Environment International 197 (2025) 109341

Contents lists available at ScienceDirect

Environment International

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

Full length article

Using artificial intelligence tools for data quality evaluation in the context of microplastic human health risk assessments

Yanning Qiu *, Svenja Mintenig, Margherita Barchiesi, Albert A. Koelmans

Aquatic Ecology and Water Quality Management Group, Wageningen University and Research, P.O. Box 47 6700 AA Wageningen, the Netherlands

ARTICLE INFO

Handling Editor: Adrian Covaci

Keywords:
Large language models
ChatGPT
Gemini
Quality Assurance and Quality Control

ABSTRACT

Concerns about the negative impacts of microplastics on human health are increasing in society, while exposure and risk assessments require high-quality, reliable data. Although quality assurance and –control (QA/QC) frameworks exist to evaluate the reliability of data for these purposes, manually assessing studies is too time-consuming and prone to inconsistencies due to semantic ambiguities and evaluator bias. The rapid growth of microplastic studies makes manually screening relevant data practically unfeasible. This study explores the potential of artificial intelligence (AI), specifically large language models (LLMs) such as OpenAI's ChatGPT and Google's Gemini, to streamline and standardize the QA/QC screening of data in microplastics research. We developed specific prompts based on previously published QA/QC criteria for the analysis of microplastics in drinking water and its sources, and used these to instruct AI tools to evaluate 73 studies published between 2011 and 2024. Our approach demonstrated the effectiveness of AI in extracting relevant information, interpreting the reliability of studies, and replicating human assessments. The findings indicate that AI-assisted assessments show promise in improving speed, consistency and applicability in QA/QC tasks, as well as in ranking studies or datasets based on their suitability for exposure and risk assessments. This groundbreaking application of LLMs in the environmental sciences suggests that AI can play a vital role in harmonizing microplastics risk assessments within regulatory frameworks and demonstrates how to meet the demands of an increasingly data-intensive application domain.

1. Introduction

Public concern is growing about the potential risks that microplastics may pose to human health (Thompson et al., 2024; Wright & Kelly, 2017). Accurately assessing these risks requires high-quality data that are specifically suited for analyzing exposure levels and health effects (Koelmans et al., 2022). However, the variability and inconsistency in analytical methods used across studies have made it challenging to assess data reliability accurately, emphasizing the need for standardized approaches for data quality screening. To address this issue, several quality assurance and −control (QA/QC) tools have been developed to ensure the reliability of analytical data used in exposure assessments (Hermsen et al., 2018; Koelmans et al., 2019; Redondo-Hasselerharm et al., 2023; WHO, 2019; Wright et al., 2021). These QA/QC tools cover various aspects of the analytical procedure, such as sampling methods, sample treatment, lab preparation, use of controls and polymer identification (De Ruijter et al., 2020; Hermsen et al., 2018; Koelmans et al., 2019). For each of these aspects, criteria are defined, allowing scores of '0' (not reliable), '1' (reliable with restrictions), or '2' (fully reliable) to be assigned (Hermsen et al., 2018). The total accumulated score (TAS) reflects the completeness of information provided by a study; however, a single 'zero' renders the data unreliable, as all criteria are considered essential. These scores facilitate the quantitative comparison of data reliability and the harmonization of microplastic risk assessments across regulatory frameworks.

However, applying these QA/QC tools usually requires a thorough review of the publication regarding multiple defined criteria. As the number of studies increases rapidly, manual assessments become increasingly impractical and time consuming. For instance, over 1000 studies on human microplastic exposure have been published over the past three years (Supplementary Text S1). Keeping up with this pace to perform QA/QC scoring on each study individually is nearly impossible.

In addition to the significant time investment required for scoring, other challenges persist with human-based scoring. For example, different researchers or research groups may arrive at varying results due to differences in prior knowledge, inconsistencies in the semantic

* Corresponding author.
E-mail address: yanning.qiu@wur.nl (Y. Qiu).

https://doi.org/10.1016/j.envint.2025.109341
Received 24 December 2024; Received in revised form 5 February 2025; Accepted 17 February 2025
Available online 17 February 2025

understanding of criteria, variations in word interpretation, and overall ambiguity in the data quality assessment tasks. Finally, we emphasize that the previously developed QA/QC evaluation tools only pertain to usability in risk assessment, based on specifically constructed criteria. However, they say nothing about the quality of the data or studies as such, in relation to the research questions as formulated by the authors. In this context, it is an ethical principle that QA/QC screening is carried out as consistently and accurately as possible.

Recent advancements in artificial intelligence (AI), such as generative AI and large language models (LLMs), have the potential to facilitate screening of whether and how studies incorporated QA/QC considerations into their work. Several chatbots utilizing LLMs have been released and have demonstrated significant potential in processing language and generating meaningful responses for various purposes. For instance, OpenAI's ChatGPT and Google's Gemini can analyze large volumes of text data and produce coherent outputs (Ali et al., 2023). The potential applications of these tools in research have been extensively discussed, demonstrating proven utility in tasks like text annotation (Gilardi et al., 2023), understanding semantics (Le Mens et al., 2023), and multilingual psychological text analysis (Rathje et al., 2024). To date, most research applications of LLMs have been concentrated in the field of social sciences, whereas in the natural sciences, only a few studies have attempted to apply LLMs to extract information from scientific literature (Dagdelen et al., 2024; Hu et al., 2024; Liang et al., 2024; Zheng et al., 2023) or to answer scientific questions (Jablonka et al., 2024; Zhu et al., 2024). The potential of LLMs in quantitative natural sciences remains to be explored further (Wu et al., 2024; Zhu et al., 2023). Applying these tools in environmental research, particularly for screening the implementation of QA/QC criteria – an inherently text-focused task – could be especially beneficial.

In this study, we aimed to test the applicability of LLMs in assessing data reliability of existing microplastic research in the context of exposure and risk assessment. We developed prompts based on our previous QA/QC criteria for the analysis of microplastics in freshwater and drinking water (Koelmans et al., 2019), and used these prompts to instruct AI tools, namely ChatGPT and Gemini. We tested the performance of these LLMs in 1) extracting relevant text from scientific publications and 2) interpreting information and assessing the data reliability based on human-provided instructions. We validated the performance of LLMs by comparing AI-generated evaluations with those conducted by human professionals using the publication set from our previous study (Koelmans et al., 2019). Additionally, we applied the LLMs to evaluate the publications on microplastics in drinking water that were published between 2021 and 2024. Our study provides a first assessment of the capability of LLMs in performing QA/QC screening tasks in microplastic research, with direct relevance for evaluating human exposure to microplastics through drinking water.

2. Methods

2.1. Literature datasets

This study assessed two datasets focusing on microplastics in drinking water and its sources. Microplastics in drinking water is a critical area for human risk assessment, since drinking water is one of the primary sources of microplastic ingestion (WHO, 2019). Dataset 1 has undergone thorough human assessments in our previous study (Koelmans et al., 2019), and was used to develop and test the applicability of the AI screening tool. After validating the AI screening tool's performance, we applied the tool to Dataset 2, a more recent dataset that had not been previously assessed manually.

2.1.1. Dataset 1

This dataset includes forty-three studies reporting microplastic concentrations in drinking water (1 tap water, 3 bottled water, 1 water from drinking water treatment plants) or its freshwater sources (26 surface water and 12 water from wastewater treatment plants). These publications were retrieved from the Scopus database using the search string: microplastic AND (bottle OR surface OR tap OR wastewater OR groundwater), and were reviewed in our previous study (Koelmans et al., 2019). In the present study, we kept only peer-reviewed publications in English, and excluded studies that addressed multiple water types (e.g. Vermaire et al., 2017, which analyzed both surface water and effluent samples from wastewater treatment plants) or were from grey literature sources. For more information regarding the inclusion standard and search procedure we refer to Koelmans et al. 2019. For a comprehensive list of selected papers see Supplementary Text S2.

2.1.2. Dataset 2

This dataset includes thirty studies published between 2021 and 2024, focusing on microplastic concentrations in drinking water and its direct sources such as bottled water, tap water, well water and water from drinking water treatment plants. These studies were retrieved from the Scopus database using the search string: microplastic AND (tap water OR bottled water OR mineral water OR drinking water OR well water OR drinking water treatment plants) AND human. The search was conducted in July 18, 2024. Only peer-reviewed publications in English (i.e., no grey literature) that focused on a single water type, consistent with the inclusion criteria for Dataset 1, were included. For a comprehensive list of selected papers see Supplementary Text S3.

For the studies from both datasets, we downloaded the main text and supplementary materials in PDF format directly from the publishers. We then converted the PDFs into plain text for analysis using Python 3.10 and the PyPDF2 package (Fenniak et al., 2022). To maximize LMM performance, we retained only the research-related sections (e.g., Introduction, Methods, Results, and Discussion) of the main text by segmenting the text and excluding non-research-related sections. The non-research-related sections were identified by searching for specific keywords such as 'References', 'Author Contributions', and 'Acknowledgements'. All supplementary materials from selected publications were included in the screening process.

2.2. QA/QC screening for data reliability assessment

We implemented screening tasks following the nine criteria designed by our previous study (Koelmans et al., 2019). Specifically, the following criteria were included and screened: 1) sampling methods, 2) sampling size, 3) sample processing and storage, 4) laboratory preparation, 5) clean air conditions, 6) negative controls (blanks), 7) positive controls (recovery tests), 8) sample treatment and 9) polymer identification. These criteria were proposed to be crucial for evaluating the reproducibility of described methods, precision, accuracy and sensitivity in the analysis and detection of microplastics, and together determine the robustness of an applied method and reliability of a study (Koelmans et al., 2019; WHO, 2019). As mentioned, for each criterion, a score of 2, 1, or 0 was assigned to the publication during screening (Hermsen et al., 2018; Koelmans et al., 2019), with the TAS of the nine criteria reflecting the completeness of the information. For data to be considered sufficiently reliable, a study should have no 'zero' values for any of the individual scores as all nine criteria are considered essential. For a detailed description, argumentation and the scoring scheme for the criteria, we refer to Koelmans et al. 2019.

2.3. AI tools and implementation

We conducted screening tasks in October and November 2024 using ChatGPT API version *gpt-4o* and the Gemini API version *gemini-1.5-pro*, hereafter referred to as GPT and Gemini, respectively. We also tested OpenAI *o1-mini*, one of the recently released reasoning models. However, we did not find an improvement in performance, which we attribute to the fact that we use very detailed instructions, apparently reducing the added value of the higher reasoning capacity of reasoning models such as *o1-mini*. For each criterion, we prompted both models with prompts derived from the criterion description and scoring scheme in Koelmans et al. 2019.

A well-structured prompt with clear logic and detailed rules is useful to guide the responses of LLMs, and to counteract the challenges of machine hallucinations (Liang et al., 2024; Tonmoy et al., 2024). In this study, we designed the prompts consisting of the following parts: 1) plain text from a publication, 2) research type identification (i.e., the source of the water sample, such as surface water, ground water, tap water, bottled drinking water or water from a water treatment plant), 3) description of the criterion and text extraction task, 4) scoring scheme of the criterion and scoring task, and 5) structured output formats. We asked LLMs to first decide which research type the publication belongs to because the scoring schemes of a few criteria are related to different research types. For full prompts per criterion we refer to Supplementary Text S4.

LLMs offer several hyperparameters that can be tuned to influence their output. One such parameter is the 'temperature', which controls the randomness of the model's responses. Higher temperature values produce more diverse and creative outputs, while lower values result in more deterministic and focused responses that might be better-suited for scientific contexts (Chen & Ding, 2023; OpenAI, 2024; Roemmele & Gordon, 2018). To evaluate the impact of different parameter settings, we tested both GPT and Gemini on Dataset 1, using two temperature settings: 1 (the default value, which balances randomness and determinism) and 0.2 (a lower value that produces more deterministic and focused outputs). For each temperature setting, we collected three responses from each model. To ensure that the AI assessments were not influenced by prior interactions, we initiated a new chat session for every publication and every criterion.

2.4. Expert-knowledge-based data reliability assessment

We used the assessment results from Koelmans et al., 2019 as a reference standard for Dataset 1. The assessments were performed in 2018 by trained microplastic experts following the same criteria and scoring scheme as in this study. Each criterion for each paper (43 papers × 9 criteria) was assessed and discussed by at least three people before a final score was assigned.

In addition, for Dataset 2, we selected and evaluated 107 individual criteria × paper combinations for validation purposes and to improve AI tool applicability. The assessments were conducted by the same group of experts using the same assessment criteria. For more details regarding this part of the assessments, we refer to the ***2.6 AI Tool Applications*** section.

2.5. Evaluation metrics

Dataset 1 was used for evaluating the LLM performance at different model settings. We evaluated the LLM performance in two steps, corresponding to the set tasks. First, we assessed whether the LLMs accurately detected and extracted the relevant text in the paper. Second, we evaluated whether the LLMs reasonably scored the criteria based on the extracted text and criteria instructions. We also assessed whether the LLMs could effectively rank the relative reliability of publications by comparing the TAS for each publication from AI assessments with those from human-assessments.

For the text extraction task, we found that LLMs usually extract the same or very similar text across trials with the same model settings. Therefore, to reduce redundancy, we randomly selected one trial per criterion × paper combination for each model setting and manually checked whether the LLMs correctly extracted the relevant text. A score of 1 was assigned for correct extractions, and a score of 0 was assigned for incorrect extractions (wrong, incomplete or no relevant text detected). We calculated the accuracy rate of text extraction per criterion as the proportion of correct extractions to the total number of extractions for each model.

For the scoring task, we evaluated the LLM performance using three metrics: intercoder agreement, consistency with human assessment results, and F2 score of the non-zero score detections. The intercoder agreement was calculated as the percentage of cases where the same score was assigned across all three trials, relative to the total number of cases. For example, if a paper receives the same score for criterion *A* across all three trials, it is considered an intercoder consistent case for criterion *A*. The intercoder agreement thus reflects stability and robustness of the LLMs, demonstrating its ability to maintain consistent judgments across different trials (Gilardi et al., 2023). In addition, intercoder agreement can also indicate the level of complexity in a task, as tasks with higher levels of difficulty or ambiguity are more likely to result in lower intercoder agreement rates under the same temperature settings. We further compiled a final score sheet by taking the majority score across the three trials for each criterion and each paper. The consistency with human assessment results was calculated as the percentage of final scores that matched expert-knowledge-based data quality assessments. Both the intercoder agreement and consistency were calculated per criterion.

Furthermore, in practice, it is crucial to identify the lack of reliability of research in a risk assessment context, i.e., distinguish between the zero (unreliable) and non-zero (2 or 1, reliable or partially reliable) scores. As mentioned, a single zero score renders the data unreliable, as all criteria are considered essential. Therefore, misclassifying non-zero scores as zero is a critical issue to avoid. Given this priority, we emphasize the importance of correctly identifying non-zero scores while ensuring minimal errors in classifying zero scores. To address this, we calculated the F2 Score of non-zero score detections (hereafter referred to as the F2 Score) as another key evaluation metric. The F2 Score is calculated based on precision and recall, the formulas are as follows:

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \tag{1}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \tag{2}$$

$$F2score = (1+2^2) \times \frac{Precision \times Recall}{(2^2 \times Precision) + Recall} \tag{3}$$

True Positives refer to instances where the LLM correctly predicts a non-zero score (reliable), while False Positives refer to instances where the LLM incorrectly predicts a zero score (unreliable) as a non-zero score (reliable). False Negatives, on the other hand, occur when the LLM incorrectly predicts a non-zero score (reliable) as a zero score (unreliable). The F2 Score integrates precision and recall, with an emphasis on recall to reflect the priority of minimizing the risk of misclassifying non-zero scores (reliable) as zero (unreliable). A higher F2 Score indicates better performance, as it reflects the LLM's ability to correctly classify non-zero scores (reliable cases) while maintaining accuracy in identifying zero scores (unreliable cases). We calculated the F2 Score using all 387 assessment cases (43 papers × 9 criteria) rather than per criterion because the distribution of zero scores is highly uneven across criteria. For example, there are 35 zero scores for the positive controls criterion but only 2 for the sampling methods criterion based on expert-knowledge based assessments results. This uneven distribution could bias the F2 Score calculation if done per criterion (explanation with examples see Supplementary Text S5).

In addition, we compared the ranking of AI assessments with those from human assessments, as this provides an intuitive measure of the LLM's overall capability in screening tasks. To achieve this, we computed the TAS of each paper as the sum of the score from 9 criteria, both for human assessment results and AI assessment results. The consistency between AI assessments and human assessments of TAS was computed using the Pearson correlation test.

2.6. AI tool applications

We applied the AI QA/QC screening to Dataset 2, a recent dataset that had not been manually assessed prior to this study. We conducted three assessment trials and selected the majority score as the final score, following the strategy described earlier, using the LLM that demonstrated the best performance during the screening of Dataset 1.

To assess the quality of the AI assessments, we selected a subset of the assessment tasks (71 criteria × paper cases) for human validation. The subset size was determined using Cochran's equation, a statistical formula for calculating the minimum sample size required for a study based on desired precision, confidence level, and estimated population proportion:

$$n_0 = \frac{Z^2 \times p \times (1-p)}{e^2} \tag{4}$$

where n_0 is the required sample size (before adjustment, assuming an infinite population size), Z-score is associated with the desired confidence level (e.g., 1.96 for 95 % confidence), p is the estimated proportion of the population (set to 0.5 for maximum variability) and e is the desired margin of error (e.g., 0.1 for ± 10 %). Since we have a finite number of assessment cases of Dataset 2 (30 papers × 9 criteria), we adjusted the sample size as follows:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \tag{5}$$

where n is the adjusted sample size for a finite population and N is the total population size. With a 10 % margin of error and a 95 % confidence level, the calculated sample size (rounded) was 71. Therefore, we randomly selected 71 assessment cases and had the same group of experts perform manual assessments for validation purposes.

We calculated the intercoder agreement (of all the assessment cases from Dataset 2), consistency with human assessments (of the selected subset), and the F2 score of non-zero score detections (of the selected subset) as the indicators of AI performance for Dataset 2.We propose that comparable values of these indicators of Dataset 2 to Dataset 1 suggest that the AI tool is reliably applicable for similar tasks across different datasets.

We further checked all the zero scores (50 cases) classified by the AI tool in Dataset 2 to improve the usability of the AI screening results. With this we minimize the chance of misclassifying reliable studies as unreliable ones, ensuring a reliable data reliability control for future risk assessments of microplastics' impact on human health. For the calculation of the probability of misclassification (reliable cases as unreliable and unreliable cases as reliable), we refer to Supplementary Text S6.

Since 14 of the above-mentioned zero-score cases have also been selected and checked during the random validation, in total 107 cases undergone manual checking in Dataset 2, equivalent to 39.6 % of the total assessment cases.

3. Results and discussion

3.1. Performance of the text extraction task

Correctly locating and comprehensively extracting the relevant text from a scientific publication is the basis for further interpretation and scoring tasks. Our results demonstrate that both LLMs exhibit a strong capacity for extracting relevant text following human instructions, with lower temperature settings yielding higher accuracy rates (Fig. 1). At temperature 0.2, Gemini shows an average accuracy rate of 0.925 ± 0.062 and GPT shows an average accuracy rate of 0.876 ± 0.118. At temperature 1, the accuracy rates slightly decrease to 0.919 ± 0.052 for Gemini and 0.870 ± 0.128 for GPT. Accuracy rates vary across criteria and models: 'Sample Treatment' shows the highest accuracy rate across all model and temperature settings, whereas 'Sampling Methods' shows the lowest accuracy. These variations may be attributed to differences in the levels of specificity or complexity of information required by certain criteria, as well as how the LLM interprets the descriptions of different criteria. Despite these variations, neither the temperature settings (0.2 vs. 1) nor the choice of LLM (Gemini vs. GPT) show significant differences in performance (*t*-test, all $p > 0.05$).

Although our present study is the first to explore the potential of LLMs in the field of microplastic research, a few previous studies have shown that LLMs can effectively extract and summarize information with efficiency and consistency in other scientific domains, such as chemistry (Dagdelen et al., 2024; Polak & Morgan, 2024), biology (Hu et al., 2024) and medical science (Singhal et al., 2023). Our results show promising consistency with those studies and enable further AI-assisted work, such as information interpretation and assessment.

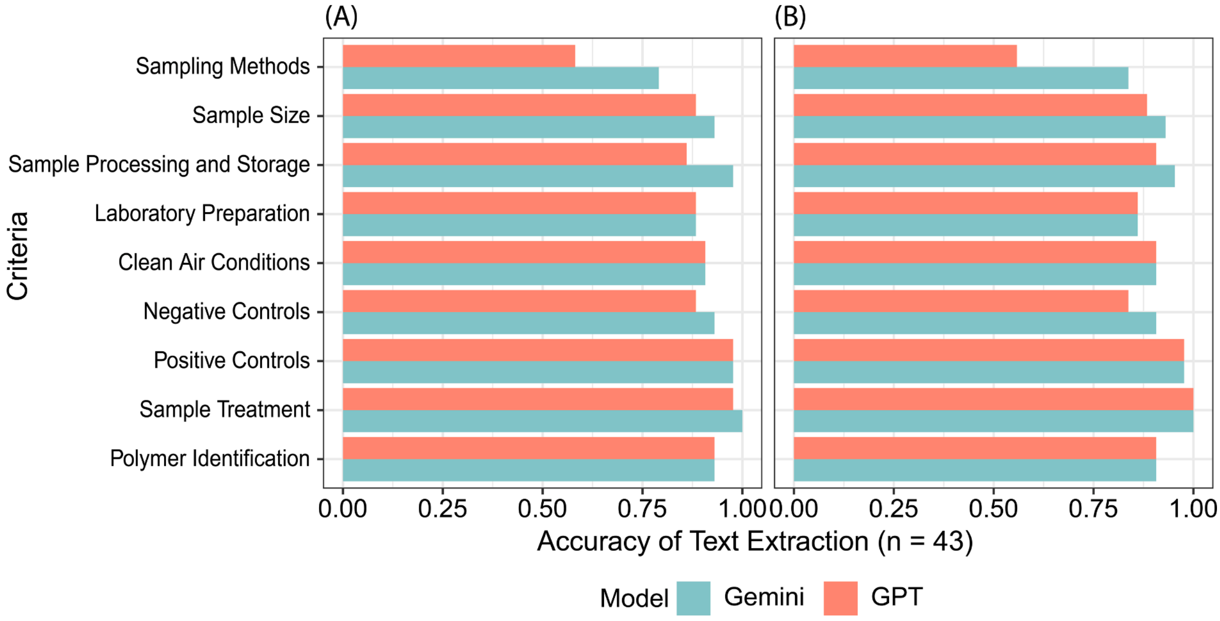

Fig. 1. Accuracy rate on text extraction tasks. Panel A: Accuracy of both models at temperature 0.2, Panel B: Accuracy of both models at temperature 1.

3.2. *Performance of the instruction-based interpretation task*

The intercoder agreement and consistency of AI assessment results compared to human assessment results are presented in Fig. 2 for different models and various temperature settings. Lower temperature settings (resulting in a more deterministic and focused model) yield slightly higher intercoder agreement, i.e. the consistency rate among three AI assessment trials. At a temperature of 0.2, Gemini achieves an average intercoder agreement of 0.822 ± 0.079, compared to GPT's 0.829 ± 0.090. In contrast, at a temperature of 1, intercoder agreement decreases, with GPT scoring 0.778 ± 0.085 and Gemini scoring 0.731 ± 0.136. Among the different criteria, 'Sample Processing and Storage' and 'Sample Size' have relatively low intercoder agreement rates whereas 'Positive Controls' has the highest intercoder agreement rate.

The Gemini model demonstrates higher consistency with human assessments, achieving an average of 0.804 ± 0.069 at temperature 0.2, outperforming GPT, which averages 0.744 ± 0.132. While Gemini's performance declines slightly at higher temperatures, GPT remains relatively unaffected. Consistencies for individual criteria range from 0.58 to 0.98 for GPT and from 0.74 to 0.93 for Gemini. Scores on the criteria 'Positive Controls' and 'Sample Treatment' consistently show high consistency with human assessments across both models and temperature settings, and 'Sample Processing and Storage' and 'Sample Size' have lower consistency rates, which is similar to the intercoder agreement results. These consistency differences are likely related to the unavoidable variations in the ambiguity of the criteria definitions. For instance, it is straightforward to determine whether a study performed a recovery test ('Positive Controls'), whereas assessing whether the samples were appropriately stored and handled to avoid contamination before analysis ('Sample Processing and Storage') is more ambiguous and open to interpretation. Similar challenges have been experienced also during the manual assessment by microplastic experts, where some criteria require more discussion than others.

The F2 score, which reflects the LLM's ability to correctly identify reliable cases (non-zero scores) with a strong emphasis on recall, are 0.961 (Gemini at temperature 0.2), 0.952 (Gemini at temperature 1), 0.926 (GPT at temperature 0.2), and 0.944 (GPT at temperature 1), respectively. The F2 score reaches its optimal value of 1 only when both recall and precision are perfect. In our case, all models achieved F2 scores close to 1, demonstrating their high effectiveness in identifying reliable cases with minimal false negatives. Among them, Gemini at temperature 0.2 achieved the closest F2 score (0.961), making it the best-performing configuration overall. In fact, Gemini at temperature 0.2 successfully detected all the reliable publications, and only

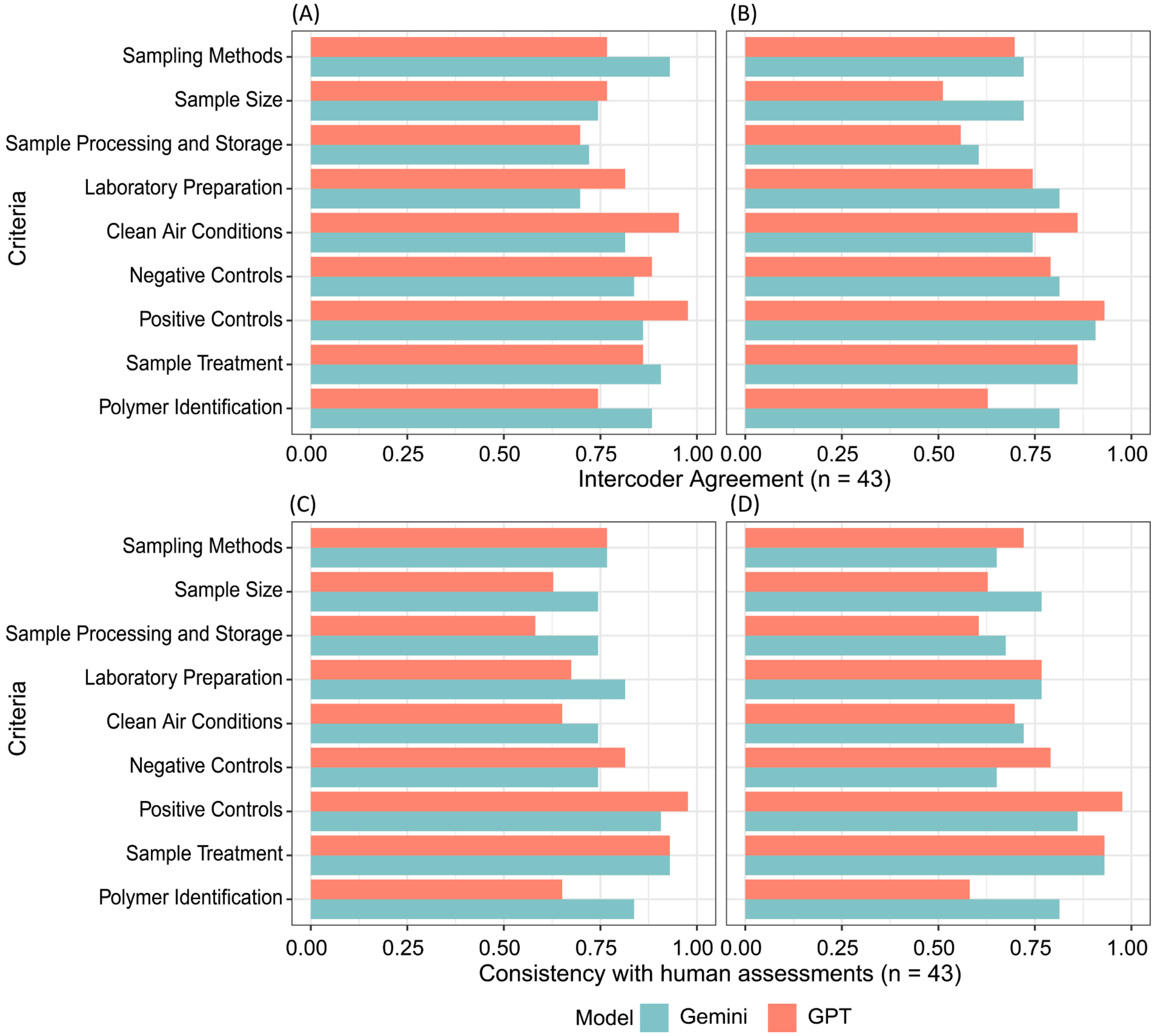

Fig. 2. Intercoder agreement and consistency with human assessment results on instruction-based interpretation tasks. Panel A: Intercoder agreement of both models at temperature 0.2; B: Intercoder agreement of both models at temperature 1; C: Consistency with human assessments of both models at temperature 0.2; D: Consistency with human assessments of both models at temperature 1.

misclassified one unreliable publications as reliable out of 43 publications from Dataset 1 (Table S1).

Overall, both models demonstrate adequate accuracy and usability, considering the challenges of defining and interpreting complex criteria under zero-shot (i.e., non-prior knowledge or scoring results given) conditions. For both intercoder agreement and consistency with human assessment results, neither different temperature settings of LLMs (0.2 vs. 1) nor different LLM implementations (Gemini vs. GPT) show significant differences (*t test*, all $p > 0.05$). However, the Gemini model and lower temperature setting exhibit a slightly better performance in our tasks.

3.3. Correlation of TAS rankings between AI and human assessments

The correlation of TAS between AI and human assessments exceeds 0.8 across all models and temperature settings (Fig. 3), suggesting a strong alignment between AI and human assessments.

While LLMs did not always replicate the exact scores of human assessments, as illustrated by the data points not fully overlapping the 1:1 line in Fig. 3, the high correlation rates indicate that both models consistently capture trends similar to human evaluations.

3.4. Application of AI assessments

We applied the Gemini model with a temperature setting of 0.2 to assess Dataset 2, which covers studies on microplastics in drinking water published between 2021 and 2024. We observed an average intercoder agreement of 0.826 ± 0.060 across different criteria (Fig. S1). A paired *t-test* revealed no significant difference in intercoder agreement between Dataset 1 and Dataset 2 ($p = 0.868$). Furthermore, the consistency rate between AI assessments and human validation was 0.803 (detailed scores are provided in Table S2). Based on a one-sample *t-test,* the consistency rate of Dataset 2 is not significantly different from that of Dataset 1 ($t = -0.957$, $p = 0.367$). The F2 score is 0.941, suggesting a high level of effectiveness in identifying reliable cases with minimal false negatives. These findings suggest that the Gemini model maintains robust performance across datasets covering similar topics, demonstrating its applicability in practical QA/QC screening.

Furthermore, we checked all zero scores classified by Gemini to minimize the misclassification of reliable publications. Among the 50

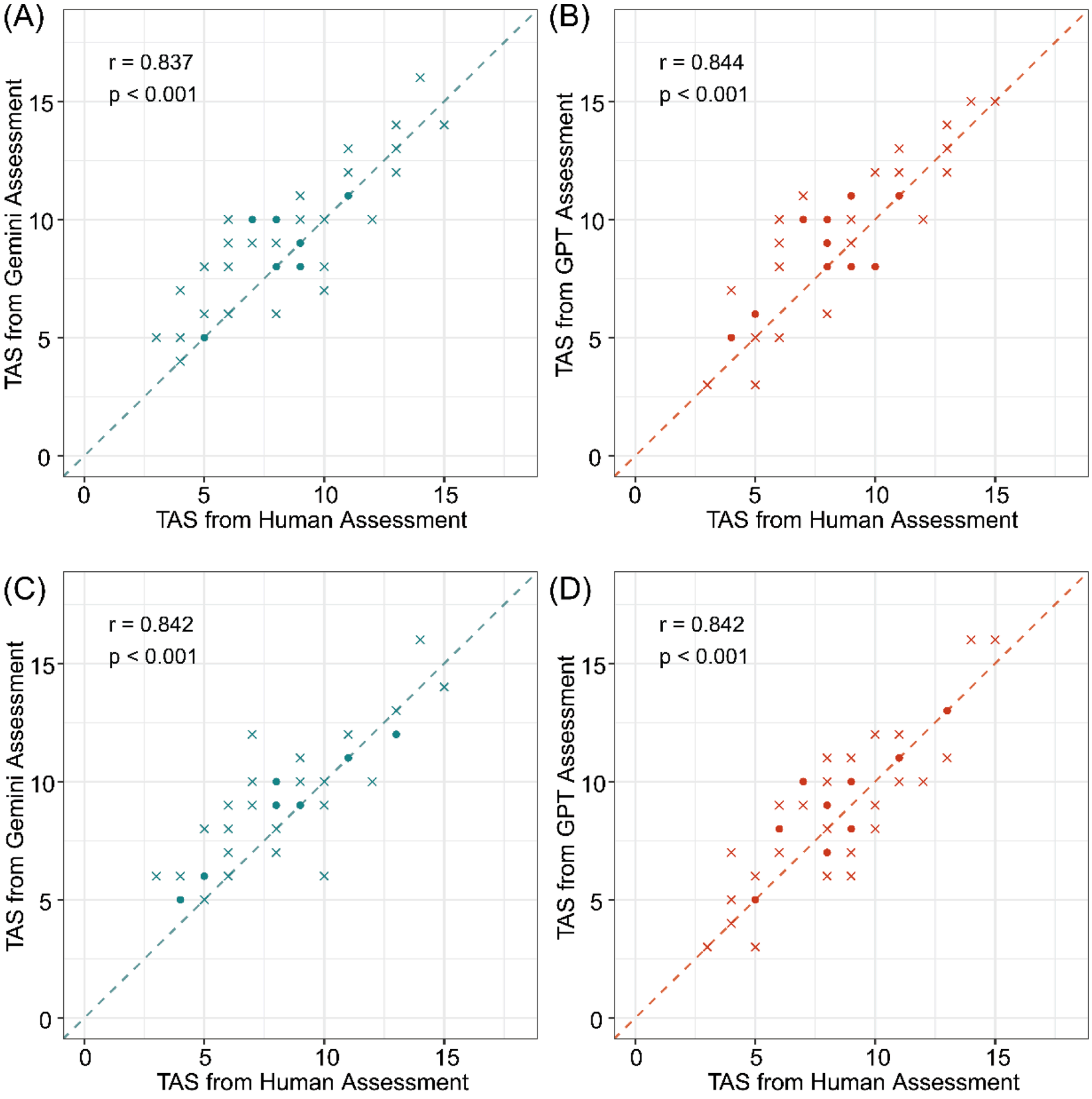

Fig. 3. Correlation of TAS from AI and human assessments at different temperature settings. Panel A: Gemini model at temperature 0.2; Panel B: GPT model at temperature 0.2; Panel C: Gemini model at temperature 1; Panel D: GPT model at temperature 1. Dots indicate overlapping cases, i.e., cases where multiple papers share the same TAS from human assessment (x-axis) and the same TAS from AI assessment (y-axis). Crosses indicate non-overlapping cases.

zero scores identified, we found only 3 cases where Gemini misclassified non-zero scores as zeros, all of which were related to 'Sample Treatment'. During the check, we observed that a greater diversity of drinking water types has been studied in recent years, potentially introducing challenges for LLMs due to their absence from our previous criteria (Table S3). For example, a few recent studies investigated drinking water from drinking water refill kiosks or fountains (Pérez-Guevara et al., 2022; Shruti et al., 2020, 2022), a type that was not covered in our screening criteria. The absence of such information may have led to errors in the AI screening process. This highlights the need of continuously updating QA/QC criteria to align with ongoing research advancements and ensure comprehensive coverage.

4. General discussion and implications

Our study presents a first application of LLMs to perform QA/QC screening tasks to assess the reliability of microplastic data for exposure- and risk assessment purposes. Our study has a main focus on microplastics in surface water and drinking water, a topic of great importance for exposure and risk assessment of microplastics on humans (WHO, 2019). However, the approach presented here can also be applied to other QA/QC criteria sets, for example, for microplastic external exposure data in other components of our diet (Cox et al., 2019; Mohamed Nor et al., 2021; WHO, 2022), for concentrations in the air (Wright et al., 2021), or for effect tests relevant to ecosystems (De Ruijter et al., 2020) or human health (Gouin et al., 2022; Wardani et al., 2024). Additionally, other QA/QC screening methods, such as Criteria for Reporting and Evaluating ecotoxicity Data or Klimisch criteria (Kase et al., 2016; Klimisch et al., 1997), can also be automated if desired.

Our results demonstrate that current LLMs can achieve strong alignment with human assessments across multiple QA/QC criteria, showing over 0.80 consistency in scoring and correlation in TAS rankings compared to human evaluations, as well as an F2 score exceeding 0.92, which reflects their capability of separating reliable cases from unreliable ones. This highlights their potential to streamline data reliability assessment tasks for specific purposes. Furthermore, in scenarios where human validation is preferred to maximize applicability, our AI-assisted approach can significantly reduce the workload. For example, in Dataset 2, only 39.6 % of the total assessment cases underwent human validation to achieve a sufficient confidence interval, a limited error rate, and minimized false negatives. In fact, this AI-assisted screening approach is unlikely to misclassify reliable cases as unreliable, and the probability of misclassifying unreliable cases as reliable is estimated to be at most 0.014 based on our calculations (Supplementary Text S6). Additionally, we emphasize that implementing LLMs in the QA/QC screening process can significantly reduce time and human effort. LLMs can assess criteria from a large number of papers within hours, with only a moderate proportion of cases requiring manual verification.

Interestingly, based on the AI assessment results, we also observed that recent studies tend to score higher with respect to the QA/QC criteria than earlier studies, as indicated by the significantly higher TAS of Dataset 2 compared to Dataset 1 (mean: 10.83 vs. 9.12; *t-test*, $p = 0.004$). Notable improvements were observed in criteria such as Sample Processing and Storage, Laboratory Preparation, Clean Air Conditions, Negative Controls, and Sample Treatment, suggesting that these criteria have been better addressed in recent research. However, some QA/QC challenges persist, particularly with criteria like Positive Controls, which received consistently low QA/QC scores across both datasets. Future research may prioritize improvements in Positive Controls and recovery tests, as these are essential for ensuring the robustness and reliability of microplastic data quality assessments.

LLMs also offer several advantages over traditional machine learning and natural language processing models. Unlike traditional models, which often require extensive training resources, high levels of coding proficiency, and large amounts of annotated data (Brady et al., 2021), LLMs are more intuitive and flexible to apply, making them accessible to a broader range of users. Another key advantage of LLMs is their ability to process large volumes of text while maintaining consistency in application. Unlike human evaluators, who may introduce variability due to different levels in prior knowledge, subjective interpretation, and semantic understandings, LLMs provide uniformity in text classification, as demonstrated by the high intercoder agreements observed across trials. High intercoder agreements are particularly beneficial in tasks involving extensive datasets, complex criteria, and high replicability requirements. QA/QC screening, which demands reliable and consistent evaluations, is especially well-suited to leverage these strengths of LLMs. In addition, we clarify that the purpose of this study is not to compare the performance of different LLMs, but rather to demonstrate the versatility and general applicability of LLMs in performing specific text extraction and interpretation tasks.

Despite LLM's strengths in QA/QC screening, several limitations were also identified. Criteria like 'Sample Processing and Storage' and 'Sample Size' show lower consistency with human assessments, highlighting challenges in handling ambiguous tasks and complex reasoning or computation. In some cases, LLMs may also struggle to determine which criteria a contamination control procedure belongs to. To address these issues, refining instructions iteratively and simplifying tasks can reduce ambiguity and improve precision. Specifically, using a step-by-step prompt (e.g., first extract relevant text, then score based on the text and criterion instructions) was particularly effective in guiding the AI tools through tasks. Additionally, breaking down tasks to assess only one criterion per publication per request allows the AI tools to focus on a single aspect at a time, resulting in improved performance compared to providing one instruction of all criteria together. However, it is important to note that some of these QA/QC criteria are often challenging for human evaluators as well. Yet, as humans we sometimes demand more precision from AI than we expect from human evaluators, raising questions about the interpretability and expectations of AI performance (Le Mens et al., 2023). Additionally, for QA/QC screening tasks, using a ranking of studies may be more favorable than relying on absolute score values. For complex criteria or critical cases such as those with lower intercoder agreement or cases identified as unreliable (zero scores), additional human validation checks may also be beneficial.

The workflow established in this study also demonstrates strong potential for adaptation and broader applications. While it was specifically designed for QA/QC screening of microplastic data reliability, the methodologies in text preprocessing and prompt engineering can also be extended to assess data in other domains. Recent studies have also utilized LLMs for automated text mining to establish data sets from scientific publications (Dagdelen et al. 2024; Hu et al. 2024; Jablonka et al. 2024; Polak and Morgan 2024; Zheng et al. 2023). These approaches can be further extended to extract concentration data of microplastics from various environmental compartments and human matrices.

For future larger-scale or more complex tasks, several techniques could be further explored. Strategies such as fine-tuning models or chain-of-thought prompting may improve the LLM performance on zero-shot reasoning (Wu et al., 2024). Furthermore, incorporating few-shot learning (Kojima et al., 2022), which utilizes a small set of labeled examples, alongside with multimodal capabilities, such as recognizing and parsing figures and tables, could enable more comprehensive workflows and support holistic applications of LLMs in environmental research.

5. Conclusion

Here, we demonstrate the potential of AI tools, for instance ChatGPT and Gemini, to streamline data quality evaluation in microplastic research, a critical component in the context of microplastic risk assessment. AI-assisted assessments show promise in improving the speed, consistency, and applicability of QA/QC tasks. This suggests that AI can play a vital role in harmonizing microplastics risk assessments within regulatory frameworks, and address the challenges of an increasingly data-intensive research domain.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT for grammar checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Yanning Qiu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Svenja Mintenig:** Writing – review & editing, Validation, Formal analysis, Data curation. **Margherita Barchiesi:** Writing – review & editing, Validation, Formal analysis, Data curation. **Albert A. Koelmans:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgements

This work was supported by the Netherlands Organisation for Health Research and Development (Project number: 4580012320003). The authors thank Ingrid van de Leemput for valuable discussions.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2025.109341.

Data availability

Data will be made available on request.

References

Ali, S.R., Dobbs, T.D., Hutchings, H.A., Whitaker, I.S., 2023. Using ChatGPT to write patient clinic letters. Lancet Digital Health 5 (4), e179–e181. https://doi.org/10.1016/S2589-7500(23)00048-1.

Brady, W.J., McLoughlin, K., Doan, T.N., Crockett, M.J., 2021. How social learning amplifies moral outrage expression in online social networks. Sci. Adv. 7 (33), eabe5641.

Chen, H., & Ding, N. (2023). Probing the "Creativity" of Large Language Models: Can models produce divergent semantic association? *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12881–12888.

Cox, K.D., Covernton, G.A., Davies, H.L., Dower, J.F., Juanes, F., Dudas, S.E., 2019. Human consumption of microplastics. Environ. Sci. Tech. 53 (12). https://doi.org/10.1021/acs.est.9b01517.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A.S., Ceder, G., Persson, K.A., Jain, A., 2024. Structured information extraction from scientific text with large language models. Nat. Commun. 15 (1), 1418. https://doi.org/10.1038/s41467-024-45563-x.

De Ruijter, V.N., Redondo-Hasselerharm, P.E., Gouin, T., Koelmans, A.A., 2020. Quality criteria for microplastic effect studies in the context of risk assessment: a critical review. Environ. Sci. Tech. 54 (19). https://doi.org/10.1021/acs.est.0c03057.

Fenniak, M., Stamy, M., pubpub-zz, Thoma, M., Peveler, M., exiledkingcc, & PyPDF2 Contributors. (2022). *The PyPDF2 library*. https://pypi.org/project/PyPDF2/.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30). doi: 10.1073/pnas.2305016120.

Gouin, T., Ellis-Hutchings, R., Thornton Hampton, L.M., Lemieux, C.L., Wright, S.L., 2022. Screening and prioritization of nano- and microplastic particle toxicity studies for evaluating human health risks – development and application of a toxicity study assessment tool. Microplast. Nanoplast. 2 (1), 2. https://doi.org/10.1186/s43591-021-00023-x.

Hermsen, E., Mintenig, S. M., Besseling, E., & Koelmans, A. A. (2018). Quality Criteria for the Analysis of Microplastic in Biota Samples: A Critical Review. In *Environmental Science and Technology* (Vol. 52, Issue 18). doi: 10.1021/acs.est.8b01611.

Hu, M., Alkhairy, S., Lee, I., Pillich, R.T., Fong, D., Smith, K., Bachelder, R., Ideker, T., Pratt, D., 2024. Evaluation of large language models for discovery of gene set function. Nat. Methods. https://doi.org/10.1038/s41592-024-02525-x.

Jablonka, K.M., Schwaller, P., Ortega-Guerrero, A., Smit, B., 2024. Leveraging large language models for predictive chemistry. Nat. Mach. Intell. 6 (2), 161–169. https://doi.org/10.1038/s42256-023-00788-1.

Kase, R., Korkaric, M., Werner, I., Ågerstrand, M., 2016. Criteria for Reporting and Evaluating ecotoxicity Data (CRED): comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. Environ. Sci. Eur. 28 (1). https://doi.org/10.1186/s12302-016-0073-x.

Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharm. 25 (1). https://doi.org/10.1006/rtph.1996.1076.

Koelmans, A. A., Mohamed Nor, N. H., Hermsen, E., Kooi, M., Mintenig, S. M., & De France, J. (2019). Microplastics in freshwaters and drinking water: Critical review and assessment of data quality. In *Water Research* (Vol. 155). doi: 10.1016/j.watres.2019.02.054.

Koelmans, A. A., Redondo-Hasselerharm, P. E., Nor, N. H. M., de Ruijter, V. N., Mintenig, S. M., & Kooi, M. (2022). Risk assessment of microplastic particles. In *Nature Reviews Materials* (Vol. 7, Issue 2). doi: 10.1038/s41578-021-00411-y.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y., 2022. Large language models are zero-shot reasoners. Adv. Neural Inf. Proces. Syst. 35.

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences of the United States of America*, *120*. doi: 10.1073/pnas.2309350120.

Liang, W., Su, W., Zhong, L., Yang, Z., Li, T., Liang, Y., Ruan, T., Jiang, G., 2024. Comprehensive characterization of oxidative stress-modulating chemicals using GPT-based text mining. Environ. Sci. Tech. https://doi.org/10.1021/acs.est.4c07390.

Mohamed Nor, N.H., Kooi, M., Diepens, N.J., Koelmans, A.A., 2021. Lifetime accumulation of microplastic in children and adults. Environ. Sci. Tech. 55 (8), 5084–5096. https://doi.org/10.1021/acs.est.0c07384.

OpenAI. (2024, December 1). *Chat API reference: Create.* https://platform.openai.com/docs/api-reference/chat/create.

Pérez-Guevara, F., Roy, P.D., Elizalde-Martínez, I., Kutralam-Muniasamy, G., Shruti, V.C., 2022. Human exposure to microplastics from urban decentralized pay-to-fetch drinking-water refill kiosks. Sci. Total Environ. 848. https://doi.org/10.1016/j.scitotenv.2022.157722.

Polak, M.P., Morgan, D., 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat. Commun. 15 (1), 1569. https://doi.org/10.1038/s41467-024-45914-8.

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *121* (34). doi: 10.1073/pnas.2308950121.

Redondo-Hasselerharm, P.E., Rico, A., Koelmans, A.A., 2023. Risk assessment of microplastics in freshwater sediments guided by strict quality criteria and data alignment methods. J. Hazard. Mater. 441, 129814. https://doi.org/10.1016/j.jhazmat.2022.129814.

Roemmele, M., & Gordon, A. S. (2018). Automated Assistance for Creative Writing with an RNN Language Model. *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 1–2. doi: 10.1145/3180308.3180329.

Shruti, V.C., Kutralam-Muniasamy, G., Pérez-Guevara, F., Roy, P.D., Elizalde-Martínez, I., 2022. Free, but not microplastic-free, drinking water from outdoor refill kiosks: a challenge and a wake-up call for urban management. Environ. Pollut. 309. https://doi.org/10.1016/j.envpol.2022.119800.

Shruti, V.C., Pérez-Guevara, F., Kutralam-Muniasamy, G., 2020. Metro station free drinking water fountain- a potential "microplastics hotspot" for human consumption. Environ. Pollut. 261, 114227. https://doi.org/10.1016/j.envpol.2020.114227.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Natarajan, V., 2023. Large language models encode clinical knowledge. Nature 620 (7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2.

Thompson, R.C., Courtene-Jones, W., Boucher, J., Pahl, S., Raubenheimer, K., Koelmans, A.A., 2024. Twenty years of microplastic pollution research—what have we learned? Science 386 (6720). https://doi.org/10.1126/science.adl2746.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. https://arxiv.org/abs/2401.01313.

Vermaire, J.C., Pomeroy, C., Herczegh, S.M., Haggart, O., Murphy, M., 2017. Microplastic abundance and distribution in the open water and sediment of the Ottawa River, Canada, and its tributaries. Facets 2 (1), 301–314.

Wardani, I., Hazimah Mohamed Nor, N., Wright, S. L., Kooter, I. M., & Koelmans, A. A. (2024). Nano- and microplastic PBK modeling in the context of human exposure and risk assessment. *Environment International*, *186*. doi: 10.1016/j.envint.2024.108504.

WHO. (2019). Microplastics in drinking-water. *World Health Organization*, 101.

WHO. (2022). *Dietary and inhalation exposure to nano- and microplastic particles and potential implications for human health.*

Wright, S.L., Gouin, T., Koelmans, A.A., Scheuermann, L., 2021. Development of screening criteria for microplastic particles in air and atmospheric deposition: critical review and applicability towards assessing human exposure. Microplast. Nanoplast. 1 (1), 6. https://doi.org/10.1186/s43591-021-00006-y.

Wright, S.L., Kelly, F.J., 2017. Plastic and human health: a micro issue? Environ. Sci. Tech. 51 (12). https://doi.org/10.1021/acs.est.7b00423.

Wu, Y., Xu, M., Liu, S., 2024. Generative artificial intelligence: a new engine for advancing environmental science and engineering. Environ. Sci. Tech. 58 (40), 17524–17528. https://doi.org/10.1021/acs.est.4c07216.

Zheng, Z., Zhang, O., Borgs, C., Chayes, J.T., Yaghi, O.M., 2023. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. J. Am. Chem. Soc. 145 (32), 18048–18062. https://doi.org/10.1021/jacs.3c05819.

Zhu, J.-J., Jiang, J., Yang, M., Ren, Z.J., 2023. ChatGPT and environmental research. Environ. Sci. Tech. 57 (46), 17667–17670. https://doi.org/10.1021/acs.est.3c01818.

Zhu, J.-J., Yang, M., Jiang, J., Bai, Y., Chen, D., Ren, Z.J., 2024. Enabling GPTs for expert-level environmental engineering question answering. Environ. Sci. Technol. Lett. https://doi.org/10.1021/acs.estlett.4c00665.