Citation Selectivity in Retraction Literature: A Large-Scale Bibliometric Analysis

Anonymous Author(s)

Abstract

The advancement of digital libraries and scholarly networks has enabled systematic scrutiny of scholarly work, including the identification and correction of its failures. Among these corrective mechanisms, retraction has been established as a safeguard for flagging flawed or fraudulent publications that may severely threaten the integrity of published literature and erode public trust in research findings. Existing retraction studies primarily focus on the reliability of a paper's own findings and offer little insight into how the paper engages with the literature it draws on. In this paper, we investigate the referencing behavior of retracted papers and formulate citation selectivity as a novel research problem theoretically grounded in motivated reasoning, where authors who shape evidence to align with a desired conclusion may curate reference lists that echo their own claims while under-citing equally relevant but less convenient work. To measure such selectivity, we introduce reference selectivity score to compare how closely a paper's cited references align with its content, against an independently constructed pool of expected references. We construct a large-scale dataset of 42,330 matched pairs of retracted and control papers to characterize and compare citation selectivity of retracted papers against their non-retracted counterparts. Evaluation results show that retracted papers exhibit significant higher citation selectivity than their matched controls. Further analyses reveal that the effect holds across all major disciplines, stronger for papers retracted with misconduct-related reasons, and has widened over the past two decades.

CCS Concepts

• **Information systems** → **Digital libraries and archives**; **Data mining**; • **Applied computing** → *Law, social and behavioral sciences*; *Document management and text processing*.

Keywords

Citation Analysis, Retracted Papers, Reference Bias, Bibliometrics, Comparative Analysis

ACM Reference Format:

Anonymous Author(s). 2026. Citation Selectivity in Retraction Literature: A Large-Scale Bibliometric Analysis. In *Proceedings of (JCDL'26).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

JCDL'26, Dallas, TX

ACM ISBN 978-1-4503-XXXX-X/2026/10
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The advancement of digital libraries and scholarly networks has transformed how the scientific community discovers, tracks, and reasons about scholarly work at scale [29, 36]. These infrastructures not only accelerate scientific discovery but also enable systematic scrutiny of the literature itself, including the identification and correction of its failures. Among these failures, research misconduct and irresponsible research practices, such as fabrication, falsification, and plagiarism, have become a growing concern in the scientific community [47]. Such behaviors severely threaten the integrity of published literature and erode public trust in research findings [26]. To mitigate these issues, retraction has been established as a safeguard mechanism through which authors, journals, and institutions can flag and withdraw publications found to be flawed, fraudulent, or otherwise unreliable [11, 15]. While retraction serves as an important corrective mechanism for scientific literature, it primarily addresses the reliability of a paper's own findings, and offers little insight into how the paper engages with the related literature it draws on. In this paper, we investigate the citation behavior of retracted papers and examine whether retracted papers exhibit distinct patterns of citation selectivity against comparable non-retracted work.

Existing studies of retraction have mainly focused on investigating the reasons for retraction and post-retraction influence on literature [44, 48]. In particular, meta-level bibliometric indicators, such as citation counts [24], author productivity [49], journal prestige [2], and self-citation rates [37], have been widely adopted to characterize retracted papers and to track how the scientific community continues to engage with them after retraction. While informative, these indicators are often agnostic of the paper content and ignore how an author selects, engages with, and situates their work within prior literature [30]. It remains unclear whether the same underlying behaviors that lead to a paper's retraction also leave a detectable trace in how that paper was written, more importantly, in how selectively it cited the literature relevant to it.

Our work is motivated by the phenomenon that a substantial body of work on research misconduct frames fabrication and falsification as extreme manifestations of *motivated reasoning* where researchers, whether deliberately or not, shape evidence to align with a desired conclusion [25, 31]. We hypothesize that the same tendency extends beyond the manipulation of data to the curation of a retracted paper's reference list. Under this view, an author engaged in motivated reasoning would be inclined to cite literature that corroborates their claims while under-citing equally relevant but less convenient work, such as foundational studies, adjacent findings, or results that complicate or contradict their own [13]. As such, the reference list of such a compromised publication (e.g., retracted paper) should look less like a representative sample of its intellectual neighborhood and more like an echo of the paper itself. We refer to this pattern as *citation selectivity*.

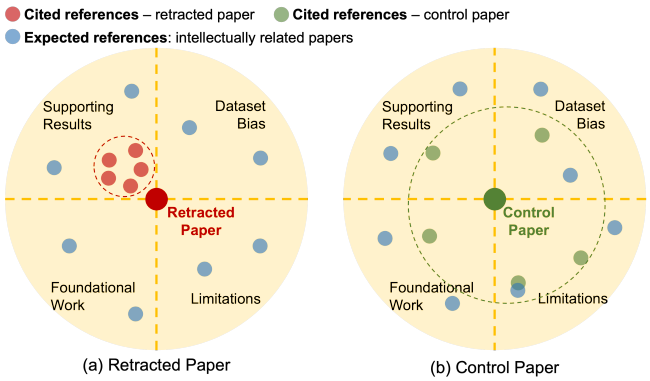


Figure 1: Overview of citation selectivity analysis

Figure 1 shows an overview of the citation selectivity analysis. Given a focal paper (e.g., retracted or control paper), we consider two sets of references: 1) *cited references*, the works a paper actually cites, and 2) *expected references*, a representative sample of the paper's broader intellectual neighborhood (e.g., foundational work, and work reporting limitations or contradictions) constructed independently of the paper's own references. As shown in Figure 1a, if a retracted paper cherry-picks its citations, its cited references should cluster closer to the paper itself than the full set of expected references, leaving foundational work and limitations underrepresented. In contrast, Figure 1b illustrates the pattern we would expect for a non-retracted control paper, whose cited references are more evenly distributed across its intellectual neighborhood.

To examine whether retracted papers exhibit citation selectivity, we introduce the reference selectivity score, which quantifies citation selectivity as the difference between how closely a paper's cited references align with its content and how closely its expected references align with its content. We construct a large-scale, matched dataset of retracted papers and non-retracted control papers with 42,330 papers in each group. Evaluation results show that retracted papers exhibit significantly higher citation selectivity than their matched controls across all three embedding models, with the effect consistent across disciplines, strongest among papers retracted with misconduct-related reasons, and widening over time. We envision that our work provides a new lens for content-aware integrity analysis in digital libraries toward understanding citation behaviors.

We summarize the key contributions of this work as follows:

- We formulate citation selectivity in retraction literature as a novel research problem grounded in motivated reasoning, extending digital libraries' role from tracking retractions to detecting selective citation behavior.
- We introduce reference selectivity score, a content-aware metric that quantifies how much more semantically aligned a paper's cited references are with the paper itself than with an independently constructed expected reference pool.
- We construct a large-scale dataset with matched retracted and control papers, to be released to support future research on scholarly integrity in the digital library community.
- Evaluation results show that retracted papers exhibit significantly higher citation selectivity than the matched controls.

2 Related Work

2.1 Research Misconduct and Retraction

Research misconduct, as defined by the Office of Science and Technology Policy (OSTP), encompasses fabrication, falsification, and plagiarism (FFP), as well as other questionable research practices that compromise the integrity of scientific literature [34, 46]. To preserve the integrity of scientific literature, publishers employ retractions to formally notify the scientific community when published findings are deemed unreliable [4, 16]. Recent data show an overall increase in retraction rates over the past two decades, highlighting the importance of understanding trends and developments in retraction research [17]. A substantial amount of work characterizes retracted papers through meta-level bibliometric analysis, including citation counts [24], author productivity [49], journal ranking [2], self-citation patterns [37], and temporal [1] and disciplinary characteristics [48]. Other studies examine the retraction process itself, documenting the limited transparency of retraction notices [38], the prevalence of misconduct among retraction causes [20], and the continued citation of papers after retraction [46]. However, existing work primarily focuses on properties of retracted papers as a whole or on post-retraction engagement. The citation behavior within a paper, namely how it selects and situates itself in prior literature, remains largely unexamined.

2.2 Motivated Reasoning in Science

Our work draws on the theory of motivated reasoning, which refers to the tendency of individuals to process and select evidence in ways that favor a desired conclusion rather than accuracy [25]. Kunda's foundational work shows that directional goals bias the beliefs and evidence that people access and treat as relevant [25]. In the scientific domain, motivated reasoning has been studied primarily on the side of evidence consumption, e.g., how prior attitudes shape the acceptance or rejection of scientific consensus [31]. On the production side, a line of work frames research misconduct as an extreme manifestation of motivated cognition. Nosek et al. [28] argue that professional incentives produce motivated reasoning whose extreme form is the fabrication or alteration of results, and empirical analyses of misconduct cases document thinking errors [12] and rationalizations [9] among researchers found to have fabricated or falsified data. However, whether motivated reasoning shapes how authors engage with related literature remains largely unexplored [27]. In this work, we introduce a measure of citation selectivity to empirically examine whether motivated reasoning leaves a detectable pattern in the reference lists of retracted papers.

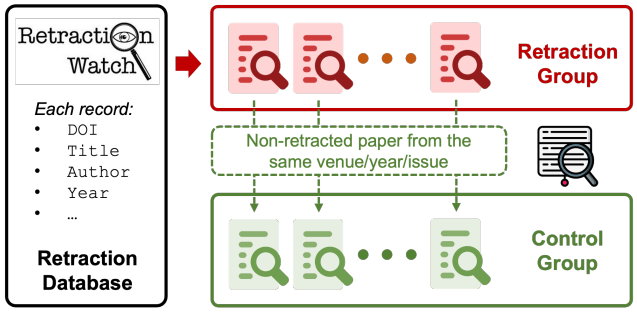
2.3 Citation Behavior and Reference Analysis

Our work also contributes to the broader study of citation behavior and biases in reference selection. Prior research has investigated citation bias, the tendency to preferentially cite studies with positive or supportive results [7, 14, 39]. For example, Greenberg [19] showed how selective citation can create unfounded authority for a claim in a biomedical citation network. Urlings et al. [42] confirmed that positive results are cited substantially more often than negative ones. Beyond bias toward outcomes, citation studies have examined self-citation [21, 22], perfunctory citation [5], and the

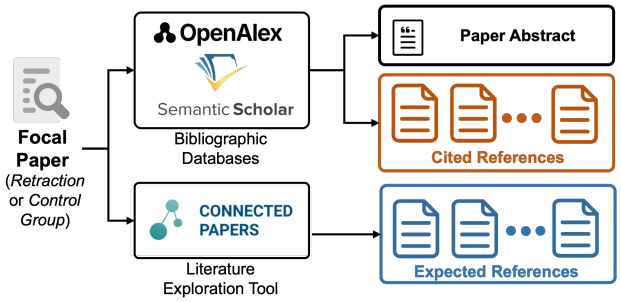
rhetorical functions that references serve [40]. However, existing studies often characterize citation bias within a single claim network or discipline, and assess bias against the outcomes of citable studies rather than against what a paper could plausibly have cited [14, 32]. In contrast, our work quantifies citation selectivity at the level of an individual paper's reference list by comparing it against an independently constructed expected reference pool, and evaluate this measure across disciplines and retraction reasons at scale.

3 Data

In this section, we describe the dataset we collected to support our analysis of citation selectivity in retracted and non-retracted papers. We first construct two groups of papers, namely a *retraction group* and a *control group*, to enable a systematic comparison of citation behavior across these two groups. For each focal paper in both groups, we retrieve its actual reference list and construct its expected reference pool. We show an overview of the data collection pipeline in Figure 2 and elaborate on each step below.



(a) Paper retrieval



(b) Reference retrieval

Figure 2: Data collection pipeline

3.1 Paper Retrieval

3.1.1 Retraction Group. As illustrated in Figure 2a, we first retrieve a list of historically retracted papers from the Retraction Watch Database [41], a comprehensive and continuously updated repository that tracks retractions across scholarly disciplines. As of 2026, the database comprises over 70,000 retraction records, each annotated with metadata such as the Digital Object Identifier (DOI), title, authors and affiliations, retraction reason and date, and original publication venue. We use this database to identify retracted papers, defined as follows.

Definition 1. **Retracted Paper:** A retracted paper is a scholarly publication that has been formally withdrawn from the scientific record by its authors, editors, or publisher, typically due to research misconduct, data fabrication, plagiarism, or honest error [38]. We denote a retracted paper as p^r. The retraction group $\mathcal{R} = \{p_1^r, p_2^r, \ldots, p_N^r\}$ is the set of all valid retracted papers in our dataset, where N is the total number of retracted papers.

3.1.2 Control Group. To enable a fair comparison, we construct a matched control group by sampling one non-retracted paper for each retracted paper from the same venue (e.g., journal or conference proceeding), publication year, and discipline/field. This matching strategy controls for field-specific and temporal citation norms that may otherwise confound the comparison [1]. Formally, we define a control paper as follows.

Definition 2. **Control Paper:** A control paper refers to a non-retracted scholarly publication matched to a retracted paper p^r based on publication year, discipline, and venue, serving as a baseline for comparing citation selectivity. We denote a control paper as p^c. The control group $C = \{p_1^c, p_2^c, \ldots, p_N^c\}$ is the set of all control papers in our dataset, where N is the total number of control papers.

The retraction group and control group form the complete set of focal papers $\mathcal{F} = \mathcal{R} \cup C$, comprising $2N$ papers in total. For each focal paper $p \in \mathcal{F}$, we subsequently retrieve its actual reference list and construct its expected reference pool, as described in the following subsection.

3.2 Reference Retrieval

For each focal paper $p \in \mathcal{F}$, we retrieve two types of references: 1) the *cited references* actually cited in the paper, and 2) the *expected references* representing the broader intellectual neighborhood of the paper. As shown in Figure 2b, the reference retrieval process consists of two steps, described as follows.

3.2.1 Cited Reference. We retrieve the cited reference (i.e., the works actually cited in the paper) of each focal paper $p \in \mathcal{F}$ using the Semantic Scholar Academic Graph API [23][1], an open academic knowledge graph that provides structured metadata and reference lists for over 200 million scholarly works. For each focal paper, we query the API using its DOI or title, and extract the full list of cited references. In cases where Semantic Scholar fails to retrieve a reference list, we complement the retrieval using the OpenAlex API [33][2], an open bibliographic catalogue that indexes over 250 million scholarly works. We denote the actual reference list of a focal paper p as $\mathcal{A}(p) = \{a_1, a_2, \ldots, a_{k_p}\}$, where k_p is the number of references cited by p. Papers for which a reference list could not be retrieved from either source are excluded from $\mathcal{F}$.

3.2.2 Expected Reference. To construct the expected reference pool for each focal paper, we leverage a literature exploration tool to identify works that are *intellectually* related to the focal paper but may not have been directly cited by it. In particular, we use Connected Papers[3], a literature exploration tool that identifies intellectually

[1] https://www.semanticscholar.org/product/api
[2] https://developers.openalex.org/
[3] https://www.connectedpapers.com

related works based on a graph-based similarity measurement algorithm. Unlike a citation tree, Connected Papers surfaces works that share an intellectual context with the focal paper without requiring a direct citation relationship, which makes it suitable for identifying the expected reference pool against which actual citation behavior is compared. For each focal paper $p \in \mathcal{F}$, we retrieve the set of papers returned by Connected Papers and denote the expected reference pool as $\mathcal{E}(p) = \{e_1, e_2, \ldots, e_{m_p}\}$, where m_p is the size of the expected reference pool.

3.3 Dataset Summary

3.3.1 Preprocessing. Starting from retraction records in the Retraction Watch Database as of May 20, 2026, we apply a series of filtering steps to obtain the final dataset. We first remove duplicated entries and exclude papers without a retrievable DOI or without an index entry in either Semantic Scholar or OpenAlex. We further remove papers with invalid abstracts, including non-English abstracts, abstracts fewer than 50 characters in length, and abstracts containing only retraction notice instead of the original paper abstract. Additionally, we exclude papers for which a reference list could not be retrieved ($k_p = 0$) or for which Connected Papers returned an empty expected reference pool ($m_p = 0$). After filtering, the retraction group $\mathcal{R}$ and the control group C comprise the same number of 42,330 focal papers.

3.3.2 Summary. We summarize the statistics of the preprocessed dataset in Table 1. On average, each focal retracted paper has 34.63 cited references and 39.80 expected references in its pool, with an average overlap of $|\mathcal{A}(p) \cap \mathcal{E}(p)| = 2.77$ papers between the two sets of references. The control group shows a comparable profile with an average of 48.74 cited references, 39.90 expected references, and an overlap of 3.37, which suggests that the expected reference pool is largely disjoint from what papers actually cite and thus provides an independent baseline against which citation behavior can be assessed. Notably, retracted papers cite fewer references on average than controls (34.63 vs. 48.74) and we examine this as a potential confounder in Section 5.6. Additionally, we report the *self-citation rate*, defined as the fraction of a focal paper's cited references that share at least one author with the focal paper itself. On average, retracted papers exhibit a substantially higher self-citation rate than control papers (3.5% vs. 0.4%).

Figure 3 summarizes the most common stated reasons for retraction across the dataset, with "investigation by journal/publisher", "concerns about peer review", and "unreliable results or conclusions" among the most frequently cited categories. Figure 4 further shows the distribution of the retraction group across academic disciplines, where "Medicine", "Biology", and "Technology" account for the largest number of retracted papers in the dataset, consistent with prior reports on the disciplinary distribution of retractions [48]. Figure 5 shows the distribution of retracted papers by publication year and retraction year. Both curves remain low until the early 2000s and grow sharply thereafter, consistent with the reported acceleration of retractions over the past two decades [17]. The retraction-year curve trails the publication-year curve, reflecting the delay between publication and retraction. Note that the drop at the right end is an artifact of incomplete data for recent years rather than an actual decline.

Table 1: Dataset Summary

	Retraction	Control
Number of papers	42,330	42,330
Publication year range	1940-2026	1940-2026
Number of disciplines	130	–*
Avg. cited reference count (k_p)	34.63	48.74
Avg. expected reference count (m_p)	39.80	39.90
Avg. overlap $\lvert\mathcal{A}(p) \cap \mathcal{E}(p)\rvert$	2.77	3.37
Avg. self-citation rate	3.5%	0.4%

* Discipline annotations are unavailable for control papers, as these labels originate from the Retraction Watch Database and are assigned only to retracted publications.

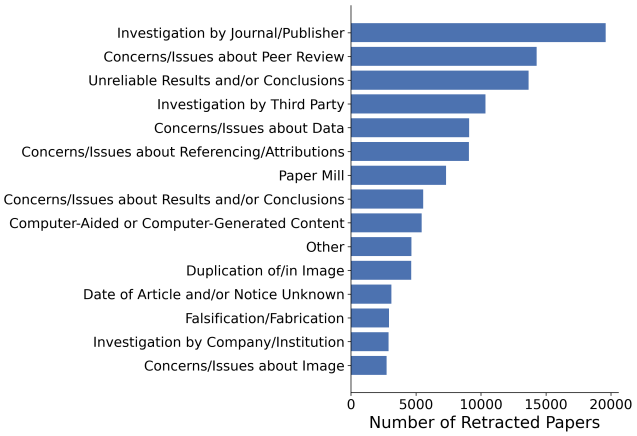


Figure 3: Distribution of retracted papers across the top 15 reasons for retraction. Papers may cite multiple reasons for retraction.

4 Methodology

In this section, we describe our methodology for measuring and comparing citation selectivity among papers in the retraction group and control group. An overview of the proposed methodology is shown in Figure 6. We first encode all focal papers and their actual and expected references into latent vector representation. We then compute the pairwise semantic similarity between each focal paper and its cited references and expected references. Based on the average similarity scores, we derive the reference selectivity score for each focal paper, which quantifies the degree to which a focal paper's cited references are more semantically aligned with itself than its expected references. Finally, we compare the reference selectivity scores across the retraction group and control group and assess whether retracted papers exhibit systematically higher citation selectivity than their matched controls.

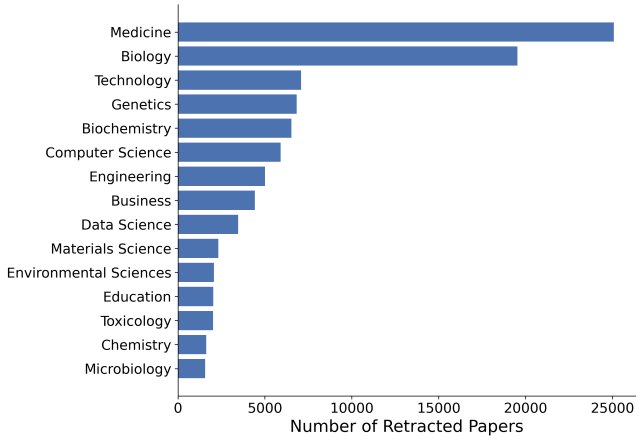


Figure 4: Distribution of retracted papers across the top 15 disciplines by retraction count. Papers may be classified under multiple disciplines.

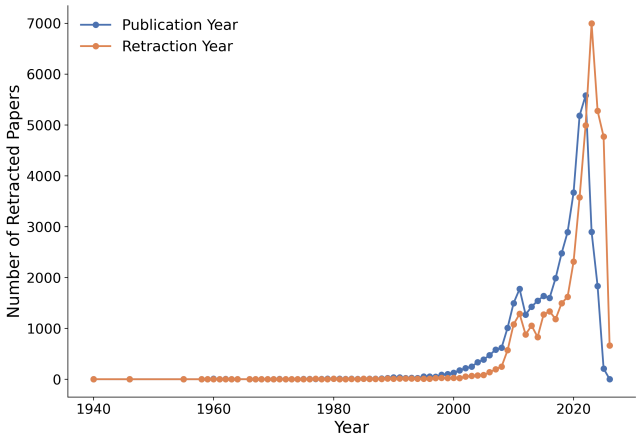


Figure 5: Distribution of retracted papers by publication and retraction year.

4.1 Latent Paper Representation

To measure the semantic similarity between a focal paper and its references, we represent each paper as a dense vector using a pre-trained language model (PLM). A language model is build on the Transformer architecture [43] and pre-trained on large corpora of text to produce dense vector representations that capture the semantic meaning of the input text. In particular, we consider the following three state-of-the-art pre-trained models to learn the latent representation of an input paper (i.e., focal paper or reference).

- **SciBERT [3]**: SciBERT is a BERT-based [10] language model pre-trained on a large corpus of scientific text spanning multiple disciplines to capture domain-specific semantic relationships between scientific concepts. We use the `[CLS]` token embedding of the abstract as the latent representation of an input paper.
- **MiniLM [45]**: MiniLM is a lightweight and computationally efficient language model distilled from a larger pre-trained Transformer-based model to produce high-quality text embeddings. We use the average pooling of the token embeddings of the abstract to encode an input paper.
- **Qwen3-Embedding [50]**: Qwen3-Embedding is a decoder-only language model trained via large-scale contrastive learning to produce general-purpose text embeddings. We use the last-token embedding of the abstract as the latent representation of an input paper.

All three models are used to encode all focal papers and their references to minimize bias introduced by the choice of language model. A detailed comparison of results across both models is presented in Section 5. In particular, we denote the latent representation of a retracted paper $p^r \in \mathcal{R}$ and a control paper $p^c \in C$ as $\vec{p}^r \in \mathbb{R}^d$ and $\vec{p}^c \in \mathbb{R}^d$, respectively, where d is the embedding dimension of the chosen model (i.e., $d = 768$ for SciBERT, $d = 384$ for MiniLM, $d = 1024$ for Qwen-Embedding). Similarly, the latent representations of a cited reference $a_i \in \mathcal{A}(p)$ and an expected reference $e_j \in \mathcal{E}(p)$ are denoted as $\vec{a}_i \in \mathbb{R}^d$ and $\vec{e}_j \in \mathbb{R}^d$, respectively.

4.2 Reference Similarity

Given the latent representation $\vec{p}$ of a focal paper $p \in \mathcal{F}$, we compute its average similarity to its cited references $\mathcal{A}(p)$ and expected references $\mathcal{E}(p)$. In particular, we adopt the cosine similarity as the similarity measure between a focal paper p and a reference paper $r_i \in \mathcal{A}(p) \cup \mathcal{E}(p)$. Formally, the reference similarity is computed as:

$$\cos(\vec{p}, \vec{r}_i) = \frac{\vec{p} \cdot \vec{r}_i}{||\vec{p}|| \cdot ||\vec{r}_i||} \tag{1}$$

The *actual* reference similarity between a focal paper $p \in \mathcal{F}$ and its cited references $\mathcal{A}(p)$ is then computed as the mean cosine similarity across all k_p cited references:

$$s_{\mathcal{A}}(p) = \frac{1}{k_p} \sum_{i=1}^{k_p} \cos(\vec{p}, \vec{a}_i) \tag{2}$$

where $k_p = |\mathcal{A}(p)|$ is the cited reference count of focal paper p.

Similarly, the *expected* reference similarity between a focal paper p and its expected references $\mathcal{E}(p)$ is computed as the mean cosine similarity across all m_p expected references:

$$s_{\mathcal{E}}(p) = \frac{1}{m_p} \sum_{j=1}^{m_p} \cos(\vec{p}, \vec{e}_j) \tag{3}$$

where $m_p = |\mathcal{E}(p)|$ is the size of the expected reference pool of focal paper p.

4.3 Citation Selectivity Measurement

Building on the actual and expected reference similarities defined in Equations 2 and 3, we now define the reference selectivity score $\delta(p)$ for each focal paper $p \in \mathcal{F}$. Intuitively, if a focal paper cherry-picks its references, its cited references should be more semantically similar to itself than its expected references, resulting in a higher $s_{\mathcal{A}}(p)$ relative to $s_{\mathcal{E}}(p)$. We formally define the reference selectivity score as follows.

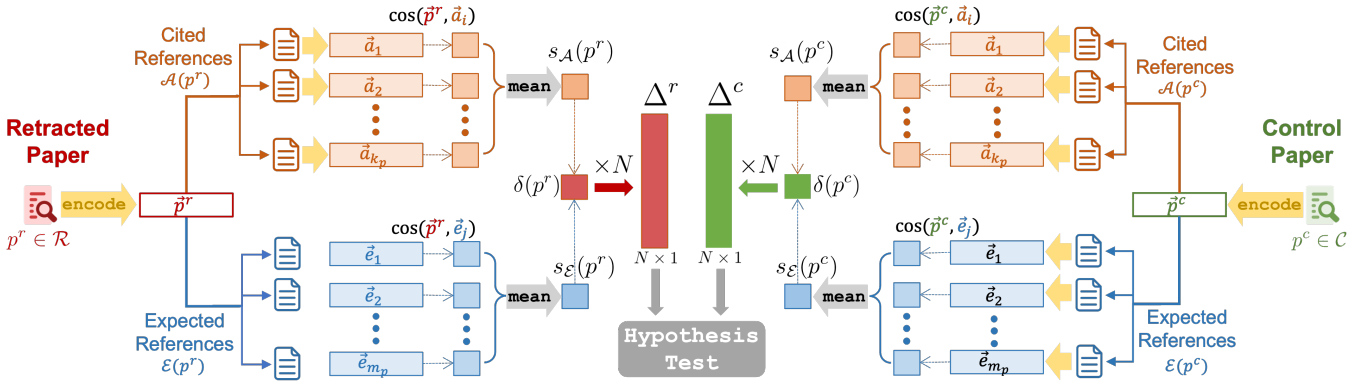


Figure 6: Overview of methodology

Definition 3. **Reference Selectivity Score:** The reference selectivity score $\delta(p)$ of a focal paper $p \in \mathcal{F}$ is the difference between its cited reference similarity $s_{\mathcal{A}}(p)$ and its expected reference similarity $s_{\mathcal{E}}(p)$:

$$\delta(p) = s_{\mathcal{A}}(p) - s_{\mathcal{E}}(p) \tag{4}$$

A positive $\delta(p)$ indicates that the focal paper's cited references are on average more semantically similar to itself than its expected references, suggesting selective citation behavior. A negative $\delta(p)$ indicates the opposite, i.e., the cited references are less topically aligned with the focal paper than the expected reference pool.

For each retracted paper $p^r \in \mathcal{R}$ and its matched control paper $p^c \in C$, we compute $\delta(p^r)$ and $\delta(p^c)$, yielding two sets of reference selectivity scores:

$$\Delta^r = \{\delta(p^r) \mid p^r \in \mathcal{R}\} \tag{5}$$

$$\Delta^c = \{\delta(p^c) \mid p^c \in C\} \tag{6}$$

where Δ^r and Δ^c are the sets of reference selectivity scores for the retraction group and control group, respectively. These scores form the basis for the statistical comparison described in the following subsection.

4.4 Statistical Testing

To determine whether retracted papers exhibit systematically higher citation selectivity than their matched control papers, we perform a paired t-test [35] on the reference selectivity scores Δ^r and Δ^c. The paired t-test is appropriate here as each retracted paper $p^r \in \mathcal{R}$ is matched to exactly one control paper $p^c \in C$ by venue and publication year, forming N matched pairs $\{(p_i^r, p_i^c)\}_{i=1}^N$. For each matched pair, we compute the difference in reference selectivity scores:

$$\Delta_i = \delta(p_i^r) - \delta(p_i^c), \quad i = 1, 2, \ldots, N \tag{7}$$

The paired t-test then assesses whether the mean difference $\bar{\Delta}$ is significantly greater than zero. Formally, we test the following one-tailed hypotheses:

$$H_0 : \mu_\Delta = 0 \tag{8}$$

$$H_1 : \mu_\Delta > 0 \tag{9}$$

where μ_Δ is the population mean of Δ_i. H_0 states that there is no difference in citation selectivity between retracted and control papers, while H_1 states that retracted papers exhibit higher citation selectivity than their matched controls, consistent with cherry-picking behavior.

The t-statistic is computed as:

$$t = \frac{\bar{\Delta}}{s_\Delta / \sqrt{N}} \tag{10}$$

where $\bar{\Delta} = \frac{1}{N}\sum_{i=1}^N \Delta_i$ is the mean difference in reference selectivity scores, s_Δ is the standard deviation of Δ_i, and N is the number of matched pairs. We reject H_0 in favor of H_1 if the p-value is below a significance threshold of $\alpha = 0.05$. A statistically significant result would indicate that retracted papers exhibit systematically higher citation selectivity than their matched controls, providing computational evidence of selective citation behavior in retracted literature.

5 Results

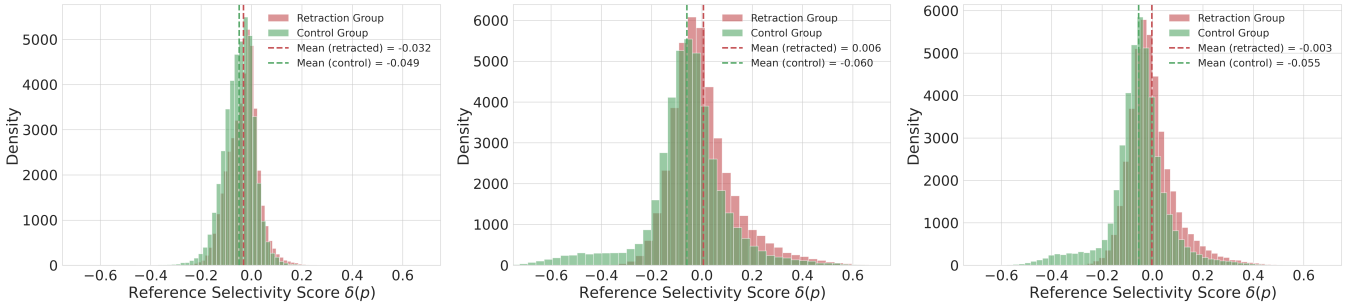
We conduct extensive experiments to study citation selectivity in retraction literature. We organize our analysis around five research questions, addressed in Sections 5.2 to 5.6, respectively:

- **RQ1**: Do retracted papers exhibit systematically higher citation selectivity than their matched non-retracted control papers?
- **RQ2**: Does the degree of citation selectivity vary with the stated retraction reason?
- **RQ3**: Does the degree of citation selectivity in retracted papers vary across academic disciplines?
- **RQ4**: Has the degree of citation selectivity in retracted papers changed over time?
- **RQ5**: Is the observed citation selectivity robust to potential confounding factors (i.e., self-citation rate, cited reference count)?

5.1 Experimental Setup

Our experiments[4] are conducted on the preprocessed dataset described in Section 3, including 42,330 retracted papers and the same

[4]Our code and dataset (with papers identified by DOI, in accordance with source licensing terms) will be made publicly available upon acceptance of the paper.



(a) SciBERT **(b) MiniLM** **(c) Qwen3-Embedding**

Figure 7: Distribution of reference selectivity score $\delta(p)$.

number of matched papers in the control group. We encode the abstract of each focal paper, along with those of its cited and expected references, into latent representations using the three pre-trained language models described in Section 4.1. Specifically, we use the open-source checkpoints `allenai/scibert_scivocab_uncased`, `sentence-transformers/all-MiniLM-L6-v2`, and `Qwen/Qwen3-Embedding-0.6B` for SciBERT, MiniLM, and Qwen3-Embedding, respectively. Abstracts are truncated to each model's maximum input length, i.e., 512, 256, and 32,768 tokens for SciBERT, MiniLM, and Qwen3-Embedding, respectively. The output embedding dimension for SciBERT, MiniLM, and Qwen3-Embedding are 768, 384, and 1024, respectively. All experiments are implemented in Python using PyTorch v2.9.0, and executed on Ubuntu 24.04.1 with four Nvidia 6000 Ada GPUs.

5.2 Reference Selectivity (RQ1)

We first study whether retracted papers exhibit systematically higher reference selectivity than their matched non-retracted control papers. We compute the reference selectivity scores $\delta(p^r)$ and $\delta(p^c)$ for every retracted focal paper $p^r \in \mathcal{R}$ and their matched control paper $p^c \in C$. We then perform the one-tailed paired t-test described in Section 4.4 over all the matched pairs, with significance threshold $\alpha = 0.05$. We additionally report Cohen's d [18] as a measure of effect size. Table 2 reports the reference selectivity score statistics for the retraction group $\mathcal{R}$ and the control group C across the three embedding models. We observe that the retraction group exhibits a significantly higher mean reference selectivity score than the control group across all three embedding models ($p < 0.001$), with small-to-moderate effect sizes, i.e., Cohen's $d = 0.226$, 0.339, and 0.373 for SciBERT, MiniLM, and Qwen3, respectively. The gap between group means is smallest under SciBERT (−0.032 vs. −0.049) and larger under MiniLM (0.006 vs. −0.060) and Qwen3 (−0.003 vs. −0.055). Recall from Definition 3 that a higher $\delta(p)$, whether positive or less negative, indicates that a paper's cited references align more closely with the paper itself relative to its expected reference pool. Thus, the reference lists of retracted papers consistently echo the paper itself more closely than those of their matched controls. Figure 7 shows the distribution of $\delta(p)$ for both groups across the three embedding models. While the two distributions overlap substantially, the retraction group's distribution is consistently shifted to the right of the control group's and consequently reflects stronger citation selectivity in the retraction group.

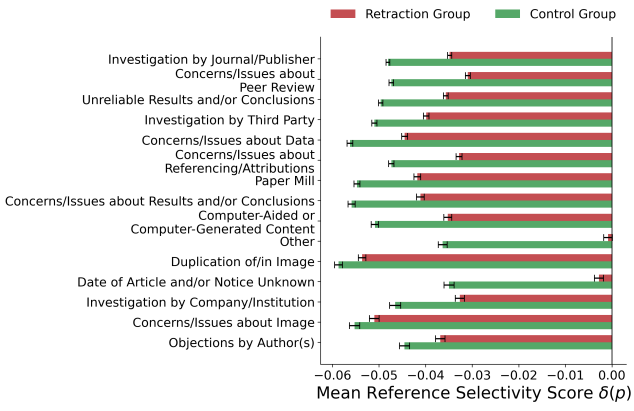
Table 2: Reference selectivity score $\delta(p)$ comparison between the retraction group $\mathcal{R}$ and control group C. * denotes $p < 0.001$.**

Model	Metric	Retraction	Control	p-value
SciBERT	Mean $\delta(p)$	-0.032	-0.049	<0.001***
	Std $\delta(p)$	0.060	0.060	–
	Median $\delta(p)$	-0.026	-0.044	–
	Cohen's d	0.226		–
MiniLM	Mean $\delta(p)$	0.006	-0.060	<0.001***
	Std $\delta(p)$	0.131	0.162	–
	Median $\delta(p)$	-0.016	-0.054	–
	Cohen's d	0.339		–
Qwen3	Mean $\delta(p)$	-0.003	-0.055	<0.001***
	Std $\delta(p)$	0.096	0.115	–
	Median $\delta(p)$	-0.018	-0.050	–
	Cohen's d	0.373		–

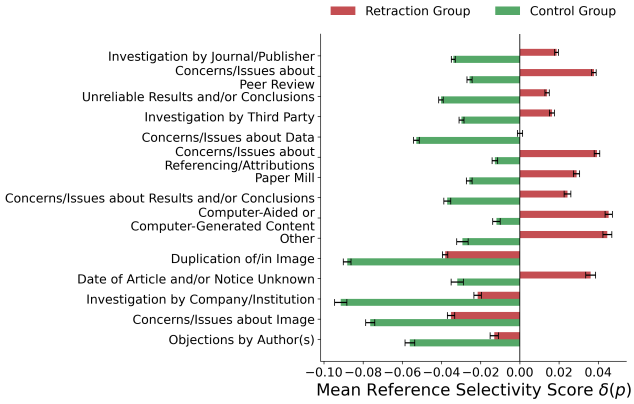
5.3 Retraction Reason Analysis (RQ2)

Next, we investigate whether reference selectivity varies with the stated reason for retraction. Since fabrication and falsification are commonly framed as extreme manifestations of motivated reasoning (Section 1), we would expect papers retracted for misconduct-related reasons to exhibit stronger citation selectivity than papers retracted for honest error or administrative reasons. Following the reason taxonomy of the Retraction Watch Database, we compute the mean reference selectivity score $\delta(p)$ within each reason category and compare it against the corresponding matched controls.

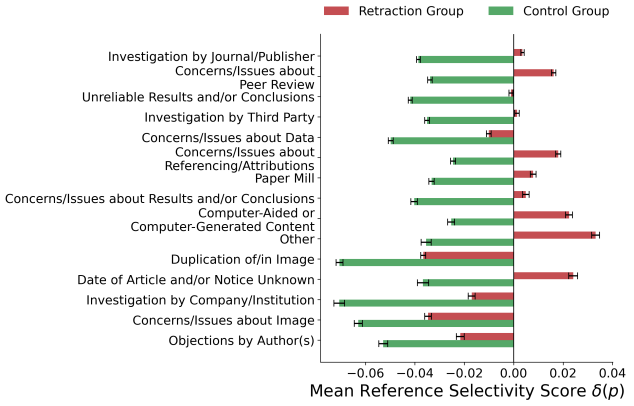
Figure 8 presents the results for the top-15 most common retraction reasons in the dataset (a paper may cite multiple reasons). We observe that the retraction group exhibits a higher mean $\delta(p)$ than the corresponding controls across nearly all of the top-15 retraction reasons and all three embedding models. Consistent with the motivated reasoning account, the gap is most pronounced for misconduct-related categories such as "paper mill" and "computer-aided or computer-generated content", where the retraction group's



(a) SciBERT



(b) MiniLM



(c) Qwen3-Embedding

Figure 8: Mean reference selectivity score $\delta(p)$ by retraction reason for the retraction group $\mathcal{R}$ and control group C.

mean $\delta(p)$ is even positive under MiniLM and Qwen3-Embedding, indicating strongly self-echoing reference lists. In contrast, the gap is smaller for categories more associated with administrative or procedural causes, such as "objections by author(s)" and "investigation by company/institution".

5.4 Cross-Discipline Comparison (RQ3)

We further examine whether the degree of reference selectivity in retracted papers varies across academic disciplines. Using the discipline annotations from the Retraction Watch Database, we group retracted papers by discipline and compare the mean reference selectivity score δp of each discipline's retracted papers against their matched controls. As a paper may be classified under multiple disciplines, it contributes to every discipline it was annotated with. Table 3 reports the mean reference selectivity scores for the top-15 most represented disciplines in our dataset, along with the p-value of the paired t-test within each discipline. The retraction group exhibits a significantly higher mean $\delta(p)$ than its matched controls in all 15 disciplines across all three embedding models ($p < 0.001$), indicating that citation selectivity in retracted papers is a cross-disciplinary phenomenon rather than an artifact of a few retraction-heavy fields. However, the magnitude of the gap varies across disciplines. It is largest in technology- and application-oriented fields (e.g., Technology, Computer Science, Engineering) where the retraction group's mean $\delta(p)$ even turns positive under MiniLM and Qwen3-Embedding. In contrast, the gap is smaller in biomedical disciplines (e.g., Medicine, Biology, Genetics), where both groups remain below the expected-reference baseline.

5.5 Temporal Analysis (RQ4)

We also examine whether the degree of reference selectivity in retracted papers has changed over time. For each publication year, we compute the mean reference selectivity score δp of retracted papers and control papers published in that year. Papers published before 1980 were dropped from the analysis due to the insufficient number of retracted papers per year, which yields unstable estimates. Figure 9 plots the mean reference selectivity score δp over time for both groups, with shaded bands denoting ±1 standard error. We observe that the retraction group maintains a consistently higher mean $\delta(p)$ than the control group across nearly the entire studied period under all three embedding models. The gap is modest and noisier before 2000, where wider standard-error bands reflect the small number of retracted papers per year (Figure 5) and becomes clearly pronounced after 2000. Notably, the widening gap is driven primarily by a decline in the control group's mean $\delta(p)$, while the retraction group remains comparatively stable. This post-2000 divergence coincides with the sharp growth of retractions shown in Figure 5 and the documented rise of large-scale misconduct, such as paper mills [8] and computer-aided or computer-generated content [6], consistent with our reason-level analysis in Section 5.3.

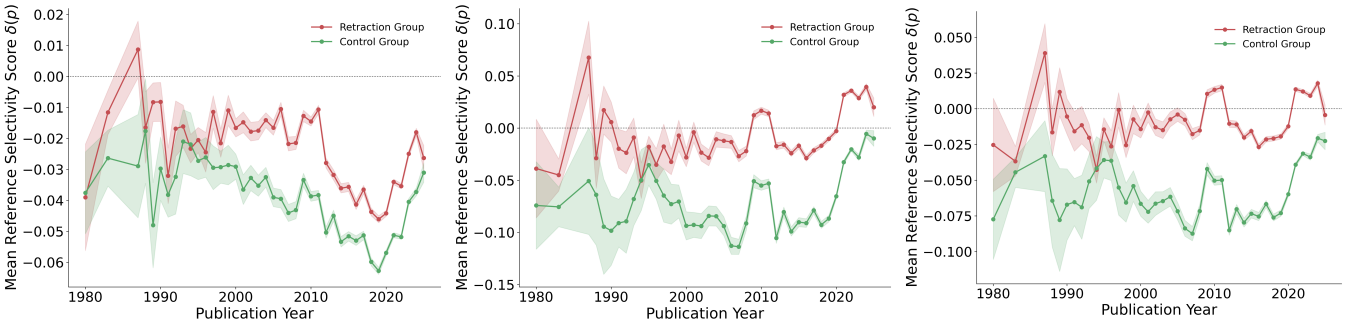
5.6 Robustness Analysis (RQ5)

Finally, we examine whether the observed difference in reference selectivity between the retraction group and the control group can be attributed to confounding factors rather than selective citation behavior. We consider two potential confounders suggested by the dataset statistics in Table 1, including self-citation rate and cited reference count.

5.6.1 Self-Citation Rate. As shown in Table 1, retracted papers exhibit a substantially higher average self-citation rate (defined in Section 3.3.2) than control papers (3.5% vs. 0.4%). As self-citations

Table 3: Mean reference selectivity score $\delta(p)$ by discipline. * denotes $p < 0.001$.**

	SciBERT			MiniLM			Qwen3-Embedding		
Discipline	**Retraction**	**Control**	***p*-value**	**Retraction**	**Control**	***p*-value**	**Retraction**	**Control**	***p*-value**
Medicine	-0.034	-0.050	<0.001***	-0.008	-0.085	<0.001***	-0.014	-0.068	<0.001***
Biology	-0.046	-0.057	<0.001***	-0.031	-0.098	<0.001***	-0.030	-0.076	<0.001***
Technology	-0.026	-0.044	<0.001***	0.048	-0.012	<0.001***	0.025	-0.025	<0.001***
Genetics	-0.051	-0.060	<0.001***	-0.032	-0.088	<0.001***	-0.034	-0.071	<0.001***
Biochemistry	-0.044	-0.056	<0.001***	-0.026	-0.096	<0.001***	-0.023	-0.075	<0.001***
Computer Science	-0.022	-0.042	<0.001***	0.048	-0.014	<0.001***	0.028	-0.026	<0.001***
Engineering	-0.019	-0.042	<0.001***	0.026	-0.031	<0.001***	0.017	-0.042	<0.001***
Business	-0.020	-0.046	<0.001***	0.042	-0.018	<0.001***	0.019	-0.031	<0.001***
Data Science	-0.029	-0.040	<0.001***	0.028	-0.006	<0.001***	0.010	-0.020	<0.001***
Materials Science	-0.020	-0.038	<0.001***	0.013	-0.036	<0.001***	0.008	-0.046	<0.001***
Environmental Sciences	-0.023	-0.045	<0.001***	0.012	-0.039	<0.001***	-0.003	-0.044	<0.001***
Education	-0.034	-0.048	<0.001***	0.055	-0.007	<0.001***	0.025	-0.024	<0.001***
Toxicology	-0.044	-0.052	<0.001***	-0.022	-0.086	<0.001***	-0.024	-0.069	<0.001***
Chemistry	-0.029	-0.047	<0.001***	-0.029	-0.081	<0.001***	-0.015	-0.069	<0.001***
Microbiology	-0.045	-0.055	<0.001***	-0.035	-0.089	<0.001***	-0.036	-0.072	<0.001***



(a) SciBERT **(b) MiniLM** **(c) Qwen3-Embedding**

Figure 9: Mean reference selectivity score $\delta(p)$ over publication year for the retraction group $\mathcal{R}$ and control group C. Shaded bands denote ±1 standard error. Papers published before 1980 were dropped from the analysis due to the insufficient number of retracted papers per year, which yields unstable estimates.

Table 4: Pearson correlation between the reference selectivity score $\delta(p)$ and potential confounding factors (self-citation rate and cited reference count) across three embedding models for the retraction group $\mathcal{R}$ and control group C. All correlations are significant at $p < 0.001$ due to the large sample size (N = 42,330 per group). We therefore focus on magnitudes.

	Self-Citation Rate		Cited Reference Count	
	Retraction	Control	Retraction	Control
SciBERT	0.144	0.055	-0.153	-0.122
MiniLM	-0.034	0.060	-0.243	-0.206
Qwen3-Embedding	-0.025	0.051	-0.251	-0.195

are, by construction, likely to be semantically close to the citing paper, a natural concern is that the higher reference selectivity score of retracted papers merely reflects their propensity to cite their own work. To examine this, we compute the Pearson correlation between the self-citation rate and the reference selectivity score $\delta(p)$ within each group separately, to rule out spurious correlation induced by group-level differences. As reported in Table 4, the within-group correlations are weak across all three embedding models ($|r| \leq 0.144$) and inconsistent in sign for the retraction group, indicating that self-citation explains little of the variance in $\delta(p)$ and cannot account for the observed group difference.

5.6.2 Cited Reference Count. Retracted papers also cite fewer references on average than control papers ($k_p = 36.56$ vs. 52.90). A shorter reference list could mechanically inflate $\delta(p)$, as a smaller set of references may be more topically concentrated around the focal paper. As shown in Table 4, $\delta(p)$ is indeed modestly and negatively correlated with k_p. However, the correlations are of comparable magnitude in both groups (−0.12 to −0.25), indicating a shared mechanical tendency rather than a factor that differentiates the two groups.

6 Discussion

6.1 Limitations and Future Work

While this work introduces a novel, theoretically grounded lens for examining citation behavior in retraction literature and demonstrates its utility at scale, we acknowledge several limitations and suggest directions for future work.

6.1.1 Dataset Coverage. Our dataset includes only retracted papers with a retrievable DOI and a valid English abstract indexed in digital bibliographic databases (e.g., Semantic Scholar, OpenAlex). This excludes older publications, non-English literature, and papers whose records were removed or degraded after retraction, which may introduce coverage bias toward recent, well-indexed venues. Consequently, our findings may not generalize to the full population of retracted papers. Future work could extend our analysis to multilingual retraction literature and incorporate additional data sources, such as publisher records and digital preservation archives, to recover papers underrepresented in current bibliographic databases.

6.1.2 Retraction as a Noisy Label. In this study, we treat retraction as a proxy for compromised research by following common practice in retraction studies [44, 48]. However, retraction encompasses a wide range of causes from deliberate fabrication to honest error, administrative issues, or publisher disputes, and only a subset involves motivated reasoning and compromised research. More importantly, such heterogeneity of retraction annotation may attenuate the observed group difference in our analysis, leading to a conservative estimate of the effect among genuinely compromised papers. Our reason-level analysis (Section 5.3) partially addresses this heterogeneity by examining selectivity within individual retraction reasons. Future work could further refine the analysis by cross-referencing retraction records with institutional investigation reports (e.g., ORI case files [9]) to isolate the subpopulation where motivated reasoning is most plausible.

6.1.3 Abstract-Level Paper Representation. We encode each paper using its abstract rather than its full text, due to the input length limits of pre-trained language models and the limited availability of open-access full texts at scale. Abstracts capture a paper's central claims but may omit the local context in which each reference is actually cited (e.g., the citing sentence or paragraph). Our measure therefore reflects topical alignment at the document level rather than the rhetorical function of individual citations. We plan to further explore the opportunities to incorporate citation contexts from full texts in our future work.

6.1.4 Correlation is not Causation. Our analysis is observational in the sense that a systematic difference in citation selectivity between retracted and control papers does not establish that motivated reasoning causes selective citation, nor that selectivity contributes to retraction. Unobserved confounders, such as field-specific citation norms, author seniority, or paper-mill authorship patterns, may drive both. Therefore, additional causality analysis, such as controlling for author-level and venue-level covariates, is a valuable direction for future work to disentangle selective citation from confounding factors.

6.2 Ethical Considerations

This work analyzes citation behavior at the population level to advance the understanding of research integrity, and we emphasize several ethical considerations in interpreting and applying our findings.

6.2.1 No Individual-Level Accusations. The reference selectivity score is a statistical signal that distinguishes groups of papers on average. Given the modest effect size and the substantial overlap between the two distributions, it cannot and should not be used to infer misconduct for any individual paper or author. A high score may arise from entirely legitimate factors, such as working in a narrow subfield, writing a focused technical contribution, or following discipline-specific citation norms. We therefore caution against using our measure, or derivatives of it, as an automated screening or accusation tool.

6.2.2 Data Privacy and Licensing. Our analysis relies exclusively on publicly available bibliographic metadata and abstracts from the Retraction Watch Database and online bibliographic databases (e.g., Semantic Scholar, OpenAlex), and involves no private or personally sensitive information beyond what is already part of the public scholarly record. We use and will release the data in accordance with the licensing terms of these sources. In particular, our released dataset will identify papers by DOI rather than redistributing copyrighted content.

6.2.3 Responsible Use and Potential Misuse. We envision citation selectivity as one complementary signal among many that could assist, rather than replace, human judgment in editorial and integrity review, and any deployment in digital library infrastructure should involve transparency about the measure's limitations, human oversight, and due process for affected authors. Meanwhile, we acknowledge that public release of paper-level selectivity scores could enable harassment of authors of retracted papers, many of whom have committed no misconduct. To mitigate this risk, we will release aggregate statistics and code for reproducibility, and will consider gating paper-level scores behind a research-use agreement.

7 Conclusion

In this paper, we investigate the problem of citation selectivity in retraction literature. We formulate citation selectivity as a novel research problem grounded in motivated reasoning and introduce the reference selectivity score to explicitly quantify how much more semantically aligned a paper's cited references are with the paper itself than with an expected reference pool. Extensive experiment results on a large-scale dataset of 42,330 matched pairs of retracted and control papers show that retracted papers exhibit significantly higher citation selectivity than their matched controls across all three embedding models. urther analyses demonstrate that the effect is consistent across disciplines, most pronounced for misconduct-related retraction reasons such as paper mills and computer-generated content, has widened since 2000, and is robust to self-citation and reference count as potential confounders. We envision that this work opens a new direction for content-aware integrity analysis in digital libraries and scholarly networks toward understanding citation behaviors.

References

[1] Judit Bar-Ilan and Gali Halevi. 2018. Temporal characteristics of retracted articles. *Scientometrics* 116, 3 (2018), 1771–1783.

[2] Prashanta Kumar Behera, Sanmati Jinendran Jain, and Ashok Kumar. 2024. Examining retraction counts to evaluate journal quality in psychology. *Current Psychology* 43, 26 (2024), 22436–22443.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 3615–3620.

[4] Mark J. Bolland, Alison Avenell, and Andrew Grey. 2025. Publication integrity: what is it, why does it matter, how it is safeguarded and how could we do better? *Journal of the Royal Society of New Zealand* 55, 2 (2025), 267–286. doi:10.1080/03036758.2024.2325004

[5] Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of documentation* 64, 1 (2008), 45–80.

[6] Guillaume Cabanac and Cyril Labbé. 2021. Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology* 72, 12 (2021), 1461–1476.

[7] María-del-Mar Camacho-Miñano and Manuel Núñez-Nickel. 2009. The multilayered nature of reference selection. *Journal of the American Society for Information Science and Technology* 60, 4 (2009), 754–777.

[8] Cristina Candal-Pedreira, Joseph S Ross, Alberto Ruano-Ravina, David S Egilman, Esteve Fernández, and Mónica Pérez-Ríos. 2022. Retracted papers originating from paper mills: cross sectional study. *bmj* 379 (2022).

[9] Mark S. Davis, Michelle Riske-Morris, and Sebastian R. Diaz. 2007. Causal Factors Implicated in Research Misconduct: Evidence from ORI Case Files. *Science and Engineering Ethics* 13, 4 (2007), 395–414.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[11] Ly Dinh, Yi-Yun Cheng, and Nikolaus Parulian. 2020. ReTracker: an open-source plugin for automated and standardized tracking of retracted scholarly publications. In *Proceedings of the 18th Joint Conference on Digital Libraries (JCDL '19)*. IEEE Press, 406–407. doi:10.1109/JCDL.2019.00092

[12] James M. DuBois, Emily E. Anderson, John Chibnall, Kelly Carroll, Tyler Gibb, Chiji Ogbuka, and Timothy Rubbelke. 2013. Understanding Research Misconduct: A Comparative Analysis of 120 Cases of Professional Wrongdoing. *Accountability in Research* 20, 5–6 (2013), 320–338.

[13] Bram Duyx, Miriam J.E. Urlings, Gerard M.H. Swaen, Lex M. Bouter, and Maurice P. Zeegers. 2017. Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of Clinical Epidemiology* 88 (2017), 92–101. doi:10.1016/j.jclinepi.2017.06.002

[14] Michael Färber and Ashwath Sampath. 2020. HybridCite: A Hybrid Model for Context-Aware Citation Recommendation. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 117–126. doi:10.1145/3383583.3398534

[15] Fabián Freijedo-Farinas, Alberto Ruano-Ravina, Mónica Pérez-Ríos, Joseph Ross, and Cristina Candal-Pedreira. 2024. Biomedical retractions due to misconduct in Europe: characterization and trends in the last 20 years. *Scientometrics* 129, 5 (2024), 2867–2882.

[16] Yuanxi Fu and Jodi Schneider. 2020. Towards Knowledge Maintenance in Scientific Digital Libraries with the Keystone Framework. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 217–226. doi:10.1145/3383583.3398514

[17] Yuki Furuse. 2024. Characteristics of retracted research papers before and during the COVID-19 pandemic. *Frontiers in Medicine* Volume 10 - 2023 (2024). doi:10.3389/fmed.2023.1288014

[18] Jean-Christophe Goulet-Pelletier and Denis Cousineau. 2018. A review of effect sizes and their confidence intervals, Part I: The Cohen'sd family. *The Quantitative Methods for Psychology* 14, 4 (2018), 242–265.

[19] Steven A Greenberg. 2009. How citation distortions create unfounded authority: analysis of a citation network. *Bmj* 339 (2009).

[20] Soo Young Hwang, Dong Keon Yon, Seung Won Lee, Min Seo Kim, Jong Yeob Kim, Lee Smith, Ai Koyanagi, Marco Solmi, Andre F Carvalho, Eunyoung Kim, et al. 2023. Causes for retraction in the biomedical literature: a systematic review of studies of retraction notices. *Journal of Korean medical science* 38, 41 (2023).

[21] Ken Hyland. 2003. Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of the American Society for Information Science and technology* 54, 3 (2003), 251–259.

[22] John PA Ioannidis. 2015. A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. *Journal of psychosomatic research* 78, 1 (2015), 7–11.

[23] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140* (2023).

[24] Anton Kühberger, Daniel Streit, and Thomas Scherndl. 2022. Self-correction in science: The effect of retraction on the frequency of citations. *Plos one* 17, 12 (2022), e0277814.

[25] Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin* 108, 3 (1990), 480.

[26] Holly Fernandez Lynch, Emily A Largent, Matthew S McCoy, and Steven Joffe. 2025. Advancing Trust in Science: Institutional Obligations to Promote Research Integrity. *Journal of Law, Medicine & Ethics* 53, 1 (2025), 1–5.

[27] Chifumi Nishioka, Michael Färber, and Tarek Saier. 2022. How does author affiliation affect preprint citation count? analyzing citation bias at the institution and country level. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 28, 8 pages. doi:10.1145/3529372.3530953

[28] Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. 2012. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science* 7, 6 (2012), 615–631.

[29] Allard Oelen, Mohamad Yaser Jaradeh, Markus Stocker, and Sören Auer. 2020. Generate FAIR Literature Surveys with Scholarly Knowledge Graphs. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 97–106. doi:10.1145/3383583.3398520

[30] Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022. Specialized document embeddings for aspect-based similarity of research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 7, 12 pages. doi:10.1145/3529372.3530912

[31] Josh Pasek. 2018. It's not my consensus: Motivated reasoning and the sources of scientific illiteracy. *Public Understanding of Science* 27, 7 (2018), 787–806.

[32] David Pride and Petr Knoth. 2020. An Authoritative Approach to Citation Classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 337–340. doi:10.1145/3383583.3398617

[33] Jason Priem, Heather Piwowar, and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833* (2022).

[34] David B Resnik. 2023. Research Misconduct and Questionable Research Practices. In *Handbook of Bioethical Decisions. Volume II: Scientific Integrity and Institutional Ethics*. Springer, 25–36.

[35] Amanda Ross and Victor L Willson. 2017. Paired samples T-test. In *Basic and advanced statistical tests: Writing results sections and creating tables and figures*. Springer, 17–19.

[36] Richard E Rubin and Rachel G Rubin. 2020. *Foundations of library and information science*. American Library Association.

[37] Kiran Sharma. 2021. Team size and retracted citations reveal the patterns of retractions from 1981 to 2020. *Scientometrics* 126, 10 (2021), 8363–8374.

[38] Martin Shepperd and Leila Yousefi. 2023. An analysis of retracted papers in Computer Science. *Plos one* 18, 5 (2023), e0285383.

[39] Ivan Stelmakh, Charvi Rastogi, Ryan Liu, Shuchi Chawla, Federico Echenique, and Nihar B Shah. 2023. Cite-seeing and reviewing: A study on citation bias in peer review. *Plos one* 18, 7 (2023), e0283980.

[40] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. 103–110.

[41] The Center for Scientific Integrity. 2018. The Retraction Watch Database. http://retractiondatabase.org/. Accessed: 5/20/2026.

[42] Miriam JE Urlings, Bram Duyx, Gerard MH Swaen, Lex M Bouter, and Maurice P Zeegers. 2021. Citation bias and other determinants of citation in biomedical research: findings from six citation networks. *Journal of Clinical Epidemiology* 132 (2021), 71–78.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[44] Quan-Hoang Vuong, Viet-Phuong La, Manh-Tung Ho, Thu-Trang Vuong, and Manh-Toan Ho. 2020. Characteristics of retracted articles based on retraction data from online sources through February 2019. *Science Editing* 7, 1 (2020), 34–44.

[45] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems* 33 (2020), 5776–5788.

[46] Zijun Wang, Qianling Shi, Qi Zhou, Siya Zhao, Ruizhen Hou, Shuya Lu, Xia Gao, and Yaolong Chen. 2022. Retracted systematic reviews continued to be frequently cited: a citation analysis. *Journal of Clinical Epidemiology* 149 (2022),

137–145. doi:10.1016/j.jclinepi.2022.05.013

[47] Yu Xie, Kai Wang, and Yan Kong. 2021. Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and engineering ethics* 27, 4 (2021), 41.

[48] Shaoxiong Xu and Guangwei Hu. 2022. A cross-disciplinary and severity-based study of author-related reasons for retraction. *Accountability in Research* 29, 8 (2022), 512–536.

[49] Qin Zhang and Hui-Zhen Fu. 2022. Productivity patterns, collaboration and scientific careers of authors with retracted publications in clinical medicine. *Scientometrics* 127, 4 (2022), 1883–1901.

[50] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176* (2025).