Regulatory Toxicology and Pharmacology 171 (2026) 106182

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

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

Automated extraction of toxicological mechanism of action from PubMed literature using Large Language Models

Arnaud Molle [a,*], Nelly D. Saenen [b], Karen Smeets [b], Karolien Bijnens [b], Marc Aerts [c], Cécile Kremer [c], Efisio Solazzo [a], José Cortiñas Abrahantes [a]

[a] European Food Safety Authority, Via Carlo Magno 1A, Parma, 43126, Italy
[b] Centre for Environmental Sciences, Universiteit Hasselt, Campus Diepenbeek, Agoralaan Gebouw D, Diepenbeek, B-3590, Belgium
[c] Data Science Institute, I-BioStat, Universiteit Hasselt, Campus Diepenbeek, Agoralaan Gebouw D, Diepenbeek, B-3590, Belgium

ARTICLE INFO

Handling Editor: Dr. Daniele Wikoff

ABSTRACT

Risk assessment activities at the European Food Safety Authority face mounting challenges from an increasing volume of compounds requiring evaluation and exponential growth in scientific literature. To address these challenges, toxicologists started grouping compounds based on their mechanism of action in the body, allowing study comparisons similar to current risk assessment strategies. The LLMs-rev pipeline was created to retrieve potentially relevant papers from PubMed, assess their relevance, and extract the mechanism of action (MoA) information from accessible publications using large language models. Applied to 121 compounds, the system processed over 400,000 papers, identifying 30,250 as containing relevant information and extracting specific MoA quotations from 4500 open-access publications. The automated approach demonstrated processing speeds exceeding 8000 papers per hour, dramatically outpacing conventional manual screening methods that typically assess 100-120 papers per hour. While the system proved particularly effective for compounds for which abundant literature was available, human expertise remained essential for interpreting complex MoAs that required contextual data analysis. The optimized prompts minimized model hallucinations by restricting outputs to direct quotations from verified sources. The methodology demonstrates considerable potential for accelerating risk assessment workflows while maintaining scientific rigor through the complementary use of human oversight.

1. Introduction

In recent years, the Risk Assessment (RA) activities of the European Food Safety Authority (EFSA) have faced two major challenges: i) high number of compounds on which risk assessment has been performed, and ii) high volume of scientific literature relative to those compounds and consequently requiring scrutiny. Both aspects negatively impact the speed of RA activities, which remains a top priority for EFSA.

The quantity of compounds requiring assessment by EFSA has been consistently high between 2018 and 2023, ranging from 586 to 741 questions closed per year (see EFSA Consolidated annual activity report, 2023). At the same pace, the number of publications in the EFSA Journal between 2021 and 2024 ranged between 589 and 625 per year. Similarly, the calls for data emitted by EFSA have steadily grown, reaching an average of 12 calls per year over the 2021-2024 period. This workload is accompanied by enhanced quality requirements in assessments, which nowadays incorporate additional factors such as environmental sustainability and animal welfare (European Food Safety Authority, 2024a; European Food Safety Authority, 2024b). Moreover, EFSA aims to report open, transparent, and reproducible science (European Parliament and Council of the European Union, 2019), which overall involves a higher workload to pursue the RA activities.

The second challenge lies in the volume of scientific literature related to the studied compounds, driven by the democratization of scientific publishing and the expansion of research fields. The scientific literature is continuously increasing with growth patterns suggesting an exponential increase (Bornmann et al., 2021).

To ensure risk assessments incorporate the most relevant and current scientific knowledge, risk assessors must strive for comprehensive literature coverage, which is becoming increasingly challenging, error-

* Corresponding author.
E-mail addresses: Arnaud.molle1@gmail.com (A. Molle), Nelly.Saenen@uhasselt.be (N.D. Saenen), Karen.Smeets@uhasselt.be (K. Smeets), Karolien.Bijnens@uhasselt.be (K. Bijnens), Marc.Aerts@uhasselt.be (M. Aerts), Cecile.Kremer@uhasselt.be (C. Kremer), Efisio.Solazzo@efsa.europa.eu (E. Solazzo), Jose.CortinasAbrahantes@efsa.europa.eu (J.C. Abrahantes).

https://doi.org/10.1016/j.yrtph.2026.106182
Received 14 February 2026; Received in revised form 22 June 2026; Accepted 19 July 2026
Available online 24 July 2026

prone, and lengthy when performed through manual screening. While various online tools have emerged to assist risk assessors in retrieving and connecting scientific papers, questions remain about their comprehensiveness and potential biases, particularly given the increasing pressure to maintain efficient assessment timelines (Landhuis, 2016). To address these challenges, emerging technologies in artificial intelligence present promising opportunities for automation and scalability.

Artificial Intelligence (AI), specifically Large Language Models (LLMs), offers promising solutions to these challenges by accelerating manual processes and efficiently extracting information from large text corpora. The application of LLMs is being extensively studied across various domains, with rapid improvements in their capacity to handle large-scale data processing (Gupta et al., 2024; Mishra et al., 2024; Delgado-Chaves et al., 2025). The integration of AI tools in toxicological assessments reflects broader trends in the field. Blümmel et al. (2023) and Blümmel et al. (2024) reviewed state-of-the-art AI tools for screening and evaluating literature in chemical risk assessment, emphasizing the potential of AI to handle large datasets and streamline literature reviews. LLMs are also used for different applications in toxicology such as the construction of adverse outcome pathways (Shi and Zhao, 2024).

Beyond conventional conversational agents (i.e. ChatGPT from OpenAI, Claude from Anthropic, or Gemini from Google), AI model providers typically offer Application Programming Interfaces (APIs) that enable programmatic model access. This capability enables the development of complex, reproducible procedures and the repeated application of queries across multiple data sources. By utilizing LLMs independently of their conversational interface, users can perform complex operations that would require significantly more time if conducted manually, and that would also be more prone to human error. For instance, LLMs could assist in identifying information about toxicological mechanisms of action (MoA) of compounds, supporting their risk assessment.

However, the use of LLMs in scientific contexts is not without limitations. A central concern is reliability: LLMs can generate plausible but factually incorrect outputs through a phenomenon known as "hallucination" (Huang et al., 2025). Beyond factual errors, LLMs may also distort scientific meaning by overgeneralizing research conclusions beyond what the original studies support, a bias particularly relevant when summarizing toxicological evidence (Peters and Chin-Yee, 2025). These reliability concerns are compounded by biases inherited from training data: LLMs are known to systematically overrepresent certain languages, viewpoints, and publication traditions, potentially skewing literature-based assessments (Gallegos et al., 2024). Transparency also remains a challenge, as most LLMs operate as black boxes with opaque internal reasoning processes, raising significant concerns regarding auditability and accountability in high-stakes domains (Chustecki, 2024). In the context of regulatory risk assessment, where conclusions carry direct public health implications, these are not merely theoretical concerns: the degree of error that is acceptable from an automated pipeline must be explicitly defined, validated against expert annotation, and distinguished from efficiency gains: speed and accuracy being distinct and independently measurable properties. Furthermore, the deployment of large-scale LLMs raises ecological concerns, as training and running such models entails substantial energy consumption and associated carbon emissions (Ren et al., 2024). Ethical considerations also arise regarding intellectual property, data provenance, and the appropriate role of automated tools in regulatory decisions (Cheong, 2024).

The study presented here leverages the capabilities of LLMs to address the aforementioned challenges in scientific literature review by developing a fast and efficient pipeline for the extraction of relevant MoA information across multiple compounds. A set of LLMs-assisted algorithms (hereafter referred to as LLMs-rev) have been developed to this aim. The case study that is presented has been used to assist a scientific report (Kremer et al., 2026) by classifying and grouping compounds by MoA on specific endpoints. Specifically, LLMs-rev has been used to retrieve potentially relevant papers from PubMed, evaluate their relevance, and extract MoA information from open-access publications. For non-open-access papers, users with appropriate credentials and under specific copyright license terms could upload PDF documents to be inspected by the system for MoA extraction. The automated approach devised through LLMs-rev aims to accelerate the RA process by providing systematic, large-scale literature screening support. The tool is designed to operate under mandatory human oversight, and its outputs are intended to assist (rather than replace) expert toxicological judgment. The scope and limitations of the approach are discussed in light of the concerns raised above.

2. Materials and methods

2.1. Mechanism of action

The MoA of a compound in this study refers to the specific biochemical, cellular, or molecular interactions and specific pathway through which a toxicant produces an adverse effect. Because of the variety of the MoA and the way it is usually described in literature, simple retrieval through keywords is not sufficient and far from exhaustive. LLM models can be a powerful tool to tackle this, since the extraction of the MoA from relevant papers would need interpretation of the context, using semantic relationships and pattern recognition, as well as a deeper understanding of the natural language used in the retrieved literature.

The present case study focuses on the extraction of MoAs for compounds identified by their Chemical Abstracts Service (CAS) number and present in the OpenFoodTOX (https://www.efsa.europa.eu/en/data-report/chemical-hazards-database-openfoodtox) and NTP (https://ntp.niehs.nih.gov/data/download) databases, as well as EPA's IRIS database (https://www.epa.gov/iris/). A broad, general term was used to account for differences in terminology and study approaches, thereby allowing inclusion of a wider range of papers. Examples of subcellular events and effects used for the determination of MoAs are oxidative stress, membrane damage or mitochondrial dysfunction.

2.2. Technical aspects

The script underlying LLMs-rev has been developed in R (R Core Team, 2023) and deployed on the Azure Databricks platform. It interfaces with the OpenAI API to utilize the GPT-4o model (128k tokens context window, OpenAI, 2024) through Python libraries, accessed in R via the 'reticulate' package (Ushey et al., 2023). This approach was adopted to circumvent stability issues encountered with the native OpenAI R package. The entire process of searching, scrutiny and extraction was carried out on a dedicated cluster and took approximately 60 h. The cluster consisted of a 4-core processor with 16 GB of active memory, using the Azure D4ds_v5 parameters. The runtime version of Databricks was 15.4 LTS (Scala 2.12, Spark 3.5.0). The full script is available at [https://doi.org/10.5281/zenodo.17546348]. The performance of the script will depend on the resources of the infrastructure hosting the LLM-rev processes.

The R environment was set up using the CRAN repository URL dated from 2024 to 09-15. This ensured that all R packages were installed from a consistent and reliable source, which is the operating system used in Azure Databricks. The OpenAI API version "2024-02-15-preview" was utilized for relevance assessment and content analysis. The LLMs-rev pipeline consists of two main phases: (1) retrieval and relevance assessment of papers from PubMed, and (2) extraction of MoA-specific quotations from relevant publications.

2.3. Part I: Retrieval and assessment of the relevance of PubMed papers

The initial phase of the workflow involves retrieving scientific papers from the PubMed database. To overcome the PubMed API's fundamental limitation of retrieving up to 10,000 records per query, the search was systematically subdivided by publication year. Papers' metadata, including titles, authors, DOIs, and abstracts, were then extracted using a combination of the easyPubMed (Fantini, 2019) and rentrez (Winter, 2021) R packages and structured into XML format with the xml2 package (Wickham et al., 2023). The query consisted in retrieving all papers that contain the studied molecule name or CAS number.

The second part of the workflow focuses on assessing the relevance of these papers. We used GPT-4o with carefully optimized prompts (Table 2) to analyze each paper's title and abstract. These prompts were designed to identify specific information related to the MoA target organs, and physiological processes. A paper was flagged as relevant if the abstract or title contained evidence of direct interaction between the compound and biological systems at the cellular, molecular, or organ level, as specified in the prompt instructions (Table 2). Papers citing the compound solely in an industrial, agricultural, or manufacturing context were excluded. Papers flagged as potentially relevant were listed in a standardized Excel file for subsequent analysis.

The entire process was automated to handle multiple target compounds, identified by their CAS numbers. Tidyverse (Wickham et al., 2019) packages were used for data manipulation and analysis, while writexl handled the final export of results to Excel, ensuring a consistent and streamlined output.

2.4. Part II: Extraction of MoA quotations from retrieved scientific papers

Once a list of relevant papers has been established, the second phase involves extracting specific quotations related to the toxicological events and effects that could be used to determine the MoA. A dedicated GPT-4o prompt (Table 2) analyses the full text content of each paper for this purpose. A key aspect of this step is implementing robust error-handling mechanisms to address common issues such as token length limitations, API communication failures, and content policy restrictions.

The extraction process follows a dual-path approach:

1. **Open-Access Papers**: For papers publicly available on PubMed, the script directly processes the full-text content to extract direct MoA-related quotations directly from the results of the easypubmed query in XML format, not saving any version of the paper locally.
2. **Non-Open-Access Papers**: For papers that are not publicly available, a separate workflow was developed. Users with appropriate access rights and copyright license can download these papers as PDFs and organize them into compound-specific directories. The script then systematically converts these local PDF documents into XML files and processes them.

To maintain scientific integrity, the process is strictly limited to extracting in text and intact quotations identified by the LLM as containing MoA related information, thereby avoiding any model interpretation, extrapolation, or potential hallucinations. Such hallucinations are described as plausible yet non factual content by Huang et al. (2025). The output from both open-access and local PDF analyses is a standardized Excel file containing paper metadata and the extracted quotations. This ensures a consistent data format for all downstream analyses, regardless of the source. The overall workflow is illustrated in Fig. 1. In this study, only open access papers have been considered.

3. Results

3.1. Query

The developed script successfully extracted MoA information from a list of 121 compounds that were needed for the case study, processing over 424,250 papers, of which 73,175 were fully available online (2A and Fig. 2:B). A manual search was also performed on a subset of targeted compounds for consistency and quality checks purposes. It provided the same results as the automated query created via the EasyPubmed packages (example in Table 1). Another check consisted of using the LLMs-rev with the Easy PubMed package and assessing if its output is equivalent to querying "compound name OR CAS_number" in the PubMed online browser. This gives confidence in the fact that the automatic querying is at least as good as a manual search, while being much faster.

3.2. Assessment of relevance

As shown in Fig. 2A, the pipeline retrieved over 400,000 papers, of which approximately 73,000 were available as full-text documents. From these, 30,250 papers were classified as containing relevant

Table 1
Statistics of the PubMed extraction, relevance assessment, and MoA extraction for Sodium dichromate dihydrate.

Metric of interest	Number
Papers retrieved	118
Available full-texts	9
Relevant papers	83
Relevant full text papers	7
Number of quotations	122

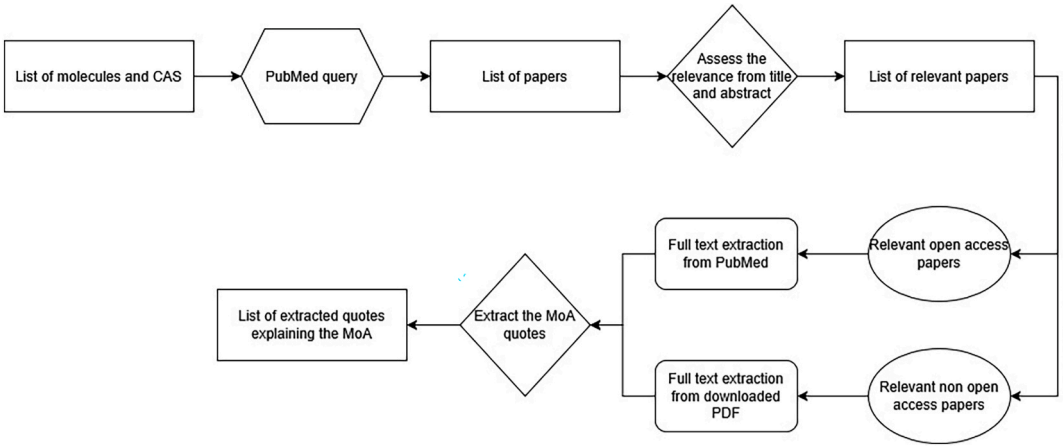

Fig. 1. Overall workflow for papers accessible via PubMed.

Table 2
Step by step prompts given to GPT-4o through the OpenAI API.

Use	Prompt
Relevance Assessment	paste("Analyze the title and abstract of this scientific paper to determine if the full paper might contain information about the mechanism of action (MOA) or mode of action or adverse outcomes for the toxicological effects of ",molecule," on humans or test subjects (mammals/birds). Based on these criteria, provide a clear YES or NO response only. Input text:", fulltext)
MoA Extraction	sprintf("Analyze the full text of this scientific paper and extract direct quotes that specifically describe the mechanism of action (MOA) for the toxicological effects of %s on humans, birds, or mammals used in testing. Focus exclusively on sentences or paragraphs that explain: Molecular pathway descriptions of toxicity Physiological toxic effects and their progression Details of toxic biological interactions Target organs or systems affected by the toxicity Cellular or organ-specific responses to toxicity Biochemical processes involved in toxicity Return only the relevant quotes that directly describe these mechanisms, maintaining the exact wording from %s The output should not include anything else than quotes", molecule, fulltext_combined)

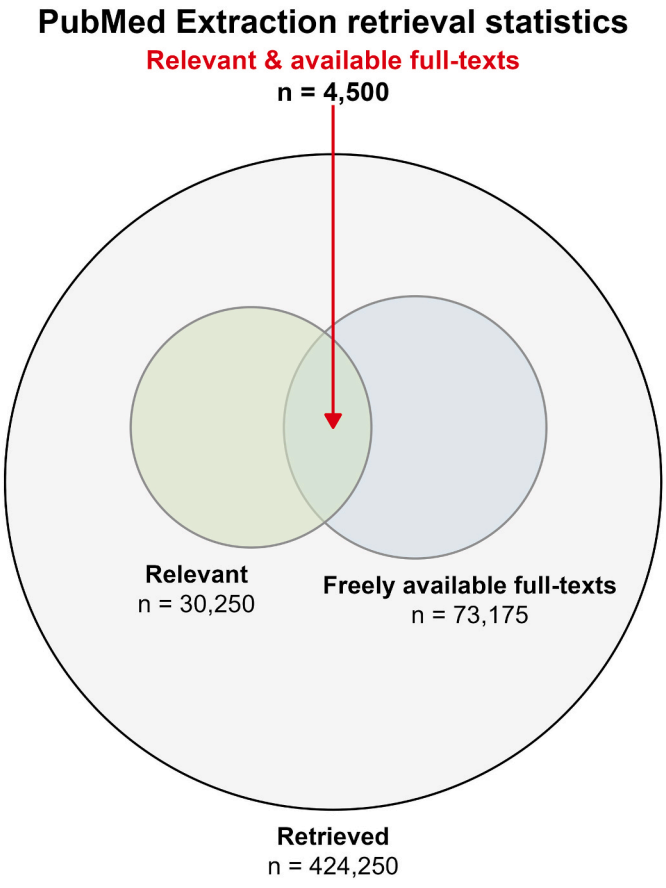

Fig. 2A. Retrieval statistics.

information for MoA classification based on title and abstract screening, and 4500 open-access papers underwent full MoA extraction, yielding specific toxicological mechanism quotations. The pipeline extracted a total of 48475 quotations for 73 molecules, ranging from 12888 (from 1209 papers) for carbon tetrachloride to <10 for the 10 least represented molecules. These numbers reflect the scale and efficiency of the pipeline as well as the broad range of literature corpus available to this day. Distribution of quotations is shown in Fig. 2B.

The prompts used for relevance assessment were systematically optimized through a rigorous, iterative human-controlled process to minimize hallucinations and extrapolations from GPT-4o, while preventing the selection of misleading information. These limitations were previously acknowledged by Sonnenburg et al. (2024) and Shaib et al. (2023). Through iterative trial, evaluation and refinement, the prompts were designed to strictly align the research objective with the compound's Mechanism of Action (MoA). Specifically, human oversight was maintained to ensure the exclusion of literature citing a molecule solely for industrial or manufacturing contexts, isolating only its toxicological effects. This calibration was executed on a limited subset of compounds; after each iteration, the generated output was manually audited, and the prompt was progressively adjusted until the resulting bibliography was entirely free of misleading citations. This optimization required less than 5 iterations. The primary objective of this human-led optimization was not to achieve a flawless definitive assessment (as full-text analysis remains mandatory) but rather to establish a robust filtering mechanism against irrelevant literature. The calibration needs accurate statements, as vague prompts could inadvertently include irrelevant information (e.g., a compound's mechanism as a pesticide rather than its toxicological effects of interest), while overly restrictive prompts risked excluding valuable insights. A balanced approach was therefore adopted to ensure accurate identification of relevant information without omitting critical details. The finalized prompts are presented in Table 2.

3.3. Quote extraction

Special attention was required for processing the output format from GPT-4o, as these responses served as the foundation for subsequent paper selection and prompt generation. While a similar methodology is expected to be compatible with next-generation LLM architectures, this step was subjected to rigorous expert supervision to ensure data integrity. Specifically, the prompt was iteratively optimized by benchmarking the AI-extracted quotations against a manual, gold-standard study previously conducted by the toxicologists on a limited list of compounds. This comparison revealed that if the prompt did not explicitly mandate the retrieval of strictly intact quotations from the text, GPT-4o tended to fabricate, extrapolate, or summarize inferred information rather than returning an empty or negative response. Such behavior introduces substantial noise into the dataset and risks biasing the selection of the correct Mechanism of Action (MoA). Consequently, through continuous human verification, the prompt was refined until the extracted quotations perfectly matched the precision of the manual toxicological study. The resulting structure required careful handling to populate the finalized extraction table, as exemplified by the data for sodium dichromate dihydrate in Table 3.

3.4. Validation and utility assessment

3.4.1. Assessment of GPT-4o-assisted literature screening for MoA classification

The extracted quotations of the GPT-4o-assisted workflow have been used to support toxicologists in classifying compounds by MoA by extracting relevant evidence from the scientific literature. To assess the reliability of the approach, the workflow was initially utilized for a limited subset of compounds. The extracted quotations and associated literature sources were qualitatively reviewed by toxicologists and considered alongside information obtained through conventional manual literature screening.

Special attention has been put to avoid model hallucination or extrapolation from the text. The optimized prompt, therefore, includes a specific mention to extract in-text quotations only. This verification process proved particularly valuable for compounds with limited available scientific literature or for compounds where different

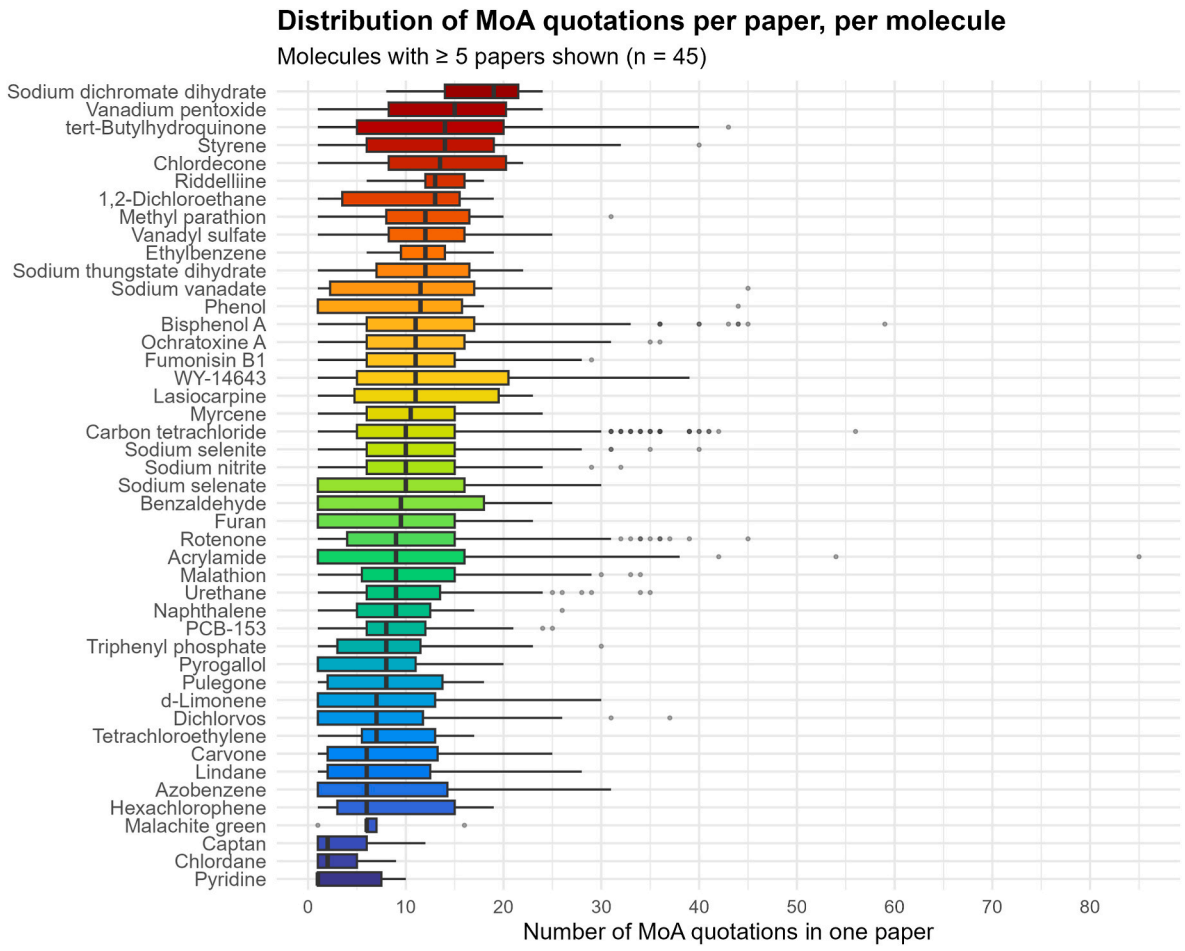

Fig. 2:B. Retrieval statistics per molecules.

parameters of an underlying mechanism were studied. By consolidating relevant evidence from diverse sources, the approach provided a comprehensive overview of the available literature and supported expert-driven MoA classification.

For selected compounds, the extracted quotations provided literature evidence consistent with and supporting the MoA classifications derived through manual screening. However, for compounds with complex MoAs requiring interpretation of figures and contextual data analysis, insufficient information was provided by the extracted quotations alone to accurately interpret the MoA. There is no convention for the wording of the description or the MoA, which makes its identification more complicated. This finding aligns with previous studies about biomedical literature (Toth et al., 2024). In these instances, targeted investigation of appropriate sources was facilitated by the comprehensive list of relevant literature, both open and closed-access. The extraction step operates on the full text of a paper once it has been flagged as relevant for the target compound, and pulls all MoA-related quotations regardless of which specific compound they describe. For papers studying multiple metals or compounds simultaneously (as in the zebrafish multi-metal exposure study cited in Table 3), this can yield quotations describing co-occurring compounds (e.g., nickel) alongside those describing the target compound. This is currently resolved through mandatory human review of all extracted quotations before MoA conclusions are drawn, rather than automated compound-specific filtering.

3.4.2. Performance with limited literature

In the present study, the query to the PubMed database retrieved fewer than 30 full-text papers for 40 of the analyzed compounds, which represented a third of the total number of compounds considered. This led to the pipeline not retrieving any quotations for 48 of the 121 analyzed molecules and less than 20 quotations for 19 of the 73 molecules for which quotes have been retrieved. This limitation tends to justify the use of the LLMs-rev mostly for compounds with abundant literature, where all potential MoAs will be thoroughly represented.

3.4.3. Limitations and expert oversight requirements

This developmental methodology was employed to validate manual searches, providing additional confidence in the findings. As found in Blümmel et al. (2023), the AI extraction tool shows its best potential as an assistant to experts, rather than a replacement. The main disadvantage of this automated data extraction from scientific literature arises from its current limitations of LLMs to access information stored in images, plots, and charts. Even though the GPT-4o model would be able to use images as input, no specific instructions have been given in the prompt to ensure it is done, as the retrieval of full text papers from easypubmed only convert texts. For toxicological studies in which the MoA is represented visually, current LLMs that rely solely on text may overlook it. This limitation is likely to be addressed as LLMs evolve towards incorporating multimodal information within a reasonable use of tokens. As LLMs advance, the range of applications and input file flexibility will broaden, and it is expected that applying the current methodology with small, relevant fixes will only increase output quality.

4. Conclusion

The LLMs-rev automated extraction system has demonstrated considerable potential for accelerating toxicological assessment workflows. Applied to 121 compounds, the system assessed over 400,000

Table 3
Sample of extracted results for Sodium dichromate dihydrate.

Paper title	Authors and DOI	Extracted quotations
Synchrotron-based imaging of chromium and γ-H2AX immunostaining in the duodenum following repeated exposure to Cr(VI) in drinking water.	Chad M Thompson; Jennifer Seiter; Mark A Chappell; Ryan V Tappero; Deborah M Proctor; Mina Suh; Jeffrey C Wolf; Laurie C Haws; Rock Vitale; Liz Mittal; Christopher R Kirman; Sean M Hays; Mark A Harris 10.1093/toxsci/kfu206	1. "Exposure to Cr(VI) induced villus blunting and crypt hyperplasia in the duodenum–the latter evidenced by lengthening of the crypt compartment by ~2-fold with a concomitant 1.5-fold increase in the number of crypt enterocytes." 2. "Cr(VI) exposure to B6C3F1 mice causes villous cytotoxicity, followed by crypt proliferation without evidence of cytogenetic damage or increases in kras mutation frequency." 3. "μ-XRF maps revealed mean Cr levels >30 times higher in duodenal villi than crypt regions; mean Cr levels in crypt regions were only slightly above background signal." 4. "These findings do not support a MOA for intestinal carcinogenesis involving direct Cr-DNA interaction in intestinal stem cells, but rather support a non-mutagenic MOA involving chronic wounding of intestinal villi and crypt cell hyperplasia." 5. "The pattern of Cr(VI)-induced lesions in mice is consistent with the fact that most dietary absorption occurs via villus enterocytes of the duodenum and proximal jejunum." 6. "Villus enterocytes are differentiated, short-lived, and destined to slough into the intestinal lumen, and therefore DNA damage in villus enterocytes poses little, if any, carcinogenic risk." 7. "At low magnification, γ-H2AX immunostaining in the crypt region appears darker in the treated mouse specimens; however, this appearance is due to the increased cellular density caused by crypt epithelial proliferation." 8. "The absence of increased γ-H2AX staining in the proliferative crypt compartment of mice exposed to Cr(VI) suggests that Cr(VI) does not directly reach crypt enterocytes in vivo." 9. "The μ-XRF maps exhibited strong calcium (Ca) and sulfur (S) signals in the crypts and villi (effectively outlining the shape of the small intestine), whereas Cr signal (determined to be Cr(III)) was limited to the intestinal villi." 10. "Chromium treatment did not appear to dramatically alter the μ-XRF maps for Ca and S; however, these maps clearly show the villus blunting in the duodena of treated animals." 11. "Quantification of the μ-XRF map indicated that the mean and median concentrations of Cr in the villi were 14.0 and 11.9 μg/g, respectively, whereas the mean and median concentration of Cr in the crypt region was 0.4 and 0.4 μg/g" 12. "F344 rats did not exhibit villus blunting, crypt hyperplasia, or develop intestinal tumors in the 13-week and 2-year NTP studies." 13. "No aberrant foci were observed in duodenal H&E stained sections from control mice or mice exposed to 0.1–180 mg/l Cr(VI) for 13 weeks." 14. "Herein, we clearly showed that high concentrations of Cr(VI) increase the number of crypt enterocytes and length of the crypt compartment after 13 weeks of exposure." 15. "The U.S. EPA regional screening level (RSL) of 0.035 μg/l for Cr(VI) in residential tap water has been proposed based on the assumption of a mutagenic MOA, that is, that the biology and risk of intestinal tumor formation from Cr(VI) exposure is linear."
Whole adult organism transcriptional profiling of acute metal exposures in male zebrafish.	Naissan Hussainzada; John A Lewis; Christine E Baer; Danielle L Ippolito; David A Jackson; Jonathan D Stallings 10.1186/2050-6511-15-15	1. "Primary mechanisms of metal toxicity include the production of free radicals which can trigger oxidative stress, induce mutagenesis by DNA-metal interactions, and impair protein function by covalently modifying proteins or competing with metal binding sites." 2. "Metal-derived reactive oxygen species (ROS) may perturb a number of tightly regulated cellular processes (e.g., cell growth and proliferation), activate transcription factors and genes, and trigger cellular adaptive programs including metal

(*continued on next page*)

Table 3 (*continued*)

Paper title	Authors and DOI	Extracted quotations
		stress response, DNA repair mechanisms, and inflammation." 3. "Exposure to all three concentrations of chromium histopathologically affected the gills, intestine, and pharynx." 4. "Both gill and pharynx epithelium exhibited mononuclear cell infiltration, which is indicative of acute inflammation." 5. "The most prominent change in intestine was moderate to moderately severe atrophy of the mucosal folds and a mild infiltration of the intestinal lamina propria with mononuclear cells." 6. "Histopathology confirmed that zebrafish exposed to all three concentrations of cobalt presented with respiratory tract lesions specific to the olfactory epithelium." 7. "Chromium redox cycling and ROS generation induce DNA damage and activate subsequent repair mechanisms." 8. "Chromium up-regulated p53 and MAPK signaling pathways." 9. "Genes associated with inflammation and acute phase stress responses were up-regulated more with chromium than the other metals." 10. "Nickel poisoning causes oxidative damage to DNA and inhibits antioxidant defenses." 11. "Chromium poisoning was associated with a marked down-regulation in genes involved in cellular metabolism, including lipid and steroid metabolic pathways." 12. "Chromium poisoning caused down-regulation of an entire set of genes encoding important regulators of energy metabolism, glucose, cholesterol, amino acid, and fatty acid metabolism and transport in many tissues, including the intestine and liver." 13. "Robust, differential down-regulation of key mediators of glucose metabolism (g6pca, gys2), lipid metabolism (fabp1), and canalicular bile acid transport (abcc2) suggest metabolic perturbations in the liver and/or gut-liver axis."

Table 3 (*continued*)

Paper title	Authors and DOI	Extracted quotations
		14. "Nickel induced transcription factor changes consistent with redox signaling, including up-regulation of hif1α and xbp1 and their associated gene targets." 15. "Cobalt showed more up-regulation of genes regulating the biological processes of the oxidative stress response than any other metal." 16. "Chromium also up-regulated cell cycle regulation and apoptosis genes not enhanced in the other metals, including genes involved in the G1/S and G2/M cell-cycle checkpoints and in the RAD6-dependent DNA repair pathway." 17. "The metals showed a significant discrepancy in the type and direction of the regulation of the transcription factors expressed." 18. "Common to all metals was the up-regulation of the highly conserved mini-chromosome maintenance 4 (mcm4) gene with DNA helicase activity essential for the inhibition of eukaryotic genome replication and the origin recognition complex (orc61) which facilitates replication." 19. "Nickel poisoning up-regulated other genes critical for redox sensing and homeostasis, including periredoxin (zgc:110343), thioredoxin (zgc:92903), thioredoxin-like 1 (txnl1), and protein disulfide isomerase 4 and 5 (pdia4, pdip5)." 20. "Chromium is readily absorbed by the intestinal tissue. Since gut mucosa represents the primary site for whole-body amino acid metabolism, mucosal atrophy can significantly diminish the gut's amino acid metabolic capacity and decrease amino acid requirements and use." 21. "Nickel dose-dependently increased expression of hspa5."

papers, identified 30,000 relevant publications, and extracted 48,000 quotations from 4500 open-access papers, achieving comparable efficiency to manual methods but at a significantly higher processing speed. By serving as a complementary tool to traditional manual screening methods, it provides toxicologists with a systematic means of validating their findings and ensuring comprehensive literature coverage. The primary value lies in augmenting expert judgment through efficient information retrieval and organization.

The LLMs-rev successfully addressed key challenges in scientific literature review through its fast and efficient MoA extraction process across multiple compounds. By automating the retrieval of relevant papers from PubMed, evaluating their relevance, and extracting MoA information from open-access publications, significant acceleration of the risk assessment process was achieved while maintaining comprehensive scientific literature coverage. No significant advantages over manual screening were demonstrated by this automated method for compounds with very limited available literature or those with multiple proposed MoAs. For non-open-access papers, the system's capability to process uploaded PDF documents by users with appropriate credentials could ensure that valuable information is not overlooked, thereby enhancing the completeness of the MoA extraction process. This integration of automated extraction with human expertise represents a significant step toward providing more efficient and thorough toxicological assessments.

The reproducibility and systematic nature of this approach address a fundamental challenge in toxicological research: the need for consistent and comprehensive literature evaluation across multiple compounds. As regulatory requirements for chemical safety assessment continue to expand, such tools become increasingly valuable for maintaining scientific rigor while managing growing information volumes.

The versatility of the presented method extends beyond its immediate application in MoA retrieval. With appropriate prompt refinement, the same methodological framework could be adapted for various applications requiring scientific literature assessment and specific information extraction. GPT-4o demonstrated the capability to process large volumes of textual information efficiently, opening new possibilities for automated information extraction from diverse document types, including legislation, institutional reports, and scientific literature. This capability could significantly accelerate research processes that traditionally require extensive human resources and time investment.

Despite its significant advantages and potential, this approach does have some important limitations. In its current form, the LLMs-rev system serves primarily as an auxiliary tool for expert toxicological assessment, particularly for compounds with complex MoAs requiring interpretative analysis of figures and contextual data. Additionally, rigorous safeguards must be maintained to prevent model extrapolation or hallucination, which has been effectively achieved by restricting outputs to direct quotations from verified sources. In information extraction, a hallucination occurs when the LLM generates information that is not present in the source papers or is factually incorrect according to the provided text. Extrapolation occurs when the LLM takes a valid finding from the paper and projects it beyond its logical or data-supported bounds. The model "reads between the lines". The results of the MoA extraction also depend on the comprehensiveness of the source database and could be implemented using other databases, eventually using non-English literature, which will also represent a further advantage of this workflow, since direct translation could be included in the process to deal with literature that would not be considered by experts who do not master the publication's language. Further developments could also include the iterative inclusion and exclusion of mechanism, events and effects keywords within the extraction prompt, as well as a contextualization of the extracted quotations within the paper. Finally, this pipeline allows only the retrieval of textual information due to the limitation of full text retrieval process, an issue which will most probably be tackled by the newest LLM architectures and alternative full text retrieval methods.

In terms of reliability, the automated approach based on LLMs-rev aligns with findings from previous studies, demonstrating significant improvements in screening efficiency. For instance, Hardy et al. (2024) reported that AI tools offer clear efficiency benefits in systematic literature reviews, although the impact on accuracy varies. LLM-rev offers a significant advantage in speed compared to conventional literature review processes, which require validation by two human reviewers. Conventional methods typically allow a single validator to assess the relevance of 100 to 120 papers per hour. In contrast, this AI script processed more than 8000 papers per hour, a figure that includes the highly time-intensive task of information retrieval from open texts. This value is highly dependent on the hardware infrastructure hosting the LLM-rev process. The efficiency of those traditional screening tools and their built-in AI features or added machine learning have been discussed by Gartlehner et al. (2019), Gates et al. (2019); Chappell et al. (2023), highlighting the potential risk of missing important and relevant papers in the process. In the presented AI technique, this risk is mitigated by an accurate optimization of the prompt, which remains the key role of the human user. This has been further corroborated by the manual screening for several compounds and a comparison made with the relevant papers retrieved by the LLMs-rev procedure. Ample detail on this can be found in Kremer et al. (2026).

In conclusion, the LLMs-rev automated approach to MoA extraction represents a valuable addition to the toxicologist's toolkit, offering enhanced efficiency, reproducibility, and comprehensiveness in scientific literature review. However, the current implementation requires expert oversight at multiple stages, and formal quantitative validation against annotated gold-standard datasets remains an essential next step before broader deployment in regulatory workflows can be recommended. Future developments including multimodal input processing, cross-lingual retrieval, and decision-support agent integration hold considerable promise for further advancing this methodology. By combining the processing power of LLMs with expert human oversight, this methodology advances the field toward more systematic, objective, and thorough toxicological assessments while maintaining scientific rigor and accuracy.

Disclaimer

The present article is published under the sole responsibility of the authors and may not be considered as an EFSA scientific output. The positions and opinions presented in this article are those of the authors alone and are not intended/do not necessarily represent the views/any official position or scientific works of EFSA. To know about the views or scientific outputs of EFSA, please consult its website under http://www.efsa.europa.eu.

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

During the preparation of this work the author(s) used GPT-4o in order to extract and analyze toxicological mechanism of action information from scientific literature. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding body information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Arnaud Molle: Formal analysis, Investigation, Software, Writing – original draft. **Nelly D. Saenen:** Data curation, Validation, Writing – review & editing. **Karen Smeets:** Data curation, Methodology, Validation. **Karolien Bijnens:** Data curation, Validation, Writing – review & editing. **Marc Aerts:** Methodology, Supervision, Writing – review & editing. **Cécile Kremer:** Formal analysis, Methodology, Writing – review & editing. **Efisio Solazzo:** Funding acquisition, Supervision, Writing – review & editing. **José Cortiñas Abrahantes:** Conceptualization, Methodology, Supervision, Writing – review & editing.

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.

Data availability

Data will be made available on request.

References

Blümmel, T., Rehn, J., Mereu, C., Graf, F., Bazing, F., Kneuer, C., Sonnenburg, A., Wittkowski, P., Padberg, F., Bech, K., Eleftheriadou, D., van der Lugt, B., Kramer, N., Bouwmeester, H., Dobrikov, T., 2024. Exploring the use of Artificial Intelligence (AI) for extracting and integrating data obtained through New Approach Methodologies (NAMs) for chemical risk assessment. EFSA Support.Pub. 2024 (1). https://doi.org/10.2903/sp.efsa.2024.EN-8567. EN-8567.

Blümmel, T., Rehn, J., Mereu, C., Graf, F., Kneuer, C., Wittkowski, P., Sonnenburg, A., Moore, A., Bech, K., van der Lugt, B., Bouwmeester, H., Kramer, N., Dobrikov, T., 2023. Review of state-of-the-art AI tools and methods for screening, extracting and evaluating NAMs literature in the context of chemical risk assessment. EFSA Support. Pub. 20 (1), 7815E. https://doi.org/10.2903/sp.efsa.2022.EN-7815.

Bornmann, L., Haunschild, R., Mutz, R., 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Hum. Soc. Sci. Commun. 8, 224. https://doi.org/10.1057/s41599-021-00903-w.

Chappell, L., Simmonds, M., O'Mara-Eves, A., Le Couteur, D., Roberts, N., Sutton, A., Weightman, A., Thomas, J., 2023. Machine learning for accelerating screening in evidence reviews. Cochrane Evid. Synth.Method. 1 (5). https://doi.org/10.1002/cesm.12021.

Cheong, B.C., 2024. Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. In: Frontiers in Human Dynamics, 6, 1421273. https://doi.org/10.3389/fhumd.2024.1421273.

Chustecki, M., 2024. Benefits and risks of AI in health care: narrative review. In: Interactive journal of medical research, 13, e53616. https://doi.org/10.2196/53616.

Delgado-Chaves, F.M., Jennings, M.J., Atalaia, A., Wolff, J., Horvath, R., Mamdouh, Z. M., Baumbach, J., Baumbach, L., 2025. Transforming literature screening: the emerging role of large language models in systematic reviews. Proc. Natl. Acad. Sci. 122 (2), e2411962122. https://doi.org/10.1073/pnas.2411962122.

European Food Safety Authority, 2024a. Consolidated Annual Activity Report 2023. https://www.efsa.europa.eu/sites/default/files/2024-03/efsa-consolidated-annual-activity-report-2023_0.pdf.

European Food Safety Authority, 2024b. EFSA's activities on emerging Risks in 2023. EFSA Support.Pub. 21 (12). https://doi.org/10.2903/sp.efsa.2024.EN-9198. EN-9198.

European Parliament and Council of the European Union, 2019. Regulation (EU) 2019/1381 of 20 June 2019 on the transparency and sustainability of the EU risk assessment in the food chain. Off. J. Eur. Union L 231/1. https://eur-lex.europa.eu/eli/reg/2019/1381/oj.

Fantini, D., 2019. *easyPubMed: search and retrieve scientific publication records from PubMed* (R package version 2.13). https://CRAN.R-project.org/package=easyPubMed.

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K., 2024. Bias and fairness in large language models: a survey. Comput. Linguist. 50 (3), 1097–1179. https://doi.org/10.1162/coli_a_00524.

Gartlehner, G., Affengruber, L., Titscher, V., Noel-Storr, A., Dooley, G., Ballarini, N., König, F., 2019. Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. Syst. Rev. 8, 277. https://doi.org/10.1186/s13643-019-1221-3.

Gates, A., Johnson, C., Hartling, L., 2019. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst. Rev. 8, 278. https://doi.org/10.1186/s13643-019-1222-2.

Gupta, S., Mahmood, A.U., Shetty, P., Adeboye, A., Ramprasad, R., 2024. Data extraction from polymer literature using large language models. Commun. Mater. 5, 269. https://doi.org/10.1038/s43246-024-00708-9.

Hardy, E.J., Jenkins, A.R., Ross, J., Lang, S., 2024. *Accuracy and efficiency of automated or artificial intelligence tools in systematic literature reviews: a systematic literature review* [poster presentation]. https://www.ispor.org/heor-resources/presentations-database/presentation/euro2024-4018/144489.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T., 2025. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. 43 (2), 42. https://doi.org/10.1145/3703155.

Kremer, C., Saenen, N.D., Bijnens, K., Faes, C., De Boever, P., Smeets, K., Aerts, M., 2026. Compilation of NTP and published dose-response data: study of informative priors in the EFSA platform for Bayesian benchmark dose analysis. EFSA Support.Pub. 23 (2), EN-9902. https://doi.org/10.2903/sp.efsa.2026.EN-9902.

Landhuis, E., 2016. Scientific literature: information overload. Nature 535 (7612), 457–458. https://doi.org/10.1038/nj7612-457a.

Mishra, T., Sutanto, E., Rossanti, R., Pant, N., Ashraf, A., Raut, A., Uwabareze, G., Oluwatomiwa, A., Zeeshan, B., Faiz, M.F., Adapa, S., Uddin, M.N., Shaik, H., Giri, R., Katageri, B., Patne, N., Kamath, R., 2024. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. Sci. Rep. 14, 31672. https://doi.org/10.1038/s41598-024-81370-6.

OpenAI, 2024. GPT-4o system card (arXiv:2410.21276). arXiv. https://doi.org/10.48550/arXiv.2410.21276.

Peters, U., Chin-Yee, B., 2025. Generalization bias in large language model summarization of scientific research. R. Soc. Open Sci. 12 (4), 241776. https://doi.org/10.1098/rsos.241776.

R Core Team, 2023. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Ren, S., Tomlinson, B., Black, R.W., et al., 2024. Reconciling the contrasting narratives on the environmental impact of large language models. Sci. Rep. 14, 26310. https://doi.org/10.1038/s41598-024-76682-6.

Shaib, C., Li, M., Joseph, S., Marshall, I., Li, J.J., Wallace, B.C., 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1387–1407. https://doi.org/10.18653/v1/2023.acl-short.119.

Shi, K., Zhao, Y., 2024. Integration of advanced large language models into the construction of adverse outcome pathways: opportunities and challenges. Environ. Sci. Technol. 58 (35), 15355–15358. https://doi.org/10.1021/acs.est.4c07524.

Sonnenburg, A., van der Lugt, B., Rehn, J., Wittkowski, P., Bech, K., Padberg, F., Eleftheriadou, D., Dobrikov, T., Bouwmeester, H., Mereu, C., Graf, F., Kneuer, C., Kramer, N.I., Blümmel, T., 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. https://doi.org/10.1016/j.tox.2024.153933.

Tóth, B., Berek, L., Gulácsi, L., Péntek, M., Zrubka, Z., 2024. Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed. Syst. Rev. 13 (1), 174. https://doi.org/10.1186/s13643-024-02592-3.

Ushey, K., Allaire, J.J., Tang, Y., 2023. *Reticulate: interface to 'Python'* (R package version 1.31.0). https://CRAN.R-project.org/package=reticulate.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., 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., et al., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.

Wickham, H., Hester, J., Ooms, J., 2023. *xml2: parse XML* (R package version 1.3.5). https://CRAN.R-project.org/package=xml2.

Winter, D.J., 2021. Rentrez: an R package for the NCBI eUtils API. R J. 9 (2), 520–526. https://doi.org/10.32614/RJ-2017-058.