Learned Publishing

LEARNED PUBLISHING Open Access

ORIGINAL ARTICLE OPEN ACCESS

Does ChatGPT Ignore Article Retractions and Other Reliability Concerns?

Mike Thelwall[1] | Marianna Lehtisaari[2] | Irini Katsirea[3] | Kim Holmberg[2] | Er-Te Zheng[1]

[1]Information School, University of Sheffield, Sheffield, UK | [2]Economic Sociology, University of Turku, Turku, Finland | [3]School of Journalism, Media and Communication, University of Sheffield, Sheffield, UK

Correspondence: Mike Thelwall (m.a.thelwall@sheffield.ac.uk)

Received: 11 October 2024 | **Revised:** 11 June 2025 | **Accepted:** 24 July 2025

Funding: This work was supported by the European Media Information Fund (316177).

Keywords: altmetrics | ChatGPT | large language models | retraction | science communication

ABSTRACT

Large language models (LLMs) like ChatGPT seem to be increasingly used for information seeking and analysis, including to support academic literature reviews. To test whether the results might sometimes include retracted research, we identified 217 retracted or otherwise concerning academic studies with high altmetric scores and asked ChatGPT 4o-mini to evaluate their quality 30 times each. Surprisingly, none of its 6510 reports mentioned that the articles were retracted or had relevant errors, and it gave 190 relatively high scores (world leading, internationally excellent, or close). The 27 articles with the lowest scores were mostly accused of being weak, although the topic (but not the article) was described as controversial in five cases (e.g., about hydroxychloroquine for COVID-19). In a follow-up investigation, 61 claims were extracted from retracted articles from the set, and ChatGPT 4o-mini was asked 10 times whether each was true. It gave a definitive yes or a positive response two-thirds of the time, including for at least one statement that had been shown to be false over a decade ago. The results therefore emphasise, from an academic knowledge perspective, the importance of verifying information from LLMs when using them for information seeking or analysis.

1 | Introduction

The knowledge in academic documents can be found directly, such as by reading journal articles, or indirectly, by reading literature reviews or textbooks on relevant topics. One new way in which the knowledge might appear is in response to a query to a large language model (LLM) like ChatGPT, Gemini, or DeepSeek, since these may even ingest some paywalled journal articles (Gibney 2024). LLMs are now integrated into some web search engines, as is the case with Microsoft Copilot, so a web user might search for information and be presented with many options, allowing them to choose between reading relevant articles or scanning the LLM synthesis. Given that LLMs are known to hallucinate (Giuffrè et al. 2024), the possibility that they may give misleading information or fake references about academic findings is a concern (Conroy 2023a; Dashti et al. 2023; De Cassai and Dost 2024), especially if there is a risk to public safety (e.g., Steen 2011). Although plugins such as ScholarAI enable LLMs to provide real references, an LLM may ingest problematic research together with any retraction/correction/warning notice and report it to users without the associated DOI, notice, or warning. It is therefore important to assess the extent to which popular LLMs recognise retracted or concerning academic work.

From a legal perspective, effective reporting of retracted information would mean that LLMs can live up to the expectation of reliability, which they seek to generate. The business model of the AI companies that create LLMs relies on the credibility and safety of the summaries that they produce (OpenAI 2024a;

Summary

- This article tested the 4o-mini version of ChatGPT to see whether it was aware that published academic articles had been retracted or were otherwise problematic.
- The results suggested that it was not aware of this information or at least did not take it into account when asked for relevant information about the articles or to check a claim made by them.
- Thus, users should be cautious of literature summaries made by large language models in case they contain false information.

Volokh 2023). If these outputs are inaccurate, they defy users' expectations and undermine the AI companies' business model. This would be the case if they portrayed unreliable science as excellent or world-leading, especially given that these companies seek to tailor their product to the needs of universities, and even invite them to "bring AI to campus at scale" (OpenAI 2024b). In so far as the scientific works evaluated by LLMs as reliable lead to personal injury based on inaccurate representation, this might expose them to liability. However, this would require causation to be proved, which may be difficult (Henderson et al. 2023). Should the LLMs present retracted, inaccurate scientific findings as sound, the AI companies could potentially also be held liable for defamation, provided fault could be established (Volokh 2023; Brown 2023).

Although ChatGPT can be helpful in academic writing, such as through language polishing (Rahimi and Talebi Bezmin Abadi 2023) and writing productivity enhancement (Noy and Zhang 2023), it may itself lead to retractions due to plagiarism (Rivera 2023; Kendall and Teixeira da Silva 2023). Some articles have already been retracted due to dishonest uses of ChatGPT, including generating unreliable figures and contents (Conroy 2023b; Tarla et al. 2023; Zhang et al. 2024). ChatGPT has also been used to help detect whether an article will be retracted by analysing its Twitter/X mentions (Zheng et al. 2024). It is also known that retractions tend to lead to a decline in new citations to the retracted article (Kühberger et al. 2022; Shuai et al. 2017), showing that the subsequent scholarly record self-corrects somewhat. However, the fact that citations to retracted research do not disappear and most post-retraction citations are positive suggests that further action is needed (Schneider et al. 2021; Yang et al. 2024). Policy citations may be reduced by retractions too, but not eliminated (Malkov et al. 2023). Nevertheless, no previous study has investigated whether retracted work can be found in LLM outputs.

From previous studies of retracted articles, the main causes are plagiarism, duplicate publication, ethical violations, misconduct/fraud, and mistakes (Ahmed et al. 2023; Audisio et al. 2022; Bolboacă et al. 2019; Campos-Varela and Ruano-Raviña 2019; Grieneisen and Zhang 2012). In contrast to fraud, mistakes are inevitable in science, and authors publicising them and retracting their work are to be commended (Bishop 2018; Fanelli 2016; Vuong 2020). Of these five reasons, the last two indicate that the findings lack evidence and may therefore be unreliable, but the first three instead point to an issue that does not invalidate the findings. In particular, the first two cases (plagiarism and duplication) suggest that something else might be cited instead, unless the plagiarism is partial, such as for a methods section or literature review. An ethical violation might leave a study's findings fully supported but still leave the work unable to enter the scholarly record. In this case, and in the case of partial plagiarism, it seems reasonable for an LLM to include the article findings or exclude them for greater ethical integrity, so the decision might be a policy call for LLM system designers, if they are able to control it.

This article includes two approaches to investigate retracted or potentially concerning information in ChatGPT. The first study takes an indirect approach. It investigates whether LLMs consider if an article has been retracted or has expressions of concern about it when carrying out a task for which that information would be highly relevant: evaluating its research quality. Scoring articles for quality is an indirect test, but it is relevant to retractions, especially if the retraction is not for duplication. This was chosen as the primary test instead of asking a question directly about the article findings, since it would be difficult to check that no other article had the same results. The focus here is on well-known articles, so overlooking their retraction status would be most concerning. Moreover, the retraction status of these articles should be the most widely publicised, and therefore easiest for an LLM to discover and absorb through repeated exposure. A follow-up study then directly checks whether key claims in retracted articles are reported as true by ChatGPT. Specifically, the research questions are as follows:

- RQ1: Does ChatGPT consider the retraction or otherwise problematic status of an article when evaluating its research quality?
- RQ2: Does ChatGPT report that claims in retracted articles are true?

2 | Methods

The overall research design for RQ1 was to obtain a set of high-profile retracted or potentially concerning articles (defined below as those with the most online mentions in mainstream media, social media and Wikipedia, as reflected by Altmetric.com scores), identify those with the greatest online attention, submit them to ChatGPT to score them for quality and then examine the highest and lowest scoring articles for evidence about whether ChatGPT knew about the retraction. High-profile articles give ChatGPT the best chance of knowing about retractions through repetition. Submitting text to the ChatGPT API conforms to UK copyright law because the text has been lawfully accessed (Hawkes 2012; Legislation.gov.uk 2014) and the results are not used to train a new model, thereby not potentially indirectly breaching copyright.

2.1 | Retracted and Potentially Concerning Article Data Set

We first identified a large sample of articles with retractions or other concerns in August 2024. We then checked them for mass

media and social media attention. We used RetractionWatch as a well-known and subjectively authoritative source of articles that have been retracted or that have substantial concerns about them. RetractionWatch has been mentioned in at least 210 Scopus-indexed articles by June 7, 2025, according to an advanced query TITLE-ABS-KEY(retractionwatch) OR TITLE-ABS-KEY("retraction watch"), giving evidence of its academic reputation. For example, its data was used for the articles, "Retractions in Otolaryngology Publications" and "Moving Open Repositories out of the Blind Spot of Initiatives to Correct the Scholarly Record." For a more complete set of potentially problematic articles, we attempted to augment this list with retractions from the two major bibliometric databases, Scopus and the Web of Science, although the latter was not technically possible due to missing DOIs in batch downloads. The data set therefore consists of articles flagged in Scopus as retracted or recorded in the RetractionWatch database, indicating a potential concern or retraction.

CrossRef maintains a list of retracted and concerning articles from the RetractionWatch Database (https://www.crossref.org/blog/news-crossref-and-retraction-watch/). This is accompanied by information from the RetractionWatch team about the nature of the retraction as well as metadata (including DOI, title, journal). This was downloaded as a file from CrossRef. Not all articles in this database have been retracted, but it is an authoritative source of articles that this well-known site has identified as potentially concerning. Some articles are also marked as retracted in Scopus, in terms of matching the advanced query DOCTYPE(tb), and the Scopus API was used to download a complete set, with metadata. This metadata was for the article rather than the retraction notice. Scopus does not record reasons for retractions, although these are sometimes briefly stated in the article abstract.

The two data sources were combined into a single file and cleaned. Titles were cleaned by removing various forms of retraction statement from article title starts (e.g., "Retraction of '…'") and article title ends (e.g., publication metadata). Characters that were systematically incorrect were also corrected, when noticed (e.g., various forms of smart quotes had been changed to non-alphabetical character sequences in some of the downloaded files).

The records in the cleaned combined database were then checked for duplicates. These could occur because the different databases contained the same article, or one database contained multiple copies of the same article. Two articles were classed as duplicates if they had the same DOI or the same title. For this data set, unusually, two articles with different DOIs could be the same if the journal published the retraction as a separate document with its own DOI rather than by replacing the article with a retraction notice, keeping its DOI. Two articles with different titles could be the same because the databases had dealt differently with non-alphabetical characters (e.g., changing β into i^2), subscripts and superscripts (marking them with tags_{and}or not), and hyphens (various lengths, with and without spaces on either side). Thus, checking that two articles were the same was imperfect, and duplicates remained in the final data set. The final sample contained 56,705 unique DOIs.

2.2 | Altmetric Scores for Articles and Final Sample Selection

For the sample, "high profile" retractions and concerning articles were operationalised as those that had been mentioned extensively online. This is a pragmatic operationalisation given that it is impractical to search for all retractions in mainstream media sources, but it is possible to search for online mentions from Altmetric.com's coverage. To identify articles with substantial media coverage from this set, the DOIs were submitted to Altmetric.com's Applications Programming Interface (API) to record their altmetrics. These included the number of mentions in Altmetric.com's curated list of mainstream media sources (scraped from web pages), and mentions in Wikipedia, Twitter/X, Facebook, Reddit, and Pinterest. The results were used to sort the 56,705 articles in descending order of mentions in mainstream media and Wikipedia, then in descending order of total mentions in social media sites, then in descending order of altmetric score. Wikipedia was included in addition to the social media sources for the pragmatic reason that Altmetric.com tracks it, and it seems to be a valid source of attention evidence. The sorting order reflected a belief that mainstream media mentions and Wikipedia mentions were more important than social media mentions, with the Altmetric.com score being a hybrid indicator and thus the least reliable.

After the above sorting, the top 250 articles were selected for submission to ChatGPT. Many of the abstracts of these articles were missing, so we manually searched for them, excluding articles that no longer had an abstract. At the same time, we searched for the article page and for retraction notices to check the status of each article (as of August 2024). While 144 articles had retraction notices and 9 had been withdrawn, the remainder were not clear retractions. Of these, 12 had been updated or corrected, 26 had been retracted and replaced/republished, 16 had an expression of concern, and 11 were not retracted and had no official publisher expressions of concern about them but had concerns elsewhere (e.g., PubPeer.com), and one had been retracted but reinstated. The final set contained 219 articles with abstracts, but when analysing the ChatGPT results, we discovered that the abstracts of two of the articles were retraction notices, so these were excluded to give a final sample size of 217.

2.3 | RQ1: Assessing ChatGPT's Quality Scores for Concerning Articles

The title and abstract of each article were combined and submitted to ChatGPT 4o-mini via its API in September 2024 with a request to evaluate its research quality. The version gpt-4o-mini-2024-07-18 was used. It has a training date cutoff of October 1, 2023, and may therefore be unaware of retractions issued after this date. This issue was not considered in the research design because, from the user's perspective, whether the article has been retracted at the point of submitting the query is the most relevant consideration.

Full text is not necessary for a quality evaluation because quality scores from ChatGPT correlate more highly with expert scores if only the title and abstract are entered (Thelwall 2024, 2025). Thus, only the title and abstract were used. Following a previous

study (Thelwall and Yaghi 2024), each request was accompanied by a system prompt defining research quality using the Research Excellence Framework (REF) 2021 guidelines for all broad areas of scholarship (REF 2019). Although there are many different definitions of research quality (Langfeldt et al. 2020), the REF guidelines have the advantage of being intended to cover all academic fields and to align with four quality levels for research, giving a clear single score (see the appendix of Thelwall and Yaghi 2024). Each article was assigned manually to the most relevant of the four REF broad areas (A: health and life sciences, B: physical sciences and engineering, C; social sciences, D: arts and humanities) for the appropriate REF system prompt. This broad area assignment was based on the Scopus fields to which the article had been assigned, the name of the publishing journal, and the title and abstract of the article. For example, "A primitive Late Pliocene cheetah, and evolution of the cheetah lineage" was classified as health and life sciences based on its title and abstract because its Scopus fields and publishing journal (PNAS) were both general. To get the most accurate scores, each article was submitted 30 times to ChatGPT, and the arithmetic mean was calculated from the individual scores. Previous studies have established the need to average scores from multiple identical queries, with 5, 15, and 30 iterations being variously used (Thelwall 2024, 2025; Thelwall and Yaghi 2024). Since precision increases with the number of iterations and there were not too many articles for the cost of 30 iterations to be prohibitive, 30 was chosen.

The primary analysis was to check how many articles' quality scores may have been influenced directly or indirectly by their retraction or concerns about them. This was checked in two ways. First, all $217 \times 30 = 6510$ reports were searched for the strings "ethical," "violation," "duplicate," "problematic," "retract," "correct," and "error" and the matches were read to identify any mentions of retraction, corrections, or errors. Second, reports on the articles with the lowest scores were read to identify the reasons for the scores, looking especially for anything related to the issue causing the retraction/correction. The number of articles to check was established from a graph of the score distribution (see Section 3).

The article scores were also correlated against their altmetric rank order as a very approximate test for more popular retractions being more likely to be recognised by ChatGPT. This is very approximate because more well-known articles might also have the most important research and tend to be better because of this.

2.4 | RQ2: Assessing Whether ChatGPT Reports Retracted Article Claims

A claim was extracted from each retracted article when this seemed possible. This was only done for articles that were retracted (with a publisher website retraction notice) rather than corrected, replaced, or flagged with an expression of concern. A claim was operationalised as a simple statement that was a primary outcome of the retracted study, and that it would be reasonable to believe based on that study. Retractions that were exclusively for duplicate publication or ethical violations were ignored (e.g., "The Publisher and the Editor-in-Chief have retracted this article due to noncompliance with our editorial policies around consent."; "The Editor-in-Chief has retracted this article because it significantly overlaps with a previously published article") as were articles where there was not a clear claim that had been supported by the retracted article (e.g., "With our results, we were not able to explain if COVID-19 mortality is reduced by staying at home in ~98% of the comparisons after epidemiological Weeks 9 to 34.") or where the evidence was inconclusive (e.g., an article with a title including "A preliminary report," and another with the imprecise claim "...which was generally associated in time with possible environmental triggers."), or where the retraction covered complex claims, such as a mediating relationship (e.g., "None of these response indicators show a clear positive trend of extreme events").

These claims were manually extracted by the first author by reading the article abstract and cross-checking with the retraction notice, resulting in 61 statements. The simplest and shortest was, "Green coffee extract reduces obesity," and the longest and most complex was, "Epinephrine (EPI) preserves autoregulation, prevents histopathology, and blocks phosphorylated c-Jun-terminal kinase (JNK) upregulation in newborn males and females and juvenile females but not juvenile males after Traumatic brain injury (TBI)." Other examples include, "No Mow May lawns have higher pollinator richness and abundances" and "Pigeons show faster responses than humans when subtasks are separated such that fast switches between processes are required". A prompt was then submitted to ChatGPT (10 times per statement) to ask whether the claim was true. This was also achieved using the ChatGPT API rather than its web interface to ensure that there was no possibility of the system learning false information from the prompts. The system component was, "You are an expert at checking the accuracy of statements about academic knowledge." and the user prompt was, "Is the following statement true? [statement]." The choice of 10 ChatGPT submissions per statement was made heuristically, using multiple submissions to improve the chance of finding rarer types of output from ChatGPT without generating an unmanageable amount of data for the subsequent manual analysis.

The ChatGPT answers were manually evaluated to assess whether ChatGPT reported that they were true or false. For this, the first author read all 610 ChatGPT responses and devised a set of 9 categories to capture how ChatGPT described the truthfulness of the statements. Eight of the categories are listed in Table 1, with the ninth category having been dropped after the process below. A second independent coder (not one of the research team) was given the scheme and independently "cold" (i.e., without training) coded the same responses with the same schema. The instructions were as follows.

1. Apply the first category that applies in the first sentence. (The first ChatGPT sentence often gives a clear answer.)
2. If no category applies in the first sentence, apply the first category that applies in the last sentence, if it is a summary. (The last ChatGPT sentence sometimes summarises its response.)
3. Apply the closest fitting category for the entire ChatGPT response.

A Cohen's Kappa agreement calculation for the two sets of results gave Kappa = 0.598, which is sometimes described as "moderate" agreement (just below "substantial" at 0.61). This seems adequate

TABLE 1 | The main types of response from ChatGPT to questions about the truth of a statement from a retracted article ($n = 61$), with 10 identical queries per question.

ChatGPT's response about the truth of the submitted statement	*N*	Percent
The statement is/is generally/can be considered to be/is likely to be/appears to be [generally/largely] true/accurate [based on research]	267	43.8%
The statement is generally consistent with/aligns with/appears to align with/is [generally] supported by/is consistent with the findings of research. OR The statement is supported by some studies.	58	9.5%
The statement is/can be considered somewhat/partially true/accurate. OR The statement is true in some/many contexts. OR The statement is not universally true. OR The statement is partially supported by research.	79	13.0%
The statement is controversial or subject to debate.	29	4.8%
The statement can't be confirmed as true/is currently uncertain/is a plausible hypothesis. OR More/Recent evidence/studies/research is needed. OR Evidence is not definitive.	38	6.2%
The statement is [generally] not [considered] [conclusively] [to be] true/accurate/established. OR The statement is not universally/widely/definitively true/accurate/established.	89	14.6%
The statement is not [definitively/generally] supported by [current] research. OR There is limited/no/lack of evidence that the statement is true. OR the statement is not supported by/does not align with/is unsupported by research.	43	7.0%
[There is a consensus that] the statement is false.	7	1.1%

Note: In the descriptions, slashes indicate alternatives, and square brackets indicate optional components.

to claim that the results are reasonably reliable. A confusion matrix between coders revealed that the second coder had never used one of the categories ("The statement is ambiguous, or its truth depends on context"), so this was removed and the answers merged into another similar category ("The statement is/can be considered somewhat/partially true/accurate. OR The statement is true in some/many contexts. OR The statement is not universally true. OR The statement is partially supported by research."). The first author revisited all cases of disagreement and selected a final code.

3 | Results

3.1 | RQ1: Does ChatGPT Consider the Retraction or Problematic Status of an Article When Evaluating Its Research Quality?

None of the 217×30 ChatGPT reports on the quality of the 217 articles mentioned retractions, corrections, or ethical problems. Although a few mentioned errors (e.g., "the article could benefit from enhanced transparency regarding data collection methods, particularly concerning dietary compliance and measurement error related to waist circumference"), none were relevant to the associated retraction or correction statement. Thus, ChatGPT seems to never directly mention retraction or correction information in response to a request to evaluate an article's quality. Possible explanations are that it does not know about the retractions/corrections (which seems unlikely since they are on the public web) or that it is not able or willing to reliably associate the retraction/correction information with an article's title and abstract. This latter point essentially suggests that ChatGPT might not fully "comprehend" the meaning of a retraction notice that has been attached to an article (usually in the title), indicating to a human reader that the article has been flagged as flawed and it should not be trusted or used.

ChatGPT might take retractions or expressions of concern into account through its language associations even if it does not directly mention them. If this happened, then lower REF scores could be expected, and many 1* scores. Nevertheless, most articles (73%) scored high average grades: between 3* (internationally excellent) and 4* (world leading) (Figure 1), so any retraction notice or expression of concern has had little effect on these (for field averages, see: Thelwall and Yaghi 2024). Nevertheless, the break in the shape of the graph at a REF score of 2.5* (Figure 1) suggests that 27 articles scoring below 2.5 are anomalous and might have been downgraded due to validity concerns. This was given accidental but partial corroboration by the two articles submitted by mistake with retraction notices instead of abstracts. These had average scores of 1.01* and 1.13*.

From reading the 30 ChatGPT reports on each article in the lowest scoring set of 27 articles (not named here to avoid further publicising potentially problematic work), the reason for the low score given in each report was a plausible explanation that did not relate to the retraction cause, except indirectly for five articles that were identified as relating to a controversial topic. The reasons included the following, with the number of articles in brackets (multiple counting). None of the reasons rule out the possibility that knowledge of retraction encouraged ChatGPT to score more negatively, however.

- [13] The study is weak overall (e.g., "the study does not significantly develop new methodologies or theoretical insights beyond the existing literature," "remains limited in scope and actionable insights," "The article could benefit from a larger sample size, richer theoretical engagement, and greater methodological sophistication"). The retraction or expression of concern may have been caused by this weakness, perhaps in combination with overclaiming.

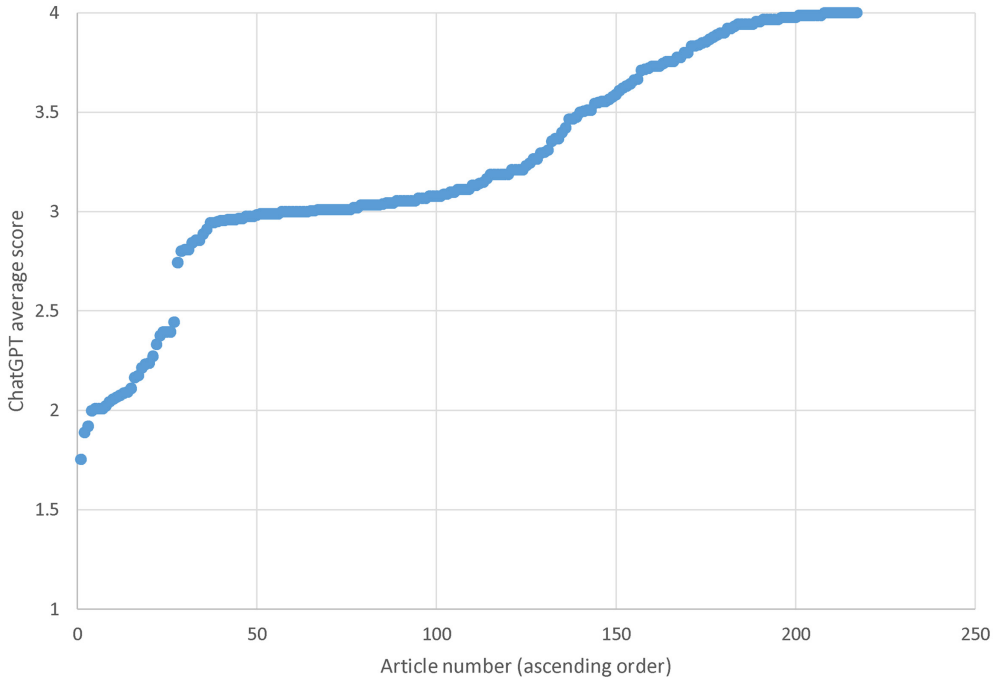

FIGURE 1 | Average ChatGPT REF score for the 217 high-profile retracted or concerning articles. Articles are listed in ascending order of ChatGPT score.

- [7] The article is criticised for a lack of novelty or relevant framing, not considering its early publication date (e.g., "the approach remains within a well-trodden area of study," "the approach appears to be similar to multiple other studies emerging during the pandemic, leading to limited originality").
- [5] The study is part of a controversy (e.g., "...contribute to a contentious public health debate regarding...," "controversial concept of...," "a drug that has witnessed both interest and controversy regarding its effectiveness," "Furthermore, it engages with themes that have been historically controversial without significantly advancing theoretical advancements or offering new insights into gender identity," and perhaps also, "the concept of [idea associated with conspiracy theories] lacks extensive empirical exploration in existing literature"). ChatGPT might be cautious with scores for sensitive or controversial topics.
- [5] The article was a review, and hence not primary research, and was criticised for a lack of contribution (e.g., "it ultimately falls short of making a distinctive contribution," "it ultimately falls short of the upper tiers of excellence due to its reliance on previous studies for originality and a lack of in-depth application and policy implications.").

Finally, the average ChatGPT research quality scores for each of the 217 articles in the retracted and potentially concerning articles set had a very weak correlation with their altmetric rank (Spearman's $\rho = -0.062$; bootstrapped 95% confidence interval: −0.199 to 0.090), so there is not a strong relationship between the awareness of a problematic or concerning article and its score.

3.2 | RQ2: Does ChatGPT Report That Claims in High-Profile Retracted Articles Are True?

For the 61 claims extracted from high-profile retracted articles, almost two-thirds of the time, ChatGPT reported that they were likely to be true (43.8%), partially true (9.5%), or consistent with research (9.5%). It was rarely reported that a statement was false (1.1%) or even that it was unsupported by current research (7.0%) or not established (14.6%). For the remaining minority of cases, it reported uncertainty. Thus, overall, ChatGPT has a positive bias towards statements from retracted articles (Table 1). The remainder of this section discusses the results but mentions the statements mostly indirectly or cautiously to avoid polluting the scientific record with new copies of potentially false statements.

ChatGPT always (i.e., 10 times out of 10) reported that a statement was true in 12 out of the 61 cases. In theory, any retracted claim could be true, unknown, or false. By default, a retracted claim would have an unknown status unless other research had shown it to be true or false. It is difficult to comprehensively check most of the claims because they are often complex or relatively general and require subject knowledge to interpret. For example, one requiring substantial subject knowledge to assess was about memory deficits in Tg2576 mice and extracellular accumulation of a 56-kDa soluble amyloid-β assembly, Aβ*56. Nevertheless, one of the claims was demonstrably false, "Acinonyx kurteni is the earliest known species of cheetah" since the sole fossil that this name is derived from was accepted as being a forgery, and the associated article was retracted in 2012. Thus, ChatGPT can report certainty about claims that are false, albeit an arguably minor claim to most people in this case.

In most cases (38 out of 61), ChatGPT reported that the statement was true at least once out of 10 iterations. In five of these cases, ChatGPT also suggested at least once that the statement was not proven or not true (i.e., one of the two penultimate categories in Table 1). Thus, the randomness within the ChatGPT algorithm can affect whether a user leaves with the impression that an academic-related claim is true or false.

Although ChatGPT never reported that a statement was false every time that it was asked (i.e., it always reported that the statement was false less than 10 times out of 10), in five cases it always (i.e., 10 times out of 10) suggested that a statement was untrue or might not be true (i.e., the bottom three categories in Table 1). Three of these statements were about COVID-19 (and Ivermectin/facemasks/vaccination), suggesting that ChatGPT had either ingested much text about these topics or had been given or learned safeguards for them as major public health issues. The other two were also health related. One was a public health issue about hearing aids and dementia, and the other was a more theoretical claim about human white blood cell surface proteins. For context, COVID-19 was mentioned in 6 of the 61 statements, although in two cases the claim was not directly relevant to public health issues.

4 | Discussion and Conclusions

This study is limited to a single LLM (ChatGPT 4o-mini), a single set of high-profile articles, a single period (September 2024), and the UK REF definition of research quality. It is possible that other LLMs will work differently, and it seems likely that LLM processing and reporting of scientific information will evolve over time. In particular, newer versions of LLMs can integrate live web information into their responses, from June 2025, combining "internal + web sources for synthesis" (OpenAI 2025). Future research may show an evolution to a stage where retraction processing is performed well by some LLMs. Finally, previous research has shown that prompt strategies can be designed to improve the consistency of information extracted from LLMs (Gundabathula and Kolar 2024), so different queries might have revealed ChatGPT's knowledge of retractions. The main source of retracted or potentially problematic articles, RetractionWatch, is also a limitation since it may contain unknown biases.

Since the ChatGPT 4o-mini version used had knowledge ending on October 1, 2023, retractions and concerns expressed after this date might be unknown to it. From the 217 articles analysed, only 14 had been retracted or corrected after this date, or had their first critical comment made after this date (if not retracted or corrected). Of these, six had expressions of concern, commenting articles or critical PubPeer comment published before October 2023 (e.g., "Change history January 26, 2023 Editor's Note: Readers are alerted that the conclusions of this paper are subject to criticisms that are being considered by editors"), so only eight (4%) would have had their first serious concerns raised after the ChatGPT knowledge cutoff date. For the 61 claims, only three had their retractions after the ChatGPT knowledge cutoff date, and one of these had an online statement of concern before this date. Thus, the primary reason for the results reported is not that problems were raised after the ChatGPT knowledge cutoff date.

ChatGPT's apparent unawareness of retractions and concerns when evaluating academic research is a new finding. Although the issue of ChatGPT's reporting of retracted academic research has not been investigated before, the results about the claims extracted from retracted articles align with previous evidence that it can be unreliable. For example, it can hallucinate when writing academic text (Alkaissi and McFarlane 2023; Giuffrè et al. 2024) and does not always answer academic questions correctly (Maitland et al. 2024; Ngo et al. 2024). In this context, the current study extends previous research by showing that it can sometimes report that retracted claims are true, even a long time after the retraction. Encouragingly, however, it was more cautious with high-profile health issues, not reporting the associated statement to be true. This caution aligns with OpenAI's claims that it has worked to reduce the chance that ChatGPT produces harmful content (openai.com/index/openai-safety-update/). It is difficult to verify OpenAI's procedures for retractions from the source since it does not mention retractions in its help documentation (https://help.openai.com/en/ as of June 7, 2025), but asking ChatGPT "How does ChatGPT deal with retractions" gives the following response:

As of now, ChatGPT does not automatically detect or flag retracted papers unless:

The retraction is widely known and occurred before its training cutoff (June 2024 for GPT-4-turbo).

You explicitly provide the article or mention that it might be retracted.

Asking the same question from the OpenAI helpbot on the site from a logged-in OpenAI account gives the following answer (https://help.openai.com/en/ on June 7, 2025), which is more cautious and more consistent with findings reported above.

> I'm an AI support agent for OpenAI products, here to help you with your questions about using ChatGPT! When it comes to academic article retractions, ChatGPT itself does not automatically detect or handle retracted articles in its responses. ChatGPT generates answers based on patterns and data it has been trained on, which includes information up to its last knowledge update (currently June 2024). If you specifically ask about the status of an article, or mention that it might be retracted, ChatGPT can look up this information if you provide sources or context. However, it won't independently alert you or filter out retracted research. If you're using ChatGPT to get information about academic papers, it's best to check the official status of the articles yourself—especially if making decisions or referencing critical work. Many publishers and platforms provide up-to-date alerts about retractions. If you have a specific article in mind or want to know how to use ChatGPT for academic research verification, let me know how you would like to proceed!

The results overall suggest that ChatGPT does not have awareness of retractions or other indicators of problematic content for high-profile academic research, although it is sensitive to some particularly important health issues associated with retractions. This lack of awareness seems to be an unfortunate limitation that should be addressed in the future. While people visiting retracted article pages would now typically see clear retraction notices, if they rely instead on LLMs for knowledge or knowledge summaries, then they can be misled. Of course, this is only one issue of false information in LLM output, but as retracted articles seem now to be clearly flagged, it seems unfortunate that this issue does not seem to have been addressed yet by ChatGPT (June 2025).

Author Contributions

Mike Thelwall: writing – original draft (lead), formal analysis (lead). **Marianna Lehtisaari:** formal analysis (lead), writing – review and editing (equal). **Irini Katsirea:** writing – original draft (supporting). **Kim Holmberg:** writing – review and editing (equal). **Er-Te Zheng:** writing – review and editing (equal).

Acknowledgements

This study was funded by the Fundação Calouste Gulbenkian European Media and Information Fund as part of project 316177: Unreliable science: Unravelling the impact of mainstream media misrepresentation.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Ahmed, F., Y. B. A. Güzelderen, Ş. Yurdakul, P. P. Kahrız, and K. M. Khawar. 2023. "Postmortem of Retracted Scientific Publications From Türkiye." *Medical Research Reports* 6, no. 3: 137–152.

Alkaissi, H., and S. I. McFarlane. 2023. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing." *Cureus* 15, no. 2: e35179. https://doi.org/10.7759/cureus.35179.

Audisio, K., N. B. Robinson, G. J. Soletti, et al. 2022. "A Survey of Retractions in the Cardiovascular Literature." *International Journal of Cardiology* 349: 109–114.

Bishop, D. V. 2018. "Fallibility in Science: Responding to Errors in the Work of Oneself and Others." *Advances in Methods and Practices in Psychological Science* 1, no. 3: 432–438.

Bolboacă, S. D., D. V. Buhai, M. Aluaș, and A. E. Bulboacă. 2019. "Post Retraction Citations Among Manuscripts Reporting a Radiology-Imaging Diagnostic Method." *PLoS One* 14, no. 6: e0217918.

Brown, N. 2023. "Bots Behaving Badly: Products Liability Approach to Chatbot-Generated Defamation." *Journal of Free Speech Law* 3, no. 2: 389–424.

Campos-Varela, I., and A. Ruano-Raviña. 2019. "Misconduct as the Main Cause for Retraction. A Descriptive Study of Retracted Publications and Their Authors." *Gaceta Sanitaria* 33: 356–360.

Conroy, G. 2023a. "How ChatGPT and Other AI Tools Could Disrupt Scientific Publishing." *Nature* 622, no. 7982: 234–236.

Conroy, G. 2023b. "Scientific Sleuths Spot Dishonest ChatGPT Use in Papers." *Nature*. https://doi.org/10.1038/d41586-023-02477-w.

Dashti, M., J. Londono, S. Ghasemi, and N. Moghaddasi. 2023. "How Much Can We Rely on Artificial Intelligence Chatbots Such as the ChatGPT Software Program to Assist With Scientific Writing?" *Journal of Prosthetic Dentistry* 133: 1082–1088. https://doi.org/10.1016/j.prosdent.2023.05.023.

De Cassai, A., and B. Dost. 2024. "Concerns Regarding the Uncritical Use of ChatGPT: A Critical Analysis of AI-Generated References in the Context of Regional Anesthesia." *Regional Anesthesia and Pain Medicine* 49, no. 5: 378–380.

Fanelli, D. 2016. "Set Up a 'Self-Retraction' System for Honest Errors." *Nature* 531, no. 7595: 415.

Gibney, N. 2024. "Has Your Paper Been Used to Train an AI Model? Almost Certainly." https://www.nature.com/articles/d41586-024-02599-9.

Giuffrè, M., K. You, and D. L. Shung. 2024. "Evaluating ChatGPT in Medical Contexts: The Imperative to Guard Against Hallucinations and Partial Accuracies." *Clinical Gastroenterology and Hepatology* 22, no. 5: 1145–1146.

Grieneisen, M. L., and M. Zhang. 2012. "A Comprehensive Survey of Retracted Articles From the Scholarly Literature." *PLoS One* 7, no. 10: e44118.

Gundabathula, S. K., and S. R. Kolar. 2024. "PromptMind Team at MEDIQA-CORR 2024: Improving Clinical Text Correction With Error Categorization and LLM Ensembles." https://doi.org/10.48550/arxiv.2405.08373.

Hawkes, N. 2012. "Government's Proposed Changes to Copyright Law Will Help Research, Wellcome Says." *BMJ (Clinical Research Ed.)* 345: e8690. https://doi.org/10.1136/bmj.e8690.

Henderson, P., T. Hashimoto, and M. Lemley. 2023. "Where's the Liability in Harmful AI Speech?" *Journal of Free Speech Law* 3, no. 2: 589–650.

Kendall, J., and J. A. Teixeira da Silva. 2023. "Risks of Abuse of Large Language Models, Like ChatGPT in Scientific Publishing: Authorship, Predatory Publishing, and Paper Mills." *Opinion* 37: 55–62.

Kühberger, A., D. Streit, and T. Scherndl. 2022. "Self-Correction in Science: The Effect of Retraction on the Frequency of Citations." *PLoS One* 17, no. 12: e0277814.

Langfeldt, L., M. Nedeva, S. Sörlin, and D. A. Thomas. 2020. "Co-Existing Notions of Research Quality: A Framework to Study Context-Specific Understandings of Good Research." *Minerva* 58, no. 1: 115–137.

Legislation.gov.uk. 2014. "Copyright, Designs and Patents Act 1988." https://www.legislation.gov.uk/ukpga/1988/48/section/29A.

Maitland, A., R. Fowkes, and S. Maitland. 2024. "Can ChatGPT Pass the MRCP (UK) Written Examinations? Analysis of Performance and Errors Using a Clinical Decision-Reasoning Framework." *BMJ Open* 14, no. 3: e080558.

Malkov, D., O. Yaqub, and J. Siepel. 2023. "The Spread of Retracted Research Into Policy Literature." *Quantitative Science Studies* 4, no. 1: 68–90.

Ngo, A., S. Gupta, O. Perrine, R. Reddy, S. Ershadi, and D. Remick. 2024. "ChatGPT 3.5 Fails to Write Appropriate Multiple Choice Practice Exam Questions." *Academic Pathology* 11, no. 1: 100099.

Noy, S., and W. Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Science* 381, no. 6654: 187–192.

OpenAI. 2024a. "GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses." https://openai.com/index/gpt-4/.

OpenAI. 2024b. "Bring AI to Campus at Scale." https://openai.com/chatgpt/education/.

OpenAI. 2025. "ChatGPT—Release Notes." https://help.openai.com/en/articles/6825453-chatgpt-release-notes.

Rahimi, F., and A. Talebi Bezmin Abadi. 2023. "Passive Contribution of ChatGPT to Scientific Papers." *Annals of Biomedical Engineering* 51, no. 11: 2340–2350.

REF. 2019. "Panel Criteria and Working Methods (2019/02)." https://2021.ref.ac.uk/publications-and-reports/panel-criteria-and-working-methods-201902/index.html.

Rivera, H. 2023. "Scientific Integrity Faces Plagiarism Fabricated With the ChatGPT." *Revista Médica del Instituto Mexicano del Seguro Social* 61, no. 6: 857–862.

Schneider, J., N. D. Woods, R. Proescholdt, Y. Fu, and RISRS Team. 2021. "Reducing the Inadvertent Spread of Retracted Science: Shaping a Research and Implementation Agenda." *F1000Research* 10: 329.

Shuai, X., J. Rollins, I. Moulinier, T. Custis, M. Edmunds, and F. Schilder. 2017. "A Multidimensional Investigation of the Effects of Publication Retraction on Scholarly Impact." *Journal of the Association for Information Science and Technology* 68, no. 9: 2225–2236.

Steen, R. G. 2011. "Retractions in the Medical Literature: How Many Patients Are Put at Risk by Flawed Research?" *Journal of Medical Ethics* 37, no. 11: 688–692.

Tarla, S., K. K. Ali, and A. Yusuf. 2023. "Exploring New Optical Solutions for Nonlinear Hamiltonian Amplitude Equation via Two Integration Schemes." *Physica Scripta* 98, no. 9: 095218.

Thelwall, M. 2024. "Can ChatGPT Evaluate Research Quality?" *Journal of Data and Information Science* 9, no. 2: 1–21. https://doi.org/10.2478/jdis-2024-0013.

Thelwall, M. 2025. "Evaluating Research Quality With Large Language Models: An Analysis of ChatGPT's Effectiveness With Different Settings and Inputs." *Journal of Data and Information Science* 10, no. 1: 7–25. https://doi.org/10.2478/jdis-2025-0011.

Thelwall, M., and A. Yaghi. 2024. "In Which Fields Can ChatGPT Detect Journal Article Quality? An Evaluation of REF2021 Results." https://arxiv.org/abs/2409.16695.

Volokh, E. 2023. "Large Libel Models? Liability for AI Output." *Journal of Free Speech Law* 3: 490–558.

Vuong, Q. H. 2020. "The Limitations of Retraction Notices and the Heroic Acts of Authors Who Correct the Scholarly Record: An Analysis of Retractions of Papers Published From 1975 to 2019." *Learned Publishing* 33, no. 2: 119–130.

Yang, S., F. Qi, H. Diao, and I. Ajiferuke. 2024. "Do Retraction Practices Work Effectively? Evidence From Citations of Psychological Retracted Articles." *Journal of Information Science* 50, no. 2: 531–545.

Zhang, M., L. Wu, T. Yang, B. Zhu, and Y. Liu. 2024. "The Three-Dimensional Porous Mesh Structure of Cu-Based Metal-Organic-Framework-Aramid Cellulose Separator Enhances the Electrochemical Performance of Lithium Metal Anode Batteries." *Surfaces and Interfaces* 46: 104081.

Zheng, E., H. Z. Fu, and Z. Fang. 2024. "Can ChatGPT Predict Article Retraction Based on Twitter Mentions?" ArXiv Preprint arXiv:2403.16851.