
Article:
Zheng, E. orcid.org/0000-0001-8759-3643, Fu, H. orcid.org/0000-0002-1534-9374, Thelwall, M. orcid.org/0000-0001-6065-205X et al. (1 more author) (2025) Can social media provide early warning of retraction? Evidence from critical tweets identified by human annotation and large language models. Journal of the Association for Information Science and Technology. ISSN: 2330-1635

https://doi.org/10.1002/asi.70028

Received: 5 December 2024 | Revised: 1 August 2025 | Accepted: 12 September 2025

DOI: 10.1002/asi.70028

RESEARCH ARTICLE

JASIST WILEY

Can social media provide early warning of retraction? Evidence from critical tweets identified by human annotation and large language models

Er-Te Zheng[1] | Hui-Zhen Fu[2] | Mike Thelwall[1] | Zhichao Fang[3,4]

[1]School of Information, Journalism and Communication, The University of Sheffield, Sheffield, UK

[2]Department of Information Resources Management, Zhejiang University, Hangzhou, China

[3]School of Information Resource Management, Renmin University of China, Beijing, China

[4]Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

Correspondence
Zhichao Fang, School of Information Resource Management, Renmin University of China, Beijing, China; Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands.
Email: z.fang@cwts.leidenuniv.nl

Funding information
National Natural Science Foundation of China, Grant/Award Number: 72304274; National Social Science Fund of China, Grant/Award Number: 22CTQ032; Fundação Calouste Gulbenkian European Media and Information Fund, Grant/Award Number: 316177; School of Information, Journalism and Communication of the University of Sheffield

Abstract

Timely detection of problematic research is essential for safeguarding scientific integrity. To explore whether social media commentary can serve as an early indicator of potentially problematic articles, this study analyzed 3815 tweets referencing 604 retracted articles and 3373 tweets referencing 668 comparable non-retracted articles. Tweets critical of the articles were identified through both human annotation and large language models (LLMs). Human annotation revealed that 8.3% of retracted articles were associated with at least one critical tweet prior to retraction, compared to only 1.5% of non-retracted articles, highlighting the potential of tweets as early warning signals of retraction. However, critical tweets identified by LLMs (GPT-4o mini, Gemini 2.0 Flash-Lite, and Claude 3.5 Haiku) only partially aligned with human annotation, suggesting that fully automated monitoring of post-publication discourse should be applied with caution. A human–AI collaborative approach may offer a more reliable and scalable alternative, with human expertise helping to filter out tweets critical of issues unrelated to the research integrity of the articles. Overall, this study provides insights into how social media signals, combined with generative AI technologies, may support efforts to strengthen research integrity.

1 | INTRODUCTION

Problematic articles are scholarly publications containing methodological flaws, data irregularities, or ethical violations, regardless of whether these issues have been formally identified by the scientific community. The presence of such articles can mislead researchers, policymakers, and the public (Candal-Pedreira et al., 2022; Larsson, 1995). When concerns identified after publication are deemed sufficiently serious, problematic articles may ultimately be retracted. Retraction serves as the primary corrective mechanism, defined as "a mechanism for

correcting the literature and alerting readers to articles that contain such seriously flawed or erroneous content or data that their findings and conclusions cannot be relied upon" (COPE Council, 2019). *Retracted articles* are thus outputs that have been officially withdrawn by journals or publishers due to their unreliability or other concerns.

Despite a notable increase in retractions in recent years (Van Noorden, 2011, 2023; Vuong et al., 2020; Zheng et al., 2025), the retraction process often remains slow and opaque. Problematic articles can remain accessible and continue exerting influence—sometimes accompanied only by a statement of concern—long before decisive editorial action is taken. This delay underscores the need for earlier detection of problematic content and mitigation of its potential harms (Bar-Ilan & Halevi, 2017).

Beyond pre-publication peer review, traditional methods for identifying problematic articles have largely focused on detecting plagiarism (Eysenbach, 2000; Wager, 2011) and image manipulation (Bik et al., 2016; Koppers et al., 2017). However, these approaches have limited scope and are less effective in detecting other forms of misconduct, such as data fabrication, falsification, or authorship disputes. In this context, social media is increasingly recognized as a component of post-publication peer review (Irawan et al., 2024).

Although criticism on platforms such as Twitter,[1] Facebook, or academic blogs does not inherently indicate that an article is problematic or destined for retraction, heightened public scrutiny may prompt further investigation by journals, institutions, or the broader scientific community. In some cases, such scrutiny has accelerated editorial actions, including retractions. Notable examples include exposure of papers containing AI-generated text (Tarla et al., 2023; Zhang et al., 2024), manipulated images (Guo, Dong, & Hao, 2024; Wu et al., 2024), or high-profile studies—such as those related to COVID-19—that were criticized for lacking robust empirical support (Mehra et al., 2020; Walach et al., 2021).

These observations suggest that critical commentary on social media may serve as an early warning signal for identifying problematic articles before formal retraction. While social media discourse has been widely studied within the framework of *altmetrics* as a measure of attention surrounding scholarly publications, it has not yet been systematically explored as a tool for surfacing concerns about scientific integrity.

1.1 | Altmetric research on retracted articles

Existing altmetric research has shown that retracted articles often attract considerable attention on social media, especially on platforms such as Twitter and among the general public (Bornmann & Haunschild, 2018; Khademizadeh et al., 2024; Khan et al., 2022). A positive correlation has been observed between high Altmetric Attention Scores (AAS) and the likelihood of retraction due to misconduct (Shema et al., 2019), suggesting that social media visibility may signal underlying issues.

Moreover, retracted articles typically receive more social media engagement than comparable non-retracted articles (Peng et al., 2022; Serghiou et al., 2021; Sotudeh et al., 2022). Metrics such as the number of tweets, retweets, likes, and replies are often higher for retracted articles, indicating greater public scrutiny or interest (Dambanemuya et al., 2024). Most of this engagement occurs before a formal retraction notice is issued (Dambanemuya et al., 2024; Serghiou et al., 2021). This suggests that monitoring social media activity may help identify problematic content at an earlier stage.

Beyond volume, several studies have examined the content of tweets referencing retracted articles. Dambanemuya et al. (2024) found that retracted articles attracted significantly more tweets containing retraction-related keywords than non-retracted ones, both before and after retraction. Similarly, Peng et al. (2022) reported that tweets about retracted articles were more likely to include critical commentary. Sentiment analysis further supports this: approximately one-third of retracted articles are associated with critical sentiment in tweets both pre- and post-retraction, with critical tweets retweeted more frequently than positive or neutral ones (Amiri et al., 2024).

Case-based evidence reinforces this trend. For example, Haunschild and Bornmann (2021) found that critical tweets questioning the validity of two out of three COVID-19 articles began circulating shortly after their publication. Despite these insights, no prior research appears to have systematically leveraged social media content to assess whether an article is likely to be problematic or at risk of retraction.

1.2 | Patterns of Twitter engagement with scholarly articles

Mentions of scholarly articles on Twitter constitute one of the most prevalent forms of altmetric data (Costas et al., 2015; Ortega, 2018; Sugimoto et al., 2017; Thelwall et al., 2013). While both scientists and the general public contribute to the dissemination of research on Twitter (Yu et al., 2019; Zhang et al., 2023), scientists remain the primary participants in science-related discussions (Carlson & Harris, 2020). They are more likely to use the platform to communicate with colleagues, establish

collaborations, engage in advocacy or political debates, share findings with broader audiences, and stay informed about developments in their fields (Bowman, 2015). The public, in turn, often follows scientists and generally responds positively to scientific engagement on social media (Côté & Darling, 2018). Notably, the topics attracting the attention of scientists and the public tend to be closely aligned (Karmakar et al., 2023).

Twitter engagement typically peaks shortly after an article's publication, making tweet counts one of the fastest-emerging forms of altmetric data (Fang & Costas, 2020; Yu et al., 2017). Around 40% of article-related tweets appear within the first week (Priem & Costello, 2010). However, most interactions consist of low-effort engagement, such as likes and retweets. Higher-level engagement—defined as actions involving discussion, interpretation, critique, or questioning of content—accounts for only a small portion of interactions (Fang et al., 2022). In certain cases, particularly those involving controversial or socially salient topics such as climate change or COVID-19, scholarly tweets have sparked broader public debates (Alperin et al., 2024; Toupin et al., 2022).

Users' motivations for tweeting about scholarly articles are diverse, encompassing scientific information seeking and sharing, professional goals such as network-building or institutional promotion, and communicative purposes like explaining findings to peers or lay audiences (Mohammadi et al., 2018). As such, Twitter plays a multifaceted role in scientific dissemination (Chatterjee et al., 2020), including informal post-publication peer review. Similarly to negative citations in the scholarly literature, tweets can take a critical tone—expressing skepticism, methodological concerns, or even accusations of misconduct (Teixeira da Silva & Dobránszki, 2019). In some of the aforementioned cases, such discussions have brought serious errors or ethical breaches to light.

1.3 | Large language models for classification tasks

Large language models (LLMs), such as ChatGPT, Claude, and Gemini, are transformer-based architectures trained on vast text corpora, enabling them to perform a wide range of natural language processing (NLP) tasks (Zubiaga, 2024). One such task is text classification, defined as the assignment of predefined labels to text, such as sentiment or topic. The ability of LLMs to generalize across domains through in-context learning and prompt engineering makes them particularly promising for classification tasks (Vajjala & Shimangaud, 2025; Wang et al., 2024).

Empirical studies have shown that LLMs often outperform traditional machine learning methods in both multiclass and binary classification tasks. For instance, LLMs have achieved higher accuracy in predicting employees' work locations based on job reviews and in classifying news articles as real or fake (Kostina et al., 2025). In zero-shot settings, where no task-specific training data or examples are provided, LLMs have surpassed support vector machines and even pretrained transformer-based models such as RoBERTa in recall, thereby reducing false negatives during preprocessing (Guo, Ovadje, et al., 2024). ChatGPT has also outperformed crowd workers in various annotation tasks, including relevance, stance, topic, and frame detection for tweets and news articles, with average zero-shot accuracy gains of around 25 percentage points (Gilardi et al., 2023). In addition, LLMs have demonstrated substantial agreement with human annotators in taxonomic labeling across diverse social science corpora (Ziems et al., 2024).

Despite these advances, LLMs have important limitations. They are prone to hallucinations, generating inaccurate or fabricated content that deviates from factual knowledge or the training data (Perković et al., 2024). They may also replicate demographic and historical biases embedded in their training corpora, potentially amplifying stereotypes and producing unfair outputs (Ranjan et al., 2024). Another constraint is the limited context window of most models, which restricts their ability to process long documents. Although recent improvements have extended input capacity, it remains a performance bottleneck for tasks requiring comprehensive document comprehension (Hosseini et al., 2025).

In industry, LLMs have recently been adopted to evaluate social media engagement with scholarly content. For example, Altmetric has introduced a beta feature that applies LLM-powered sentiment analysis to posts on X (formerly Twitter) and Bluesky.[2] This tool classifies posts into seven categories—ranging from strong negative to strong positive—in an effort to capture more nuanced public discourse about research. While this represents a practical step forward, it remains unclear how effectively LLMs can detect *critical content* specifically directed at scholarly articles. If LLMs can reliably identify such criticism, they may offer a scalable method for early detection of potentially problematic research, thereby contributing to efforts to uphold research integrity.

1.4 | Objectives of this study

The widespread engagement with retracted articles on Twitter, along with the growing application of LLMs for

sentiment analysis by platforms such as Altmetric, highlights the need to assess whether social media commentary can serve as an early warning signal for problematic research, and whether LLMs can support this effort.

This study addresses these questions in two parts. First, we examine whether retracted articles attract more critical tweets *prior to* retraction than non-retracted articles, based on human annotation. *Critical tweets* are operationally defined as posts that express criticism, accusations, doubts, sarcasm, irony, or mockery directed at a scholarly article—a definition partially overlapping with Altmetric's negative sentiment categories.

Second, we evaluate the accuracy of LLMs in identifying critical content in tweets, using human annotation as the benchmark. This evaluation not only provides insights into the reliability of LLM-powered sentiment systems for detecting potentially problematic research, but also contributes empirical evidence to support or challenge the validity of using LLMs in the analysis of science communication.

Specifically, we address the following research questions (RQs):

> **RQ1.** To what extent do critical tweets appear in Twitter discussions of retracted versus non-retracted articles?
>
> **RQ2.** How accurately can LLMs detect whether a tweet about a scholarly article is critical of it, using human annotation as the benchmark?

2 | CONCEPTUAL FRAMEWORK

Scholarly peer review is widely recognized as an imperfect process (Heesen & Bright, 2021). Many articles are retracted post-publication despite having passed peer review, leading to wasted research resources—especially for those building upon or attempting to replicate flawed findings—and ultimately undermining trust in the scientific enterprise (Peterson, 2018). In response to concerns about articles that have not been retracted, a complementary evaluation mechanism has emerged: post-publication peer review (PPPR).

Platforms such as PubPeer offer dedicated spaces for ongoing critical discussion of published academic work. These forums often help to publicize methodological or factual errors and facilitate constructive critique, which can contribute to article retractions (Bordignon, 2020). However, relatively few researchers actively participate on formal PPPR platforms. Writing thoughtful and methodologically sound critiques is time-consuming, and early-career researchers may hesitate to publicly criticize

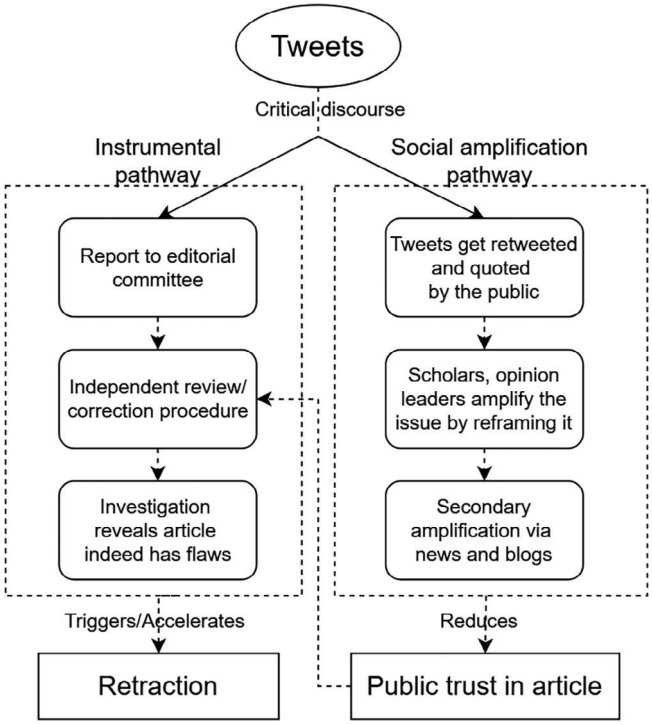

FIGURE 1 Conceptual model of Twitter's influence pathways in the post-publication peer review (PPPR) process of problematic articles.

senior academics due to fear of reputational or professional repercussions (Bastian, 2014).

Informal PPPR has long existed within academia, for example through discussions at conferences or within research groups (Peterson, 2018). With the rise of social media as a tool for informal scholarly communication, these discussions also occur online and can be conceptualized as a form of PPPR (Irawan et al., 2024). Compared to formal platforms, social media offers several unique advantages: its brevity lowers participation barriers, and the possibility of anonymity may encourage researchers—especially junior ones—to voice criticism without fear of professional consequences. Twitter, in particular, plays a central role, accounting for perhaps 80% of social media discussions about scholarly publications (Peng et al., 2022).

Despite its accessibility, the reliability of social media commentary is variable. Users without domain expertise may post unsubstantiated, irrelevant, or misleading content (Ali & Watson, 2016). Unlike specialized PPPR platforms, which often feature detailed critiques that can directly lead to retractions, the role of social media in influencing retraction decisions has yet to be systematically studied at scale. Given Twitter's prominence in science communication, this study focuses on critical tweets as a lens for examining how social media may contribute to the retraction process.

We conceptualize tweets as playing a dual role in the post-publication lifecycle of scholarly articles (see Figure 1), corresponding to two distinct yet interconnected influence pathways through which critical Twitter discussions may affect retraction outcomes:

- *Instrumental pathway*: Tweets that explicitly identify methodological flaws, fabricated data, or ethical violations may function as whistle-blowing signals. Once detected—whether by experts, editorial monitoring tools, or through direct tagging of editors or journals—such tweets may trigger formal investigations by journals or institutions.
- *Social amplification pathway*: Drawing on the *social amplification of risk framework* (Kasperson et al., 1988), initial critical tweets may be retweeted and discussed by others, extending their reach beyond the original audience. Over time, scholars, influencers, or topic-focused accounts may reframe and reinforce the concerns. This cascade of attention can eventually involve mainstream media or influential blogs, amplifying perceived risks and potentially exerting indirect pressure on journals or funders to take formal action, sometimes culminating in article retraction.

3 | DATA AND METHODS

3.1 | Dataset of retracted articles

To identify retracted articles, we used a snapshot of the Web of Science (WoS) database (version dated March 2025), maintained by the Centre for Science and Technology Studies (CWTS) at Leiden University. From the WoS *Core Collection*, we selected articles published in 2019 that were classified as "Retracted Publication" under the document type field. The publication year 2019 was chosen to avoid anomalies in retraction patterns associated with the surge of COVID-19-related publishing in subsequent years. This process yielded a dataset of 2387 distinct retracted articles, for which bibliographic metadata were also extracted from the WoS database. To determine the exact retraction dates, we consulted *Retraction Watch* (accessed in June 2025), which serves as the most comprehensive repository of retracted articles (Brainard, 2018).

Using DOI-based searches in the Altmetric database (snapshot dated November 2022), also maintained by CWTS, we retrieved the tweet IDs of original tweets referencing these retracted articles, as recorded by Altmetric up to that date. The corresponding tweet content was then retrieved via the Twitter API in March 2023,[3] based on the retrieved tweet IDs. The resulting dataset included each tweet's full text and timestamp (i.e., date and time of posting).

As this study focuses specifically on *pre-retraction* discourse, we excluded any retracted articles that lacked original tweets posted before their official retraction date. This filtering step resulted in a dataset of 712 retracted articles that had at least one pre-retraction tweet, representing 29.8% of the initial sample.

3.2 | Methods

Figure 2 illustrates the workflow of data sampling, cleaning, and analysis. Each methodological step is described in detail below.

3.2.1 | Matching method: Coarsened exact matching

To assess whether Twitter discussions can serve as early warning signals for articles that are subsequently retracted, we analyzed tweets referencing either retracted or non-retracted articles (i.e., those not retracted at the time of data collection). Coarsened exact matching (CEM) was applied to construct a comparison group of non-retracted articles matched to those in the retracted dataset. CEM is a matching technique that involves coarsening the values of covariates into groups and then performing exact matching within each group (Blackwell et al., 2009; Iacus et al., 2012). This method reduces the imbalance of covariates between treatment and control groups by excluding observations that lack suitable matches.

Non-retracted articles were drawn from the WoS database and matched to retracted articles based on four covariates: (1) publication venue (i.e., journal), (2) publication year, (3) total number of tweets received, and (4) number of original tweets (excluding retweets). This ensured that the matched retracted and non-retracted articles were published in the same journal and year (2019) and had comparable levels of Twitter attention. Of the 591,527 non-retracted articles published in the same journal issues in 2019 as the 712 retracted articles in our dataset, 213,320 had at least one tweet recorded by Altmetric.

After applying CEM, a total of 1362 articles were successfully matched, comprising 681 retracted and 681 non-retracted articles. Bibliographic and Twitter data for the non-retracted articles were collected using the same procedures as for the retracted articles.

3.2.2 | Tweet filtering

We refined the dataset through three filtering steps: First, for retracted articles, only tweets posted before the

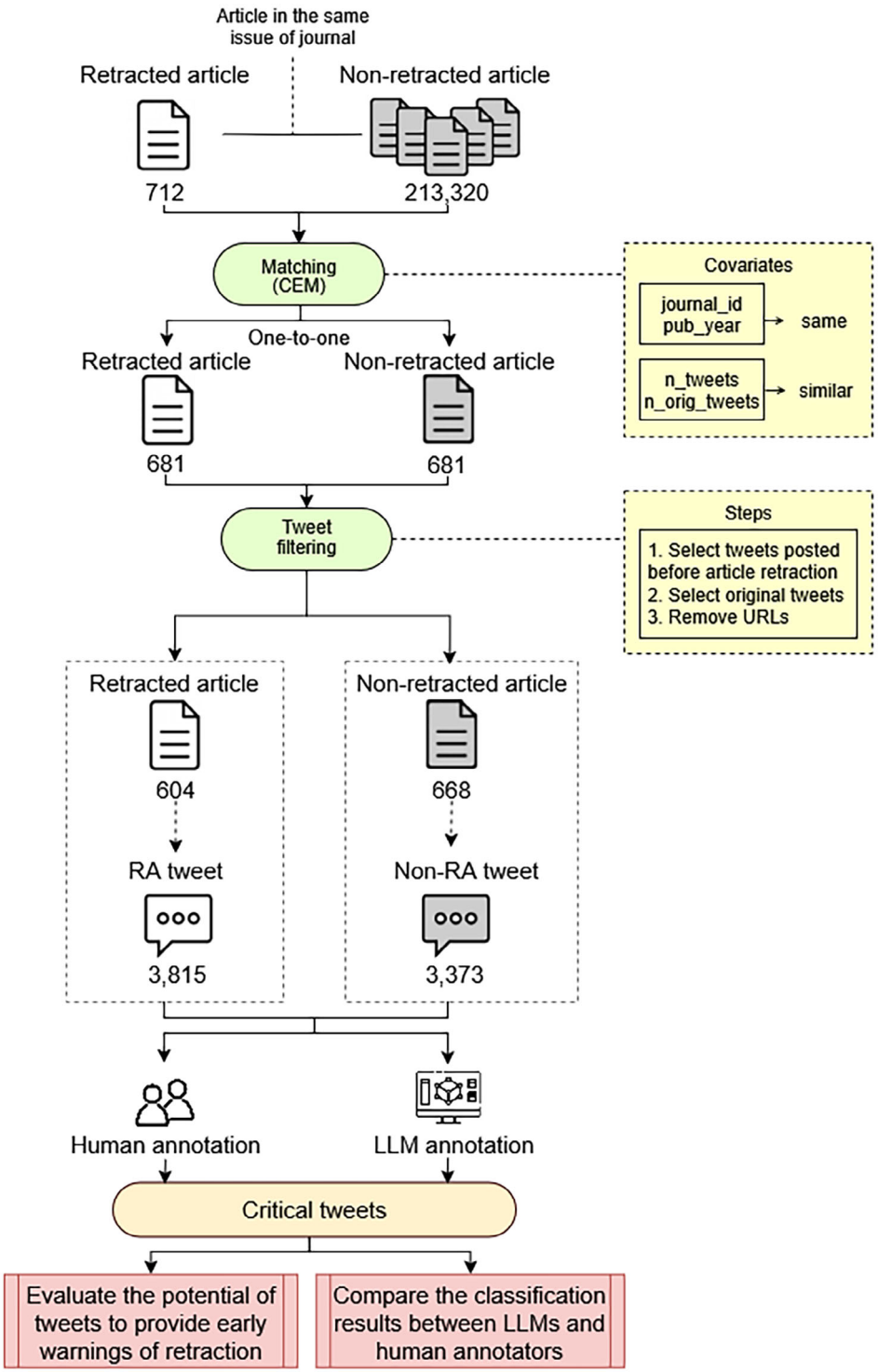

FIGURE 2 Research workflow of the study. "RA" refers to "retracted articles" and "non-RA" to "non-retracted articles," both here and throughout the text.

retraction date were retained. Second, retweets were excluded to ensure that all analyzed tweets contained original textual content. Third, URLs were removed from tweet texts to focus exclusively on semantic content. The final dataset comprised 3815 tweets associated with 604 retracted articles and 3373 tweets associated with 668 comparable non-retracted articles.

3.2.3 | Baseline: human annotation

Two authors (Zheng and Fang) independently annotated all 7188 tweets to determine whether they expressed criticism of the referenced article (see Table 1). Based solely on tweet text, tweets were annotated as *critical* if they conveyed criticism, accusations, doubts, sarcasm, irony,

TABLE 1 Confusion matrix of human annotation by the two annotators.

	Fang: Non-critical	Fang: Critical
Zheng: Non-critical	6785	71
Zheng: Critical	154	178

TABLE 2 Examples of tweets annotated as critical by the two human annotators.

Category	Example tweet
Criticism, accusations	"This paper's methodology is seriously flawed, they cherry-picked data to fit their narrative."
Doubts	"I bet the authors fudged the results to get published in that journal, smells fishy."
Sarcasm, irony, or mockery	"Impressive! Anyone can get into PubMed these days by tossing a draft to some shady open-access journal."

Note: Tweet content has been paraphrased to protect user privacy.

or mockery directed at the article—features considered indicative of potential issues with the article. Tweets meeting this criterion were assigned a label of 1 (*critical*); otherwise, they were labeled 0 (*non-critical*). To ensure consistency, the annotation process followed predefined criteria, and any uncertainties were flagged for subsequent discussion and resolution.

To evaluate inter-coder reliability, Gwet's AC1 coefficient was calculated. This chance-adjusted metric offers greater stability than Cohen's Kappa (Wongpakaran et al., 2013). The resulting Gwet's AC1 value of 0.969 indicates a high level of agreement. Remaining discrepancies were resolved through discussion to finalize the annotation. Table 2 reports examples of tweets annotated as critical, illustrating the range of rhetorical styles included in the classification.

3.2.4 | Large language models

Three LLMs—GPT-4o mini, Gemini 2.0 Flash-Lite, and Claude 3.5 Haiku—were employed to classify tweets as either critical or non-critical. These transformer-based models represent leading LLMs for natural language understanding tasks (Akpan, 2025). The selected model versions were chosen for their high performance-to-cost ratio, making them well-suited for tasks that require nuanced language comprehension under moderate computational constraints.

For each LLM, we conducted experiments using both zero-shot and few-shot prompting strategies (Figure 3). In zero-shot prompting, the model performs the task without being given any exemplars. In contrast, few-shot prompting provides the model with a small set of illustrative examples embedded in the prompt to guide its output (Brown et al., 2020). Employing both strategies allowed for a systematic evaluation of model performance with and without exemplar-based guidance.

All LLM runs were conducted via their respective APIs,[4] with the temperature parameter set to 0 to maximize output consistency and reproducibility. Experiments were performed on June 10, 2025. For each tweet, three independent classification runs were executed per model under both zero-shot and few-shot prompt settings. The majority vote across the three runs was recorded as the model's consensus output for each prompt setting, which was then compared against the human-annotated ground truth. In addition to individual model evaluations, we generated a cross-model consensus output—defined as the majority vote across the consensus outputs of all three LLMs—to assess whether ensemble agreement yielded better alignment with human annotation than any single model.

The consistency rate—defined as the proportion of identical classification outputs across repeated runs of the same model—is reported in Table 3. The results show high internal consistency across all models and prompt settings, suggesting that the models produced highly reproducible outputs under fixed conditions.

3.3 | Indicators

To evaluate the performance of the LLMs in identifying tweets containing critical commentary, we employed three widely used evaluation metrics: precision, recall, and F1-Score. These indicators offer complementary perspectives on classification performance and enable robust model comparisons. In all cases, human annotation was treated as the ground truth.

- *Precision*, also known as positive predictive value, represents the proportion of correctly predicted positive cases among all cases predicted as positive. In this study, it reflects the probability that tweets classified as critical by an LLM were also labeled as critical by human annotators. A precision score of 100% would indicate that all tweets classified as critical by the model were indeed judged critical by humans, with no false positives.
- *Recall*, also known as sensitivity, denotes the proportion of correctly predicted positive cases among all actual positive cases. Here, it reflects the probability that tweets labeled as critical by human annotators

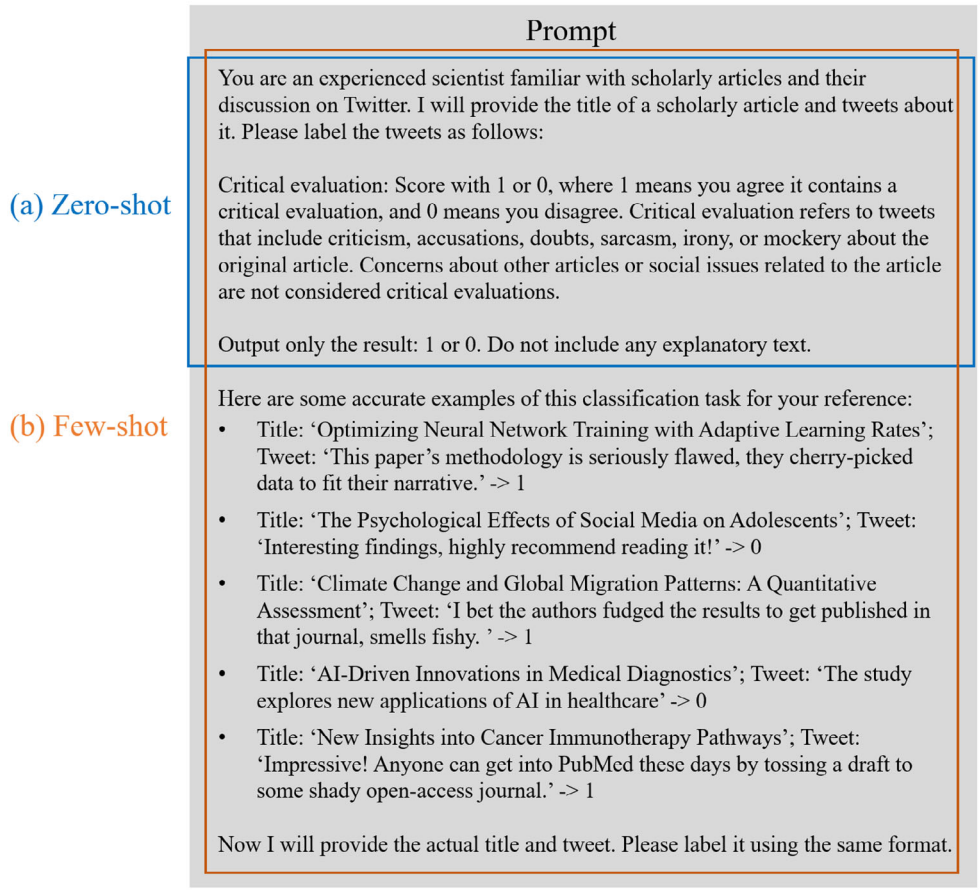

FIGURE 3 Prompt settings used for the LLMs: (a) zero-shot and (b) few-shot.

TABLE 3 Consistency rates across repeated runs for each model.

Prompt setting	GPT-4o mini (%)	Gemini 2.0 Flash-Lite (%)	Claude 3.5 Haiku (%)
Zero-shot	98.8	99.9	99.6
Few-shot	99.4	99.8	99.5

were also classified as critical by the model. A recall score of 100% would indicate that the model successfully captured all tweets deemed critical by humans.

- *F1-Score* combines both precision and recall into a single harmonic mean, offering a balanced measure of classification performance. It is calculated as:

$$\mathrm{F1-Score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

The F1-Score ranges from 0 to 1, with higher values indicating better overall performance. In this study, F1-Scores are reported in percentage format for consistency with other metrics. An F1-Score of 100% would signify that the model achieves perfect precision and recall, flawlessly identifying all critical tweets with no misclassifications.

4 | RESULTS

4.1 | Distribution of critical tweets across retracted and non-retracted articles

Approximately 6.2% of tweets referencing retracted articles were manually classified as critical of the corresponding research. In contrast, only 1.1% of tweets referencing non-retracted articles were annotated as critical, revealing a substantial disparity in critical discourse between the two groups (Figure 4a).

At the article level, 8.3% of retracted articles were associated with at least one critical tweet prior to retraction, compared to only 1.5% of non-retracted articles (Figure 4b). These results suggest that nearly one in twelve retracted articles in the sample could have been flagged for scrutiny based on pre-retraction Twitter activity. Importantly, the 1.5% of non-retracted articles that attracted critical commentary may also warrant closer

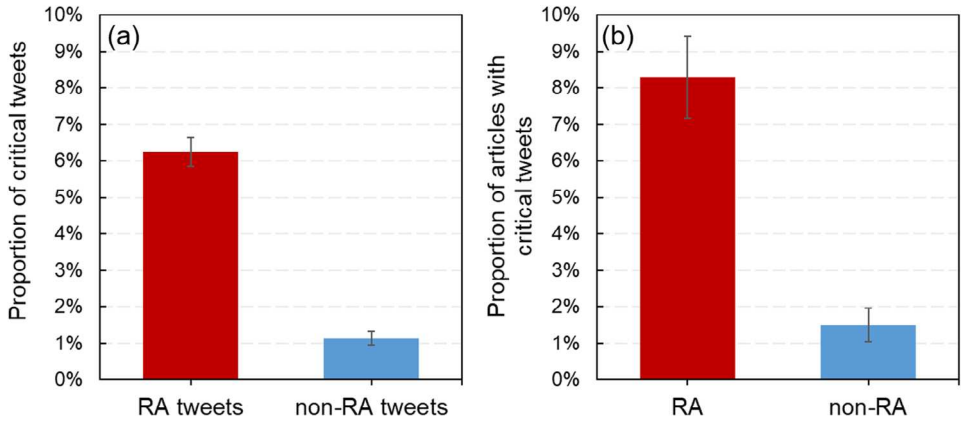

FIGURE 4 (a) Proportion of critical tweets, and (b) proportion of retracted (RA) and non-retracted (non-RA) articles associated with at least one critical tweet.

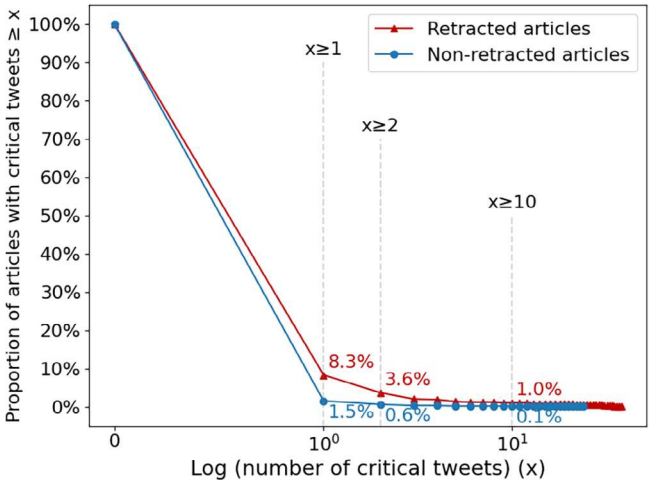

FIGURE 5 Distribution of critical tweets across retracted and non-retracted articles.

scrutiny, as they could potentially harbor unresolved issues.

To explore this pattern further, we plotted a survival function showing the cumulative distribution of critical tweets across articles (Figure 5). The results indicate that 3.6% of retracted articles received two or more critical tweets, and 1.0% received ten or more. In contrast, only 0.6% of non-retracted articles received at least two critical tweets, and just 0.1% received ten or more. These results suggest not only that retracted articles are more likely to receive critical attention overall, but also that they tend to accumulate a greater volume of critical commentary at the individual article level.

4.2 | Comparison of classification results between human annotation and LLMs

Table 4 presents the performance of each LLM in classifying tweets, using human annotation as the benchmark. Across all LLMs, few-shot prompts consistently yielded superior overall performance compared to zero-shot prompts, as reflected by higher F1-Scores.

However, regardless of the prompt setting, all LLMs exhibited relatively low precision—ranging from 22.10% to 49.85%—indicating a substantial rate of false positives. In other words, many tweets classified as critical by the models were not judged as such by human annotators. In contrast, recall values were generally higher (61.59% to 77.54%), suggesting that the models were moderately effective in capturing tweets considered critical by human annotators. Nevertheless, the recall scores remained far from prefect, reflecting the limitations of LLMs in reliably detecting critical commentary directed at scholarly articles.

Importantly, the consensus output across the three LLMs did not outperform the best-performing individual model, Gemini 2.0 Flash-Lite, which achieved the highest F1-Score across all configurations. This suggests that simple ensemble voting does not necessarily enhance classification performance in this context.

To better understand the nature of classification errors, we conducted a qualitative analysis of selected false positives, that is, tweets marked as critical by the LLMs but judged non-critical by human annotators. As shown in Table 5, these misclassifications often stemmed from the model's difficulty in distinguishing tweets that directly critique an article from those that merely reference it. In many instances, tweets used the article as a springboard to discuss broader topics, such as media coverage, ethical debates, or other research, rather than criticizing the article itself.

5 | DISCUSSION

5.1 | The potential of Twitter discussions to support research integrity

Human annotation revealed that 8.3% of retracted articles were associated with at least one critical tweet prior to

Method	Prompt setting	Precision (%)	Recall (%)	F1-Score (%)
GPT-4o mini	Zero-shot	31.33	77.54	44.63
	Few-shot	40.24	73.91	52.11
Gemini 2.0 Flash-Lite	Zero-shot	49.85	61.59	55.10
	Few-shot	48.59	68.84	56.97
Claude 3.5 Haiku	Zero-shot	22.10	77.17	34.36
	Few-shot	30.12	73.55	42.74
LLMs consensus	Zero-shot	34.92	74.28	47.51
	Few-shot	43.78	73.91	54.99

TABLE 4 Performance of different LLMs using human annotation as the benchmark.

TABLE 5 Examples of tweets misclassified by LLMs as critical of articles, with explanations from human annotators.

Example tweet	Explanation by human annotators
"Has Jurassic Park become a reality?"	This expresses astonishment at the attempt described in the article, not criticism
"Our findings uncovered substantial evidence warranting a closer examination of the research conducted by this team. This evidence raises serious doubts about the reliability of most of their published studies."	The tweet promotes the authors' own critique of other studies, not the article itself
"You can find the original research paper via this link. Upon review, it's clear that the actual findings significantly underdeliver compared to the overstated claims in the news."	The critique targets media misrepresentation, not the article directly
"I'm no specialist by any means, but it seems like the kind of information you'd want to have in hand prior to experimenting with CRISPR-engineered babies."	The tweet prompts ethical reflection, not direct criticism of the article's content

Note: All tweets have been paraphrased to protect user privacy.

retraction, compared to only 1.5% of non-retracted articles. This substantial difference indicates that critical tweets are more commonly associated with articles ultimately retracted. These tweets often highlight methodological flaws, express concerns about misconduct, or employ sarcasm and irony to question an article's credibility. Consequently, critical discourse on social media may serve as an early—albeit partial—indicator of potential problematic research.

Our findings, based on a large, systematically constructed dataset, corroborate insights from earlier small-scale case studies (Haunschild & Bornmann, 2021), reinforcing the value of social media as a supplementary channel for post-publication scrutiny. Twitter thus has potential as a low-cost, real-time early warning system for stakeholders concerned with research integrity. Journal editors could integrate Twitter monitoring into post-publication workflows to identify articles warranting further review. Funding agencies and research integrity offices might use such signals to prioritize investigative resources. For researchers, early critical feedback via social media could reveal methodological weaknesses, limit the dissemination of flawed findings, and strengthen the self-correcting mechanisms of science.

Previous studies have demonstrated that negative sentiment and critical keywords are more prevalent in tweets referencing retracted articles (Amiri et al., 2024; Dambanemuya et al., 2024; Peng et al., 2022). Our manual and LLM-based classifications refine these observations by highlighting that negative sentiment or critical keywords do not necessarily reflect criticism of the article itself. Many tweets direct criticism at related societal issues, media representations, or broader debates, rather than at the article's content or methods. This insight underscores the need to move beyond simplistic sentiment analysis and instead consider the referential focus and rhetorical intent of social media discourse.

Despite the presence of some critical tweets, the majority of tweets about retracted articles are not critical. Several factors may account for this. First, Twitter is not optimally designed for scholarly discussions. Many users are not academics and may lack the expertise to offer critical evaluations (Haustein, 2019; Zhang et al., 2023). Second, scholarly Twitter activity is often geared toward dissemination rather than critique (Didegah et al., 2018; Mohammadi et al., 2018), with many tweets merely sharing article links without substantive commentary (Robinson-Garcia et al., 2017). Third, bot accounts play a notable role in disseminating scientific content on Twitter—both for regular articles (Didegah et al., 2018) and for retracted ones (Dambanemuya et al., 2024)—yet

rarely engage in substantive or critical assessments. Together, these factors limit the overall prevalence of critical tweets, even for articles that are eventually retracted.

5.2 | The potential of LLMs to support research integrity

Although concerns have arisen regarding the misuse of LLMs in academic writing and data fabrication (Conroy, 2023a; Naddaf, 2023; Silva et al., 2023), these tools offer promising avenues for supporting research evaluation and scientific standards. For example, LLMs have been used to predict future Nobel Prize laureates by identifying major discoveries made by living scientists (Conroy, 2023b), and to generate peer review comments and quality assessments for scholarly articles, often aligning well with human evaluations (Liang et al., 2024). LLM-generated quality scores have also been shown to correlate positively with expert assessments in evaluative contexts (Thelwall et al., 2025).

In more direct applications related to research integrity, LLMs have been used to detect predatory journals (Al-Moghrabi et al., 2024), identify retracted COVID-19 articles (Jan, 2025), and assist with error detection through AI-based tools such as the *Black Spatula Project* and *YesNoError* (Gibney, 2025). Our study extends this emerging literature by exploring whether LLMs can classify tweets as critical of scholarly articles, thereby assessing the feasibility of scalable, automated monitoring of social media as a supplementary mechanism for identifying potential research misconduct.

Despite prior evidence that LLMs can match or exceed human performance in various annotation tasks (Gilardi et al., 2023; Ziems et al., 2024), our findings reveal only partial alignment between LLM-generated labels and human annotations for critical tweets. This suggests that caution is warranted when interpreting the outputs of LLMs in this domain.

Nevertheless, LLMs could be used to triage social media data by filtering out non-critical tweets, directing human attention to potentially concerning posts. Such a hybrid approach could reduce the workload of human reviewers while maintaining high sensitivity. Although misclassifications may still occur, a two-step workflow—initial classification by LLMs followed by human validation—offers a scalable and efficient alternative to human annotation alone. Articles generating a high volume of tweets flagged as critical by LLMs could be prioritized for further investigation, with human reviewers distinguishing tweets that directly critique the article from those addressing unrelated issues.

A key contribution of this study lies in shifting the focus of AI-based detection from article content to the surrounding social media discourse. Instead of solely scrutinizing articles, we examined public conversations that may reveal otherwise hidden concerns. This approach expands existing research integrity monitoring strategies and highlights the role of social media not only in dissemination, but also in mobilizing collective intelligence to safeguard scientific credibility.

However, several cautions must be highlighted regarding the use of LLMs in this context. First, LLMs are trained on broad and heterogeneous corpora, which may embed social, cultural, or epistemic biases that can affect classification accuracy (Gallegos et al., 2024; Guo, Guo, et al., 2024). Second, deploying advanced models such as GPT-4.5, Gemini 2.5 Pro, or Claude Opus 4 at scale can be computationally intensive and costly (Sathish et al., 2024), raising concerns about the accessibility and scalability of such approaches, particularly for resource-constrained institutions. Third, tweets may contain copyrighted, personal, or sensitive information (Kierkegaard, 2010; Small et al., 2012), and their automated processing by LLMs raises privacy and intellectual property concerns. Ensuring compliance with data protection regulations, such as the General Data Protection Regulation (GDPR),[5] is essential in real-world monitoring scenarios. Finally, while automation can enhance efficiency, overreliance on LLMs may erode critical human judgment, especially in cases requiring contextual understanding or disciplinary expertise. Persistent trust in model outputs could diminish researchers' agency and creativity in making independent critical evaluations (Kumar et al., 2025).

5.3 | Limitations of the study

Several limitations should be considered when interpreting the findings of this study. First, although human annotation was used as the benchmark for evaluating LLM performance, human judgment is inherently subjective and susceptible to bias. While the two annotators discussed uncertainties to reach agreement, some misclassifications may remain, potentially influencing the results.

Second, not all critical tweets reflect verifiable issues in the referenced articles, nor do they reliably predict eventual retractions. In some cases, genuine criticism may arise from ideological disagreement rather than methodological flaws. For instance, scientifically rigorous climate research may attract negative responses from climate change denialists. As such, the presence of critical tweets should be interpreted as a preliminary signal

rather than a definitive judgment of an article's quality or integrity.

Third, this study did not differentiate between retraction reasons, such as methodological errors, data fabrication, ethical violations, or plagiarism (Fang et al., 2012). These distinct causes may elicit varying patterns of social media engagement. Future research could investigate whether specific categories of retraction are more readily detectable through critical online discourse.

Fourth, we did not analyze the thematic content of the articles, which likely affects both the frequency and tone of Twitter engagement. Controlling for thematic variation could help clarify the relationship between content and critique. Additionally, our analysis focused exclusively on tweet text and did not incorporate engagement metrics (e.g., likes, retweets, and replies) or contextual data about users (e.g., whether the account belongs to a bot, journalist, or academic). These factors may shape both the interpretation and amplification of tweets and should be considered in future work (Dambanemuya et al., 2024).

Lastly, recent structural changes to Twitter—such as the shift to paid API access and stricter rate limits—have constrained researchers' ability to collect large-scale, up-to-date datasets. These changes may also influence the volume and nature of academic discourse on the platform. In parallel, many researchers are migrating to alternative platforms such as Bluesky or Mastodon (Bittermann et al., 2023; Kupferschmidt, 2024; Mallapaty, 2024; Vidal Valero, 2023), reducing Twitter's representativeness as a venue for science communication. Nonetheless, the methodological framework developed in this study is adaptable and can be applied to other platforms. This presents opportunities for future research to extend the investigation of scholarly discourse and critical engagement across a broader social media ecosystem.

6 | CONCLUSIONS

This study examined whether critical tweets can serve as early warning signals for retracted scholarly articles and assessed the potential of LLMs to automate the detection of such tweets. Manual analysis revealed that 8.3% of retracted articles were associated with at least one critical tweet prior to retraction, compared to just 1.5% of non-retracted articles. These findings underscore Twitter's potential as a supplementary channel for post-publication review and research integrity monitoring.

LLMs' ability to detect critical content aligned only moderately with human annotation. In particular, the models' low precision indicates a relatively high rate of false positives. This suggest that the selected LLMs, in their current form, are not yet reliable as standalone tools for critical tweet detection. However, they hold promise as components of hybrid workflows that combine automated filtering with human validation to support scalable monitoring.

This study contributes a novel perspective by shifting attention from analyzing article content to examining surrounding social media discourse. Despite certain limitations—including the relative rarity of critical tweets and the evolving platform constraints—our methodological framework offers a scalable, transferable approach for identifying potentially problematic research through real-time public commentary. Future work could extend this framework to emerging platforms and incorporate additional contextual signals to improve detection accuracy and enhance practical utility.

ACKNOWLEDGMENTS

Zhichao Fang is financially supported by the National Natural Science Foundation of China (No. 72304274). Hui-Zhen Fu is supported by the National Social Science Foundation of China (No. 22CTQ032). Er-Te Zheng is financially supported by the GTA scholarship from the School of Information, Journalism and Communication of the University of Sheffield. Mike Thelwall is supported by the Fundação Calouste Gulbenkian European Media and Information Fund (No. 316177). The authors thank Altmetric and Retraction Watch for providing the data for research purposes, and the anonymous reviewers for their valuable comments.

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.

ORCID

Er-Te Zheng https://orcid.org/0000-0001-8759-3643
Hui-Zhen Fu https://orcid.org/0000-0002-1534-9374
Mike Thelwall https://orcid.org/0000-0001-6065-205X
Zhichao Fang https://orcid.org/0000-0002-3802-2227

ENDNOTES

[1] Twitter was the official name of X at the time of data collection, so we use the terms "Twitter" and "tweets" in this study.

[2] For more information on the sentiment analysis introduced by Altmetric, see: https://help.altmetric.com/support/solutions/articles/6000279392-sentiment-analysis-in-altmetric (accessed on June 19, 2025).

[3] Following the acquisition of Twitter by Elon Musk and its subsequent rebranding as X, the platform discontinued free access to its API in March 2023. Consequently, the tweet data collected in

March 2023 represents the most recent large-scale dataset obtainable under the previous access conditions. The lack of up-to-date tweet data constitutes a key limitation of this study. However, since the majority of tweets referencing scholarly articles typically accumulate shortly after publication (Fang & Costas, 2020; Priem & Costello, 2010; Yu et al., 2017), the 4-year window (2019–2022) is likely sufficient for capturing most Twitter engagement with the articles published in 2019 that are included in our datasets. Additionally, given that Twitter (now X) remains the most prominent source of altmetric data, the methodology developed in this study remains applicable to scholarly discourse on the platform and is potentially transferable to other similar social media environments.

[4] ChatGPT: https://platform.openai.com/docs/api-reference/assistants/createAssistant (accessed on March 31, 2025); Gemini: https://ai.google.dev/gemini-api/docs/models/generative-models?hl=en (accessed on March 31, 2025); Claude: https://docs.anthropic.com/en/api/complete (accessed on March 31, 2025).

[5] For more information on this regulation, see: https://gdpr-info.eu/ (accessed on June 24, 2025).

REFERENCES

Akpan, M. (2025). Have we reached artificial general intelligence? Comparison of ChatGPT, Claude, and Gemini to human literacy and education benchmarks. *Corporate Ownership and Control*, *22*(1), 103–110. https://doi.org/10.22495/cocv22i1art8

Ali, P. A., & Watson, R. (2016). Peer review and the publication process. *Nursing Open*, *3*(4), 193–202. https://doi.org/10.1002/nop2.51

Al-Moghrabi, D., Abu Arqub, S., Maroulakos, M. P., Pandis, N., & Fleming, P. S. (2024). Can ChatGPT identify predatory biomedical and dental journals? A cross-sectional content analysis. *Journal of Dentistry*, *142*, 104840. https://doi.org/10.1016/j.jdent.2024.104840

Alperin, J. P., Fleerackers, A., Riedlinger, M., & Haustein, S. (2024). Second-order citations in altmetrics: A case study analyzing the audiences of COVID-19 research in the news and on social media. *Quantitative Science Studies*, *5*(2), 366–382. https://doi.org/10.1162/qss_a_00298

Amiri, M., Yaghtin, M., & Sotudeh, H. (2024). How do tweeters feel about scientific misinformation: An infoveillance sentiment analysis of tweets on retraction notices and retracted papers. *Scientometrics*, *129*(1), 261–287. https://doi.org/10.1007/s11192-023-04871-7

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: A case study. *Scientometrics*, *113*(1), 547–565. https://doi.org/10.1007/s11192-017-2242-0

Bastian, H. (2014). A stronger post-publication culture is needed for better science. *PLoS Medicine*, *11*(12), e1001772. https://doi.org/10.1371/journal.pmed.1001772

Bik, E. M., Casadevall, A., & Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *MBio*, *7*(3), e00809-16. https://doi.org/10.1128/mbio.00809-16

Bittermann, A., Lauer, T., & Peters, F. (2023). Academic #TwitterMigration to mastodon: The role of influencers and the open science movement. *PsychArchives*. https://doi.org/10.23668/psycharchives.13062

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). Cem: Coarsened exact matching in Stata. *The Stata Journal*, *9*(4), 524–546. https://doi.org/10.1177/1536867X0900900402

Bordignon, F. (2020). Self-correction of science: A comparative study of negative citations and post-publication peer review. *Scientometrics*, *124*(2), 1225–1239. https://doi.org/10.1007/s11192-020-03536-z

Bornmann, L., & Haunschild, R. (2018). Allegation of scientific misconduct increases Twitter attention. *Scientometrics*, *115*(2), 1097–1100. https://doi.org/10.1007/s11192-018-2698-6

Bowman, T. D. (2015). Differences in personal and professional tweets of scholars. *Aslib Journal of Information Management*, *67*(3), 356–371. https://doi.org/10.1108/AJIM-12-2014-0180

Brainard, J. (2018). Rethinking retractions. *Science*, *362*(6413), 390–393. https://doi.org/10.1126/science.362.6413.390

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems, 1877–1901.

Candal-Pedreira, C., Pérez-Ríos, M., & Ruano-Ravina, A. (2022). Retraction of scientific papers: Types of retraction, consequences, and impacts. In J. Faintuch & S. Faintuch (Eds.), *Integrity of scientific research: Fraud, misconduct and fake news in the academic, medical and social environment* (pp. 397–407). Springer International Publishing. https://doi.org/10.1007/978-3-030-99680-2_40

Carlson, J., & Harris, K. (2020). Quantifying and contextualizing the impact of bioRxiv preprints through automated social media audience segmentation. *PLoS Biology*, *18*(9), e3000860. https://doi.org/10.1371/journal.pbio.3000860

Chatterjee, S., Rana, N. P., & Dwivedi, Y. K. (2020). Social media as a tool of knowledge sharing in academia: An empirical study using valance, instrumentality and expectancy (VIE) approach. *Journal of Knowledge Management*, *24*(10), 2531–2552. https://doi.org/10.1108/JKM-04-2020-0252

Conroy, G. (2023a). How ChatGPT and other AI tools could disrupt scientific publishing. *Nature*, *622*(7982), 234–236. https://doi.org/10.1038/d41586-023-03144-w

Conroy, G. (2023b). Can AI predict who will win a Nobel Prize? *Nature*. https://www.nature.com/articles/d41586-023-03074-7

COPE Council. (2019, November 2). *COPE retraction guidelines—English*. https://doi.org/10.24318/cope.2019.1.4

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019. https://doi.org/10.1002/asi.23309

Côté, I. M., & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops? *FACETS*, *3*(1), 682–694. https://doi.org/10.1139/facets-2018-0002

Dambanemuya, H. K., Abhari, R., Vincent, N., & Horvát, E.-Á. (2024). Online engagement with retracted articles: Who, when, and how? (No. arXiv:2203.04228; Version 3). *arXiv*. http://arxiv.org/abs/2203.04228

Didegah, F., Mejlgaard, N., & Sørensen, M. P. (2018). Investigating the quality of interactions and public engagement around

scientific papers on Twitter. *Journal of Informetrics*, *12*(3), 960–971. https://doi.org/10.1016/j.joi.2018.08.002

Eysenbach, G. (2000). Report of a case of cyberplagiarism—And reflections on detecting and preventing academic misconduct using the Internet. *Journal of Medical Internet Research*, *2*(1), e793. https://doi.org/10.2196/jmir.2.1.e4

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033. https://doi.org/10.1073/pnas.1212247109

Fang, Z., & Costas, R. (2020). Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics*, *123*(2), 1077–1101. https://doi.org/10.1007/s11192-020-03405-9

Fang, Z., Costas, R., & Wouters, P. (2022). User engagement with scholarly tweets of scientific papers: A large-scale and cross-disciplinary analysis. *Scientometrics*, *127*(8), 4523–4546. https://doi.org/10.1007/s11192-022-04468-6

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524/2471010

Gibney, E. (2025). AI tools are spotting errors in research papers: Inside a growing movement. *Nature*. https://doi.org/10.1038/d41586-025-00648-5

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Guo, X., Dong, L., & Hao, D. (2024). RETRACTED: Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. *Frontiers in Cell and Developmental Biology*, *11*, 1339390. https://doi.org/10.3389/fcell.2023.1339390

Guo, Y., Guo, M., Su, J., Yang, Z., Zhu, M., Li, H., Qiu, M., & Liu, S. S. (2024). Bias in large language models: Origin, evaluation, and mitigation (No. arXiv:2411.10915). *arXiv*. https://doi.org/10.48550/arXiv.2411.10915

Guo, Y., Ovadje, A., Al-Garadi, M. A., & Sarker, A. (2024). Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, *31*(10), 2181–2189. https://doi.org/10.1093/jamia/ocae210

Haunschild, R., & Bornmann, L. (2021). Can tweets be used to detect problems early with scientific papers? A case study of three retracted COVID-19/SARS-CoV-2 papers. *Scientometrics*, *126*(6), 5181–5199. https://doi.org/10.1007/s11192-021-03962-7

Haustein, S. (2019). Scholarly Twitter metrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 729–760). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_28

Heesen, R., & Bright, L. K. (2021). Is peer review a good idea? *The British Journal for the Philosophy of Science*, *72*(3), 635–663. https://doi.org/10.1093/bjps/axz029

Hosseini, P., Castro, I., Ghinassi, I., & Purver, M. (2025). Efficient solutions for an intriguing failure of LLMs: Long context window does not mean LLMs can analyze long sequences flawlessly. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 1880–1891). Association for Computational Linguistics.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Irawan, D. E., Pourret, O., Besançon, L., Herho, S. H. S., Ridlo, I. A., & Abraham, J. (2024). Post-publication review: The role of science news outlets and social media. *Annals of Library and Information Studies*, *71*(4), 465–474. https://doi.org/10.56042/alis.v71i4.14254

Jan, R. (2025). Examining the reliability of ChatGPT: Identifying retracted scientific literature and ensuring accurate citations and references. In *Impacts of Generative AI on the Future of Research and Education* (pp. 367–392). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0884-4.ch014

Karmakar, M., Banshal, S. K., & Singh, V. K. (2023, September 27). Exploring twitter for scientific and public engagement with scholarly articles. *27th International Conference on Science, Technology and Innovation Indicators* (STI 2023), Leiden, The Netherlands. https://doi.org/10.55835/6442b5d9f1b0f0c586cb14ac

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis*, *8*(2), 177–187. https://doi.org/10.1111/j.1539-6924.1988.tb01168.x

Khademizadeh, S., Danesh, F., Esmaeili, S., Lund, B., & Santos-d'Amorim, K. (2024). Evolution of retracted publications in the medical sciences: Citations analysis, bibliometrics, and altmetrics trends. *Accountability in Research*, *31*(8), 1182–1197. https://doi.org/10.1080/08989621.2023.2223996

Khan, H., Gupta, P., Zimba, O., & Gupta, L. (2022). Bibliometric and altmetric analysis of retracted articles on COVID-19. *Journal of Korean Medical Science*, *37*(6), e44. https://doi.org/10.3346/jkms.2022.37.e44

Kierkegaard, S. (2010). Twitter thou doeth? *Computer Law and Security Review*, *26*(6), 577–594. https://doi.org/10.1016/j.clsr.2010.09.002

Koppers, L., Wormer, H., & Ickstadt, K. (2017). Towards a systematic screening tool for quality assurance and semiautomatic fraud detection for images in the Life Sciences. *Science and Engineering Ethics*, *23*(4), 1113–1128. https://doi.org/10.1007/s11948-016-9841-7

Kostina, A., Dikaiakos, M. D., Stefanidis, D., & Pallis, G. (2025). Large language models for text classification: Case study and comprehensive review (No. arXiv:2501.08457). *arXiv*. https://doi.org/10.48550/arXiv.2501.08457

Kumar, H., Vincentius, J., Jordan, E., & Anderson, A. (2025). Human creativity in the age of LLMs: Randomized experiments on divergent and convergent thinking. *Proceedings of the 2025 CHI conference on human factors in computing systems*, 1–18. https://doi.org/10.1145/3706598.3714198

Kupferschmidt, K. (2024). Researchers and scientific institutions flock to Bluesky. *Science*, *386*(6725), 950–951. https://doi.org/10.1126/science.adu8276

Larsson, K. S. (1995). The dissemination of false data through inadequate citation. *Journal of Internal Medicine*, *238*(5), 445–450. https://doi.org/10.1111/j.1365-2796.1995.tb01222.x

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), AIoa2400196. https://doi.org/10.1056/AIoa2400196

Mallapaty, S. (2024). 'A place of joy': Why scientists are joining the rush to Bluesky. *Nature*, *636*(8041), 15–16. https://doi.org/10.1038/d41586-024-03784-6

Mehra, M. R., Desai, S. S., Kuy, S., Henry, T. D., & Patel, A. N. (2020). Retracted: Cardiovascular disease, drug therapy, and mortality in Covid-19. *New England Journal of Medicine*, *382*, e102. https://doi.org/10.1056/NEJMoa2007621

Mohammadi, E., Thelwall, M., Kwasny, M., & Holmes, K. L. (2018). Academic information on Twitter: A user survey. *PLoS One*, *13*(5), e0197265. https://doi.org/10.1371/journal.pone.0197265

Naddaf, M. (2023). ChatGPT generates fake data set to support scientific hypothesis. *Nature*, *623*(7989), 895–896. https://doi.org/10.1038/d41586-023-03635-w

Ortega, J. L. (2018). The life cycle of altmetric impact: A longitudinal study of six metrics from PlumX. *Journal of Informetrics*, *12*(3), 579–589. https://doi.org/10.1016/j.joi.2018.06.001

Peng, H., Romero, D. M., & Horvát, E.-Á. (2022). Dynamics of cross-platform attention to retracted papers. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(25), e2119086119. https://doi.org/10.1073/pnas.2119086119

Perković, G., Drobnjak, A., & Botički, I. (2024). Hallucinations in LLMs: Understanding and addressing challenges. 2024 47th MIPRO ICT and Electronics Convention (MIPRO), 2084-2088. https://doi.org/10.1109/MIPRO60963.2024.10569238

Peterson, G. I. (2018). Postpublication peer review: A crucial tool. *Science*, *359*(6381), 1225–1226. https://doi.org/10.1126/science.aas9490

Priem, J., & Costello, K. L. (2010). How and why scholars cite on Twitter: How and why scholars cite on Twitter. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4. https://doi.org/10.1002/meet.14504701201

Ranjan, R., Gupta, S., & Singh, S. N. (2024). A comprehensive survey of bias in LLMs: Current landscape and future directions (No. arXiv:2409.16430). *arXiv*. https://doi.org/10.48550/arXiv.2409.16430

Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PLoS One*, *12*(8), e0183551. https://doi.org/10.1371/journal.pone.0183551

Sathish, V., Lin, H., Kamath, A. K., & Nyayachavadi, A. (2024). LLeMpower: Understanding disparities in the control and access of large language models (No. arXiv:2404.09356). *arXiv*. https://doi.org/10.48550/arXiv.2404.09356

Serghiou, S., Marton, R. M., & Ioannidis, J. P. A. (2021). Media and social media attention to retracted articles according to Altmetric. *PLoS One*, *16*(5), e0248625. https://doi.org/10.1371/journal.pone.0248625

Shema, H., Hahn, O., Mazarakis, A., & Peters, I. (2019). Retractions from altmetric and bibliometric perspectives. *Information – Wissenschaft & Praxis*, *70*(2–3), 98–110. https://doi.org/10.1515/iwp-2019-2006

Silva, T. P., Ocampo, T. S. C., Alencar-Palha, C., Oliveira-Santos, C., Takeshita, W. M., & Oliveira, M. L. (2023). ChatGPT: A tool for scientific writing or a threat to integrity? *British Journal of Radiology*, *96*(1152), 20230430. https://doi.org/10.1259/bjr.20230430

Small, H., Kasianovitz, K., Blanford, R., & Celaya, I. (2012). What your tweets tell us about you: Identity, ownership and privacy of Twitter data. *International Journal of Digital Curation*, *7*(1), 174–197. https://doi.org/10.2218/ijdc.v7i1.224

Sotudeh, H., Barahmand, N., Yousefi, Z., & Yaghtin, M. (2022). How do academia and society react to erroneous or deceitful claims? The case of retracted articles' recognition. *Journal of Information Science*, *48*(2), 182–198. https://doi.org/10.1177/0165551520945853

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062. https://doi.org/10.1002/asi.23833

Tarla, S., Ali, K. K., & Yusuf, A. (2023). Retracted: Exploring new optical solutions for nonlinear Hamiltonian amplitude equation via two integration schemes. *Physica Scripta*, *98*(9), 095218. https://doi.org/10.1088/1402-4896/aceb40

Teixeira da Silva, J. A., & Dobránszki, J. (2019). A new dimension in publishing ethics: Social media-based ethics-related accusations. *Journal of Information, Communication and Ethics in Society*, *17*(3), 354–370. https://doi.org/10.1108/JICES-05-2018-0051

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS One*, *8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841

Thelwall, M., Jiang, X., & Bath, P. A. (2025). Estimating the quality of published medical research with ChatGPT. *Information Processing & Management*, *62*(4), 104123. https://doi.org/10.1016/j.ipm.2025.104123

Toupin, R., Millerand, F., & Larivière, V. (2022). Who tweets climate change papers? Investigating publics of research through users' descriptions. *PLoS One*, *17*(6), e0268999. https://doi.org/10.1371/journal.pone.0268999

Vajjala, S., & Shimangaud, S. (2025). Text classification in the LLM era—Where do we stand? (arXiv:2502.11830). *arXiv*. https://doi.org/10.48550/arXiv.2502.11830

Van Noorden, R. (2011). Science publishing: The trouble with retractions. *Nature*, *478*(7367), 26–28. https://doi.org/10.1038/478026a

Van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023—A new record. *Nature*, *624*(7992), 479–481. https://doi.org/10.1038/d41586-023-03974-8

Vidal Valero, M. (2023). Thousands of scientists are cutting back on Twitter, seeding angst and uncertainty. *Nature*, *620*(7974), 482–484. https://doi.org/10.1038/d41586-023-02554-0

Vuong, Q.-H., La, V.-P., Ho, M.-T., Vuong, T.-T., & Ho, M.-T. (2020). Characteristics of retracted articles based on retraction data from online sources through February 2019. *Science Editing*, *7*(1), 34–44. https://doi.org/10.6087/kcse.187

Wager, E. (2011). How journals can prevent, detect and respond to misconduct. *Notfall + Rettungsmedizin*, *14*(8), 613–615. https://doi.org/10.1007/s10049-011-1543-8

Walach, H., Klement, R. J., & Aukema, W. (2021). RETRACTED: The safety of COVID-19 vaccinations—We should rethink the

policy. *Vaccines*, *9*(7), 693. https://doi.org/10.3390/vaccines9070693

Wang, Z., Pang, Y., Lin, Y., & Zhu, X. (2024). Adaptable and reliable text classification using large language models (No. arXiv:2405.10523). *arXiv*. https://doi.org/10.48550/arXiv.2405.10523

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*(1), 61. https://doi.org/10.1186/1471-2288-13-61

Wu, Y., Pang, S., Guo, J., Yang, J., & Ou, R. (2024). Assessment of the efficacy of alkaline water in conjunction with conventional medication for the treatment of chronic gouty arthritis: A randomized controlled study [RETRACTED]. *Medicine*, *103*(14), e37589. https://doi.org/10.1097/MD.0000000000037589

Yu, H., Xiao, T., Xu, S., & Wang, Y. (2019). Who posts scientific tweets? An investigation into the productivity, locations, and identities of scientific tweeters. *Journal of Informetrics*, *13*(3), 841–855. https://doi.org/10.1016/j.joi.2019.08.001

Yu, H., Xu, S., Xiao, T., Hemminger, B. M., & Yang, S. (2017). Global science discussed in local altmetrics: Weibo and its comparison with Twitter. *Journal of Informetrics*, *11*(2), 466–482. https://doi.org/10.1016/j.joi.2017.02.011

Zhang, L., Gou, Z., Fang, Z., Sivertsen, G., & Huang, Y. (2023). Who tweets scientific publications? A large-scale study of tweeting audiences in all areas of research. *Journal of the Association for Information Science and Technology*, *74*(13), 1485–1497. https://doi.org/10.1002/asi.24830

Zhang, M., Wu, L., Yang, T., Zhu, B., & Liu, Y. (2024). RETRACTED: 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. https://doi.org/10.1016/j.surfin.2024.104081

Zheng, E.-T., Fu, H.-Z., Thelwall, M., & Fang, Z. (2025). Do male leading authors retract more articles than female leading authors? *Journal of Informetrics*, *19*(3), 101682. https://doi.org/10.1016/j.joi.2025.101682

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Zubiaga, A. (2024). Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence*, *6*, 1350306. https://doi.org/10.3389/frai.2023.1350306

How to cite this article: Zheng, E.-T., Fu, H.-Z., Thelwall, M., & Fang, Z. (2025). Can social media provide early warning of retraction? Evidence from critical tweets identified by human annotation and large language models. *Journal of the Association for Information Science and Technology*, 1–16. https://doi.org/10.1002/asi.70028