often reported only in figures and may be missed by text-only workflows. Multimodal LLMs can jointly interpret figure images and contextual information, including captions, axes, legends, panel labels, and methods text, and reconstruct the recovered values in predefined structured fields. Moreira-Filho et al. implemented an image mode within a KNIME workflow to process graphical content and non-linear document layouts using extraction prompts and structured output schemas^[9]. Park et al. combined full text with figures, tables, captions, legends, and methods sections to extract nanotoxicity data. In a benchmark of 17 oxide nanotoxicity articles, GPT-5.4 achieved a mean extraction F1 score of 96.1%, although the F1 score for quantitative assay results was lower at 86.5%. The workflow was subsequently applied to 821 articles and generated 16,786 structured records^[54].

However, high average performance does not ensure accurate numerical recovery from figures. Wang et al. reported accuracies above 85% for plot recognition and domain-knowledge interpretation with GPT-4V, but below 60% for color perception, with additional weaknesses in positional inference and counting^[55]. In CharXiv, GPT-4o achieved 47.1% accuracy on reasoning questions based on realistic scientific charts, compared with 80.5% for human evaluators^[56]. Overlapping curves, similar colors or markers, error bars, logarithmic or dual axes, multi-panel layouts, and incomplete

labels can therefore lead to incorrect pairing of values with pollutants, conditions, or treatment groups. Figure-derived records should retain the source figure, panel, caption, axis, and legend information and be validated against manually digitized values or available source data using explicit numerical tolerances. At present, multimodal LLMs are better suited to candidate data recovery and structured reconstruction than fully autonomous figure digitization.

Domain adaptation strategies: fine-tuning, retrieval-augmented generation, and prompt engineering

Although LLMs show clear potential for record-level quantitative data extraction, the performance of general-purpose LLMs on specialized environmental science text is constrained by domain terminology density, heterogeneous reporting formats, and numerical expression conventions^[47]. Therefore, domain adaptation is essential to improve extraction accuracy and output consistency. Common strategies include task-specific fine-tuning, retrieval-augmented generation (RAG), and prompt engineering.

Task-specific fine-tuning enables models to learn domain terminology, annotation boundaries, and task-specific output patterns. It is particularly useful for large-scale, repetitive extraction tasks when sufficiently large and high-quality annotated datasets are available. In a closely related chemical text-mining task, Zhang et al. compared prompt-engineered GPT-3.5 with the same model after task-specific fine-tuning for extracting structured metal-organic framework synthesis information, including reagent amounts, solvent volumes, reaction temperatures, and reaction times. Fine-tuning increased exact-match accuracy by more than 20 percentage points for both single- and multi-reaction paragraphs, reaching 82.7% and 68.8%, respectively^[57]. This controlled comparison provides direct evidence that task-specific fine-tuning can improve structured quantitative data extraction, although its performance in environmental literature requires domain-specific validation.

RAG improves evidence grounding by retrieving relevant passages or domain knowledge before generation, thereby reducing dependence on the model's internal knowledge and allowing the supporting evidence to be inspected and the external knowledge base to be updated without retraining the generative model^[58]. Zou et al. developed ESGReveal by integrating report preprocessing, targeted retrieval, and ESG metadata within an enhanced RAG framework. In ablation tests with GPT-4, enhanced RAG increased data-extraction accuracy from 52.2% to 74.0%, while incorporating ESG-specific knowledge further increased it to 76.9%^[59].

Prompt engineering can improve extraction consistency by decomposing complex tasks, specifying output formats, and introducing iterative verification without changing model parameters. Cheng and Zhang developed the L2D pipeline by combining domain-specific multistep prompting with self-validation, achieving 93% accuracy in extracting eight biodegradation-related features from 79 articles^[48]. Similarly, Polak and Morgan developed ChatExtract, which uses engineered prompts and iterative follow-up questions to identify, extract, and verify quantitative materials data, achieving precision and recall close to 90% with GPT-4^[60].

Overall, task-specific fine-tuning, RAG, and prompt engineering have demonstrated potential to improve quantitative data extraction from specialized texts. Fine-tuning primarily improves adaptation to domain terminology and task-specific output patterns; RAG strengthens evidence grounding and facilitates knowledge updating; and prompt engineering enhances consistency, traceability, and numerical verification. However, their applicability varies according

to task scale, terminology density, document structure, and the availability of annotation and validation resources.

Quality control and external constraints

Quality control in quantitative data extraction is a critical step in constructing reliable and reusable databases. Because model outputs may contain numerical misextraction, unit errors, field misalignment, unsupported values, and inconsistent terminology, external constraints are required beyond the extraction model itself^[24]. Ontological constraints can support the consistency and traceability of term recognition. Specifically, they map synonymous expressions to standardized entities through controlled vocabularies and canonical identifiers. Predefined entity types, attributes, and relationships can then be used to identify undefined terms, missing attributes, and schema-incompatible records. Ontological relations and semantic rules can further flag logically inconsistent combinations. Hodgson et al. evaluated OntoGPT, an ontology-grounded AI tool for environmental evidence synthesis, and found higher agreement with human reviewers when extracting standardized information than when identifying study-specific or interpretation-intensive attributes^[61]. As an example of external validation, Bauer et al. combined literature mining with AI language models to derive hepatotoxicity scores for more than 50,000 compounds. Validation against a drug-induced liver injury (DILI) reference set yielded an AUC of 0.85 for the LLM approach and 0.87 when combined with text mining and word embeddings^[62].

A practical deployment mode is a collaborative framework combining model extraction, rule validation, and human review: LLMs handle initial extraction and record reconstruction; rule systems handle numerical range checking, unit consistency verification, and required field completeness checking; and human review handles final quality control and anomalous records processing. Specifically, a post-extraction quality-control workflow comprises three stages: data validation, consistency alignment, and error correction. First, each extracted record should retain its supporting text and page, table, or figure location to enable evidence tracing. Automated validation should further check data types, plausible ranges, cross-field relationships, and duplicate records. Second, consistency alignment should map chemical names, environmental media, endpoints, and units to controlled vocabularies and standardized formats. Unit conversion should preserve the original reported value and unit alongside the harmonized value to maintain traceability. Third, records failing these checks should be re-extracted from the original context or corrected using deterministic rules when the supporting evidence is unambiguous. Because exhaustive expert review is infeasible for large and structurally complex datasets, automated checks should cover all records, whereas expert review should prioritize low-confidence extractions, rule conflicts, implausible values, and records with potentially high-impact errors. Low-risk records can instead be evaluated through random sampling to estimate the residual error rate. This risk-based workflow shifts the role of experts from checking every record to defining validation rules, resolving ambiguous cases, and auditing recurring error patterns. Finally, the proportion of records reviewed, correction rate, and residual error rate should be reported to make the workload and reliability of quality control transparent. Source evidence, correction decisions, and model and rule versions should also be retained for auditing. Overall, the practical positioning of LLMs in quantitative data extraction is most appropriately as an assistive tool rather than a fully automated system, with term mapping, parameter standardization, and quality control still requiring external knowledge resources and rule systems.

Current application value and limitations

Overall, the methodological value of LLMs in quantitative data extraction for environmental research is primarily reflected in joint field extraction, structured data reconstruction, and database update support. However, related research also clearly reveals their boundaries: although LLMs achieve relatively high accuracy on discrete information and process substantially faster than humans, they perform markedly worse on coordinate conversion, time ranges, and some continuous quantitative information^[63]. LLMs depend on relatively stable document structures, well-defined data dictionaries, and strict field constraints; once tasks shift toward complex literature, cross-paragraph information pairing, or multi-source context integration, error risk increases substantially. For environmental research, the more operationally feasible deployment mode remains a collaborative framework combining model extraction, rule validation, and human review, rather than treating model outputs directly as final data ready for database entry or evaluation.

Challenges and future directions of LLM applications in environmental research

Task evaluation and result reliability

A key limitation of LLM applications in environmental research is that task evaluation frameworks, result reliability, and cross-study comparability remain insufficient^[19]. A systematic review by Raeissi et al. on generative LLMs in environmental science showed that while existing research has covered multiple directions including literature analysis, information extraction, and knowledge support, there is substantial variation in task definition, evaluation metrics, and validation approaches, and no stable consensus framework has yet emerged^[19]. Kleinstreuer et al. noted that the application value of artificial intelligence in environmental research should be assessed through evaluation systems consistent with real-world tasks, rather than relying solely on single-metric accuracy performance^[64]. On this basis, differentiated evaluation systems should be established for different tasks: literature screening should report recall, precision, specificity, F1, and changes in manual workload; mechanistic knowledge mining should focus on relationship correctness, directional consistency, and source traceability; and quantitative data extraction should distinguish between extraction and matching accuracy, and identify key error types including numerical mis-extraction, unit errors, and field mismatching.

Reproducibility, traceability, and reporting standards

Methodological transparency and result auditability are fundamental requirements for deploying LLMs in environmental research. Hartung et al. noted in a review of validation of AI-based new approach methods that applicability domain definition, external validation, error analysis, and independent replication are basic prerequisites for AI methods entering environmental applications^[65]; Gant et al. similarly emphasized, in the context of AI integration into chemical risk assessment, that transparent task definitions, clearly documented input sources, traceable output pathways, and explicit human review mechanisms are essential prerequisites for building regulatory confidence and practical trust^[66]. Accordingly, LLM research in the environmental domain should at minimum explicitly report model name, exact version, access date, prompting strategy, input preprocessing, validation sample sources, and major error types, and should preserve source tracking and human verification mechanisms wherever possible.

Model-specific variability and legal/ethical considerations

Reporting the exact model name, version, and access date is particularly important because model selection can materially affect task performance. Performance differences among LLMs reflect not only the model itself but also differences in training strategies, context capacity, reasoning ability, and task-specific configurations. Model performance should therefore be interpreted within the specific dataset, prompt, adaptation strategy, and evaluation setting rather than inferred from the model family alone. Comparative studies have demonstrated the influence of model selection on environmental and toxicological information extraction tasks. Using the same dataset and prompt, Nykvist et al. found substantial differences between GPT-3.5 and GPT-4 in environmental literature screening^[26]. Moreira-Filho et al. compared proprietary and open-weight models under a common toxicological data-extraction workflow. The evaluated models included Claude 3 Opus, Claude 3.5 Sonnet, Claude 3 Haiku, GPT-4o, GPT-4o mini, and Llama 3.1^[9]. Among these models, Claude 3 Opus and Claude 3.5 Sonnet performed best in text-mode extraction, both achieving an accuracy of 81.14% and an F1 score of 84.70%. These studies suggest that larger proprietary models often perform better in complex reasoning, instruction-following, and zero- or few-shot extraction tasks, whereas open-weight models provide advantages in local deployment, customization, and reproducibility. Therefore, model selection should be determined by the specific application scenario, considering factors such as task complexity, data sensitivity, available computational resources, and the need for validation and long-term maintenance.

Reliance on proprietary API-based models also raises copyright, confidentiality, privacy, and data-governance concerns. Transmitting full-text articles, figures, unpublished manuscripts, or identifiable environmental-health data to third-party services may conflict with applicable licenses, institutional policies, or data-protection requirements^[67]. Locally deployed open-weight models can reduce external data transfer and improve version control, but they do not automatically resolve licensing, cybersecurity, maintenance, or performance limitations^[68]. Model selection should therefore consider accuracy, reproducibility, reuse rights, data sensitivity, provider terms, and deployment requirements.

Applicable boundaries

LLMs should currently be regarded as assistive tools rather than fully autonomous systems for environmental assessment. Given the complexity of environmental mechanisms, data heterogeneity, and regulatory requirements, mature end-to-end LLM workflows for complete environmental risk assessment are not yet available^[66]. The value of LLMs in environmental research is primarily reflected in improving data discovery efficiency, facilitating structured data generation, and enhancing the operability of data integration, but their application remains constrained by result traceability, domain knowledge constraints, absence of evaluation benchmarks, and regulatory acceptability.

Future application directions in the environmental domain

Future progress will depend on integrating LLMs more deeply with environmental domain knowledge and validation systems. LLMs in environmental research are better understood as collaborative tools rather than fully autonomous analytical agents. Kim et al. noted that high-quality databases, standardized formats, and domain knowledge constraints remain the foundational conditions for converting AI methods into usable results^[69]. Four directions warrant particular

attention. First, deep integration of retrieval-augmented generation with domain knowledge bases—combining LLMs with environmental ontologies, chemical databases, and standard terminology tables can improve term recognition accuracy and result consistency. The reliability of this approach will depend on the quality, coverage, and updating of the underlying knowledge resources. Second, because different LLMs may exhibit complementary strengths and failure modes, heterogeneous models can independently extract the same records, with disagreements examined through cross-critique, source verification, or human review. Cross-domain studies suggest that multi-LLM consensus can improve extraction accuracy and reduce hallucination, although stricter agreement may reduce data yield^[70,71]. Such agreement should therefore be treated as a validation signal rather than proof of correctness. Third, cross-scenario transfer should be evaluated beyond current domain-specific applications, such as water-conservancy risk decision support^[72]. This requires testing whether models developed for one pollutant, environmental medium, or endpoint retain their performance under different terminology and evidence structures. Fourth, task-specific benchmarks are needed for literature screening, relational knowledge mining, and quantitative data extraction^[19]. Shared datasets, common metrics, human-reference baselines, and task-specific error analyses would enable more reliable cross-study comparison. Overall, future applications should prioritize constrained, traceable, validated, and expert-guided workflows rather than fully automated knowledge generation.

Conclusion

This review systematically examined advances in LLM applications for literature screening, relational knowledge mining, and quantitative data extraction, starting from three core areas of environmental research: environmental exposure, toxicological effects, and pollution control. In literature screening, LLMs have demonstrated methodological feasibility in large-scale record pre-screening; however, given the potential bias introduced by false negatives, their role should remain limited to front-end pre-screening and priority ranking. In relational knowledge mining, mechanistic relationship extraction under the AOP framework has relatively mature implementation pathways, while process-oriented relationship mining for pollutant source–transport–fate and degradation transformation remains at an exploratory stage. In quantitative data extraction, domain-adapted tools have emerged across the exposure, toxicity, and degradation directions, but maintaining record-level completeness and cross-study comparability remain common bottlenecks.

The more feasible development pathway for LLMs in environmental research is not the pursuit of full automation, but rather the construction of a collaborative data processing framework jointly composed of retrieval augmentation, structured output, rule constraints, knowledge graphs, resource integration, and expert review. Within this framework, the role of LLMs should be defined as a supportive tool for improving data discovery and structured output efficiency, rather than an independent decision-making unit operating outside domain knowledge and quality control.

Author contributions

The authors confirm their contributions to the paper as follows: draft manuscript preparation: Yunlong Li; review and editing: Jing Guo and Wei Shi. All authors reviewed the results and approved the final version of the manuscript.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Funding

This study was supported by the National Major Program of China (42394153), the Basic Science Center Project of the National Natural Science Foundation of China (52388101), the National Natural Science Foundation of China (42477290, 22306093, 22376098, and 22276094), the National Key Research Program of China (2023YFC3709600 and 2024YFF1106703), the Jiangsu Province Key Research and Development Program (BE2023805), the Natural Science Foundation of Jiangsu Province (BK20230796), the Fundamental Research Funds for the Central Universities (2026300361), and the Jiangsu Provincial Young Scientific and Technological Talent Support Project.

Declarations

Competing interests

The authors declare that they have no conflict of interest.

Author details

¹State Key Laboratory of Water Pollution Control and Green Resource Recycling, School of the Environment, Nanjing University, Nanjing 210023, China; ²Jiangsu Province Ecology and Environment Protection Key Laboratory of Chemical Safety and Health Risk, Nanjing 210023, China

References

- [1] Landrigan PJ, Fuller R, Acosta NJR, Adeyi O, Arnold R, et al. 2018. The Lancet Commission on pollution and health. *Lancet* 391:462–512
- [2] Liu N, Jin X, Johnson AC, Zhou S, Liu Y, et al. 2025. Pharmaceutical and personal care products (PPCPs) in global surface waters: risk and drivers. *Environmental Science & Technology* 59:19146–19159
- [3] Wang Z, Walker GW, Muir DCG, Nagatani-Yoshida K. 2020. Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. *Environmental Science & Technology* 54:2575–2584
- [4] Keshava C, Davis JA, Stanek J, Thayer KA, Galizia A, et al. 2020. Application of systematic evidence mapping to assess the impact of new research when updating health reference values: a case example using acrolein. *Environment International* 143:105956
- [5] Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. *Environmental Health Perspectives* 122:711–718
- [6] Woodruff TJ, Sutton P. 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. *Environmental Health Perspectives* 122:1007–1014
- [7] Whaley P, Halsall C, Ågerstrand M, Aiassa E, Benford D, et al. 2016. Implementing systematic review techniques in chemical risk assessment: Challenges, opportunities and recommendations. *Environment International* 92-93:556–564
- [8] Pelch KE, Reade A, Kwiatkowski CF, Merced-Nieves FM, Cavalier H, et al. 2022. The PFAS-Tox Database: A systematic evidence map of health studies on 29 per- and polyfluoroalkyl substances. *Environment International* 167:107408
- [9] Moreira-Filho JT, Ranganath D, Tieghi RS, Patton R, Sutherland V, et al. 2025. Automating data extraction from scientific literature and general PDF files using large language models and KNIME: an application in toxicology. *WIREs Computational Molecular Science* 15:e70047

- [10] Olker JH, Elonen CM, Pilli A, Anderson A, Kinziger B, et al. 2022. The ECOTOXicology Knowledgebase: a curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. *Environmental Toxicology and Chemistry* 41:1520–1539
- [11] Wolffe TAM, Whaley P, Halsall C, Rooney AA, Walker VR. 2019. Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. *Environment International* 130:104871
- [12] Schmidt L, Finnerty Mutlu AN, Elmore R, Olorisade BK, Thomas J, et al. 2021. Data extraction methods for systematic review (semi) automation: Update of a living systematic review. *F1000Research* 10:401
- [13] Haddaway NR, Macura B, Whaley P, Pullin AS. 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
- [14] Borah R, Brown AW, Capers PL, Kaiser KA. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 7:e012545
- [15] Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, et al. 2021. Resource use during systematic review production varies widely: a scoping review. *Journal of Clinical Epidemiology* 139:287–296
- [16] Walker VR, Schmitt CP, Wolfe MS, Nowak AJ, Kulesza K, et al. 2022. Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr. *Environment International* 159:107025
- [17] O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 4(1):5
- [18] Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, et al. 2024. Structured information extraction from scientific text with large language models. *Nature Communications* 15:1418
- [19] Raeissi MM, Knapen R. 2025. Applications of generative large language models in environmental science: a systematic review. *Advances in Environmental and Engineering Research* 6(3):028
- [20] Delgado-Chaves FM, Jennings MJ, Atalaia A, Wolff J, Horvath R, et al. 2025. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences* 122:e2411962122
- [21] Lieberum JL, Toews M, Metzendorf MI, Heilmeyer F, Siemens W, et al. 2025. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *Journal of Clinical Epidemiology* 181:111746
- [22] Galli C, Gavrilova AV, Calciolari E. 2025. Large language models in systematic review screening: opportunities, challenges, and methodological considerations. *Information* 16:378
- [23] Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, et al. 2024. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods* 15:576–589
- [24] Schmidt L, Campbell F, Craig D, Walker VR, Rooney AA, et al. 2025. Large language models for data extraction in toxicology: implications and lessons learned from the clinical evidence domain. *Evidence-Based Toxicology* 3:2546791
- [25] Zuo C, Yang X, Erickson J, Li J, Hong Y, et al. 2025. AI-assisted evidence screening method for systematic reviews in environmental research: integrating ChatGPT with domain knowledge. *Environmental Evidence* 14(1):5
- [26] Nykvist B, Macura B, Xylia M, Olsson E. 2025. Testing the utility of GPT for title and abstract screening in environmental systematic evidence synthesis. *Environmental Evidence* 14:7
- [27] Nicholson Thomas I, Roche P, Grêt-Regamey A. 2024. Harnessing artificial intelligence for efficient systematic reviews: A case study in ecosystem condition indicators. *Ecological Informatics* 83:102819
- [28] Marshall IJ, Kuiper J, Banner E, Wallace BC. 2017. Automating biomedical evidence synthesis: RobotReviewer. Proc Conf Assoc Comput Linguist Meet 2017. Vancouver, Canada: Association for Computational Linguistics. pp. 7–12 <https://doi.org/10.18653/v1/P17-4002>
- [29] Baudiffier D, Audouze K, Armant O, Frelon S, Charles S, et al. 2024. Editorial trend: adverse outcome pathway (AOP) and computational strategy - towards new perspectives in ecotoxicology. *Environmental Science and Pollution Research* 31:6587–6596
- [30] Corradi MPF, de Haan AM, Staumont B, Piersma AH, Geris L, et al. 2022. Natural language processing in toxicology: Delineating adverse outcome pathways and guiding the application of new approach methodologies. *Biomaterials and Biosystems* 7:100061
- [31] Corradi M, Luechtefeld T, de Haan AM, Pieters R, Freedman JH, et al. 2024. The application of natural language processing for the extraction of mechanistic information in toxicology. *Frontiers in Toxicology* 6:1393662
- [32] Chen Z, Zhong R, Long W, Tanga H, Wang A, et al. 2025. Advancing oil and gas emissions assessment through large language model data extraction. *Energy and AI* 20:100481
- [33] Xu B, Wu G, Li Z, Xu G, Zeng H, et al. 2025. Towards domain-adapted large language models for water and wastewater management: methods, datasets and benchmarking. *npj Clean Water* 8:82
- [34] Li Q, Cheng F, You J. 2026. Large language models in aquatic risk assessment: research status and future perspectives. *Environmental and Biogeochemical Processes* 2:e007
- [35] Liang W, Su W, Zhong L, Yang Z, Li T, et al. 2024. Comprehensive characterization of oxidative stress-modulating chemicals using GPT-based text mining. *Environmental Science & Technology* 58:20540–20552
- [36] Shi H, Zhao Y. 2024. Integration of advanced large language models into the construction of adverse outcome pathways: opportunities and challenges. *Environmental Science & Technology* 58:15355–15358
- [37] Jaylet T, Jornod F, Capdet Q, Armant O, Audouze K. 2025. AOP-helpFinder 3.0: from text mining to network visualization of key event relationships, and knowledge integration from multiple sources. *Bioinformatics* 41:btaf381
- [38] Carvillio JC, Barouki R, Coumoul X, Audouze K. 2019. Linking bisphenol S to adverse outcome pathways using a combined text mining and systems biology approach. *Environmental Health Perspectives* 127:47005
- [39] Kaiser AM, Zare Jeddi M, Uhl M, Jornod F, Fernandez MF, et al. 2022. Characterization of potential adverse outcome pathways related to metabolic outcomes and exposure to per- and polyfluoroalkyl substances using artificial intelligence. *Toxics* 10:449
- [40] James F, Pratt D, Churas C, Luna A. 2026. textToKnowledgeGraph: generation of molecular interaction knowledge graphs using large language models for exploration in Cytoscape. *Bioinformatics* 42:btag031
- [41] El Khettari O, Nishida N, Liu S, Munne RF, Yamagata Y, et al. 2024. Mention-agnostic information extraction for ontological annotation of biomedical articles. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. Bangkok, Thailand: Association for Computational Linguistics. pp. 457–473 <https://doi.org/10.18653/v1/2024.bionlp-1.37>
- [42] Sepehri S, Heymans A, Win D, Maushagen J, Sanctorum A, et al. 2025. The TOXIN knowledge graph: supporting animal-free risk assessment of cosmetics. *Database (Oxford)* 2025:baae121
- [43] Luechtefeld T, Hartung T. 2025. Navigating the AI frontier in toxicology: trends, trust, and transformation. *Current Environmental Health Reports* 12:51
- [44] Zhu Z, Hu G, Jia G. 2026. Artificial intelligence in adverse outcome pathways: a review of strategies for automated information extraction, quantitative analysis, and iterative optimization. *Occupational Health* 1:9
- [45] Lai PT, Wei CH, Tian S, Leaman R, Lu Z. 2025. Enhancing biomedical relation extraction with directionality. *Bioinformatics* 41:i68–i76
- [46] Wall JT, Sayre RR, Smith D, Winter S, Groover M, et al. 2025. Development of the toxicity values database, ToxValDB: A curated resource for experimental and derived human health-relevant toxicity data. *Computational Toxicology* 35:100365
- [47] Schilling-Wilhelmi M, Rios-Garcia M, Shabih S, Gil MV, Miret S, et al. 2025. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews* 54:1125–1150
- [48] Cheng Y, Zhang H. 2025. L2D: a versatile literature-to-data set pipeline for complex, user-specific data extraction in environmental research. *ACS ES&T Water* 5:4897–4907

- [49] Shah-Mohammadi F, Im S, Facelli JC, Cummins MR, Gouripeddi R. 2026. Scaling sensor metadata extraction for exposure health using LLMs. *Exposome* 6:osag008
- [50] Sonnenburg A, van der Lugt B, Rehn J, Wittkowski P, Bech K, et al. 2024. Artificial intelligence-based data extraction for next generation risk assessment: Is fine-tuning of a large language model worth the effort? *Toxicology* 508:153933
- [51] Paula AJ, Petry R, Almeida JM, Caetano AA, Sales J, et al. 2026. Large-scale meta-analysis of nanomaterials toxicity based on natural language processing of scientific articles. *ACS Applied Nano Materials* 9:28–44
- [52] Ha E, Ha SM, Gerelkhuu Z, Kim H-Y, Yoon TH. 2025. AI-based nanotoxicity data extraction and prediction of nanotoxicity. *Computational and Structural Biotechnology Journal* 29:138–148
- [53] Wu J, Shi R, Zhou X, Zhang L, Chen K, et al. 2026. Empowering chemistry experts with large language models for literature interpretation in single-atom catalysis toward advanced oxidation. *Angewandte Chemie-International Edition* 65:e202520525
- [54] Park S-G, Ha SM, Gerelkhuu Z, Yoon TH. 2026. Automated large-scale extraction of nanotoxicity data from the literature using multimodal large language models. *Computational and Structural Biotechnology Journal* 35:0161
- [55] Wang J, Ye Q, Liu L, Guo NL, Hu G. 2024. Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception. *npj Precision Oncology* 8:84
- [56] Wang Z, Xia M, He L, Chen H, Liu Y, et al. 2024. CharXiv: charting gaps in realistic chart understanding in multimodal LLMs. *Advances in Neural Information Processing Systems* 37. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS). pp. 113569–113697 <https://doi.org/10.52202/079017-3609>
- [57] Zhang W, Wang Q, Kong X, Xiong J, Ni S, et al. 2024. Fine-tuning large language models for chemical text mining. *Chemical Science* 15:10600–10611
- [58] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* 33:9459–9474
- [59] Zou Y, Shi M, Chen Z, Deng Z, Lei Z, et al. 2025. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *Journal of Cleaner Production* 489:144572
- [60] Polak MP, Morgan D. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications* 15:1569
- [61] Hodgson RY, Robinson SA, Boutin AC, Chan FK, Bennett JR, et al. 2026. Assessing the effectiveness of ontology-grounded AI term extraction using OntoGPT for environmental evidence synthesis. *Environmental Evidence* 15:1
- [62] Bauer C, Duc Dang LT, van den Beucken T, Schuchhardt J, Herwig R. 2025. Systematic analysis of hepatotoxicity: combining literature mining and AI language models. *Frontiers in Artificial Intelligence* 8:1561292
- [63] Gougherty AV, Clipp HL. 2024. Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *npj Biodiversity* 3:13
- [64] Kleinstreuer N, Hartung T. 2024. Artificial intelligence (AI)-it's the end of the tox as we know it (and I feel fine). *Archives of Toxicology* 98:735–754
- [65] Hartung T, Kleinstreuer N. 2025. Challenges and opportunities for validation of AI-based new approach methods. *Altex* 42:3–21
- [66] Gant TW, Boxall A, Burgwinkel D, Zare Jeddi M, Djidrovski I, et al. 2026. Building trust in the integration of artificial intelligence into chemical risk assessment: findings from the 2024 ECETOC workshop. *Archives of Toxicology* 100:2149–2167
- [67] Bjelobaba S, Waddington L, Perkins M, Foltýnek T, Bhattacharyya S, et al. 2025. Maintaining research integrity in the age of GenAI: an analysis of ethical challenges and recommendations to researchers. *International Journal for Educational Integrity* 21:18
- [68] Wiest I, Ferber D, Zhu J, van Treeck M, Meyer S, et al. 2024. Privacy-preserving large language models for structured medical information retrieval. *npj Digital Medicine* 7:257
- [69] Kim D, Choi J. 2025. Big data and AI: Potential and challenges for digital transformation in toxicology. *Environmental Analysis Health and Toxicology* 40:e2025s2007
- [70] Khan MA, Ayub U, Naqvi SAA, Khakwani KZR, Sipra ZbR, et al. 2025. Collaborative large language models for automated data extraction in living systematic reviews. *Journal of the American Medical Informatics Association* 32:638–647
- [71] MacKay EJ, Goldfinger S, Chan TJ, Grasfield RH, Eswar VJ, et al. 2025. Automated structured data extraction from intraoperative echocardiography reports using large language models. *British Journal of Anaesthesia* 134:1308–1317
- [72] Yang L, Li Y, Tan J, Mao L. 2025. Research on risk decision-making generation method for water conservancy project based on multimodal knowledge graph and large language model. *PLoS One* 20:e0330258



Copyright: © 2026 by the author(s). Published by NEW HORIZON PRESS LIMITED, TSIM SHA TSUI, HK. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.