publications

MDPI

Article

The Attention Mismatch: Mapping the Structural Academic Governance Deficit in the Age of Generative AI

Zhenning Guo [1], Haoran Mao [1,*] and Fang Zhang [2]

[1] School of Foreign Studies, China University of Petroleum (East China), Qingdao 266580, China; guozhenning2024@163.com
[2] Library, China University of Petroleum (East China), Qingdao 266580, China; im02@163.com
* Correspondence: mao@upc.edu.cn

Abstract

With the rapid advancement in Generative Artificial Intelligence (GenAI), AI-generated content (AIGC) lacking human cognitive oversight is increasingly permeating open web environments and academic communication systems. This study integrates longitudinal retraction data (Retraction Watch Database, 1990–2026), web-scale analyses of AI-content penetration (Common Crawl, 2013–2026), and bibliometric mapping of governance scholarship (Web of Science Core Collection, Scopus, Google Scholar, 2020–2026) to diagnose the cross-level misalignment between synthetic-content diffusion, AI-related misconduct pressure, and governance attention. On this basis, it proposes a Normalized Coverage Index (NCI) to measure the relative relationship between scholarly attention to AI-related academic misconduct governance and the level of misconduct pressure observed through retraction data across disciplines. The results reveal pronounced asymmetries at the disciplinary level. Fields such as chemistry (0.04), physics, mathematics & statistics (0.11), and life sciences & biology (0.34) exhibit clear governance gaps, whereas Education shows a comparatively excessive level of attention (NCI = 29.26). Since 2022, AIGC has expanded rapidly across open web corpora, accompanied by a sharp rise in AI-related retractions, which also exhibit a longer detection lag than traditional forms of misconduct (2.77 years vs. 1.91 years). Although the volume of academic governance-related research has grown rapidly, its proportion within the broader body of AI-related research has declined, suggesting that scholarly attention to governance has not kept pace with technological diffusion. Consequently, a structural misalignment in governance—closely tied to the allocation of attention—has emerged within the academic system in the era of GenAI. This misalignment may pose potential risks to the robustness of the knowledge production system. Addressing it requires rebuilding epistemic infrastructure through provenance transparency, auditable workflows, and governance-aware seed corpora aligned with empirically concentrated risks.

Keywords: synthetic contamination; AI-related academic misconduct; governance misalignment; retraction lag; Normalized Coverage Index; generative AI governance

Check for updates

Academic Editor: Eungi Kim

Received: 4 March 2026
Revised: 7 April 2026
Accepted: 10 April 2026
Published: 17 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid proliferation of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) has fundamentally transformed global knowledge production (Zhang et al., 2025). While these technologies offer unprecedented capabilities in accelerating literature synthesis, code generation, and linguistic polishing, they simultaneously introduce systemic risk—the infiltration of AIGC lacking human epistemic oversight into

the digital and academic ecosystem (Benítez et al., 2024). This risk not only undermines public databases but also poses an equivalent challenge to the academic ecosystem. In this study, we define it as "synthetic contamination," referring to the large-scale influx of AIGC lacking human cognitive oversight into open web environments and scholarly communication systems. Its impacts operate cross-level characteristics. First, it raises the risk of "model collapse" as future AI systems are increasingly trained on synthetic rather than human-produced data (Shumailov et al., 2024). Second, it may erode academic credibility as the boundary between authentic scholarship and machine-generated output becomes increasingly blurred (Ong et al., 2024). Beyond amplifying conventional misconduct such as plagiarism and data fabrication, GenAI enables novel forms of violation that are inherently more difficult to detect—including AI-authored manuscripts with fabricated citations, "hallucinated" results presenting as empirical findings, and sophisticated paper mill operations at scale (Perkins et al., 2023).

In the realm of research integrity, the emergence of AI-related misconduct represents a "Version 2.0" of traditional academic misconduct, characterized by three fundamental distinctions: scalability and automation; increased concealment and detection difficulty; and the blurring of responsibility attribution. Beyond conventional plagiarism and data fabrication, the integration of GenAI into scholarly workflows has facilitated the rise in sophisticated paper mills, untraceable hallucinations in peer-reviewed literature, and the manipulation of authorship processes (Sridharan & Sivaramakrishnan, 2026). Although international regulatory bodies and publishing ethics committees, such as the Committee on Publication Ethics (COPE), have issued preliminary guidelines for AI use, the reactive nature of policy-making often lags behind the velocity of technological misuse (Kocak, 2024). When an LLM generates significant portions of a manuscript's text, analysis, or even its core argument, the line between tool and contributor becomes blurred. This has led to cases of "authorship manipulation," where researchers fail to disclose AI assistance, thereby presenting AIGC as their own intellectual product—a clear violation of transparency norms (Pellegrina & Helmy, 2025). The problem is exacerbated by the fact that many current journal policies, while requiring disclosure of AI use, do not mandate the submission of the specific prompts and raw AI outputs, making it impossible for reviewers or readers to assess the extent and nature of the AI's contribution (Kendall, 2024). This lack of a verifiable audit trail creates a significant gap in accountability, allowing for the potential laundering of AIGC into the scholarly record under the guise of human authorship. Against this backdrop, a key question arises: does academic attention align with the actual distribution of risk? If the disciplinary distribution of governance research fails to reflect the empirically observed pressure of misconduct across fields, this suggests a deviation in the allocation of cognitive resources within the academic system. This study further conceptualizes this phenomenon as "academic attention mismatch," defined as a systematic inconsistency between the attention allocated by governance research and the risk distribution inferred from AI-related retraction data.

Empirical studies have begun to document the infiltration of AI-generated text within specific disciplines (Mortlock & Lucas, 2024; Yao et al., 2025), but such research remains largely fragmented across different levels of analysis. One line of research focuses on potential contamination at the corpus level. For instance, analyses of large-scale web archives suggest that, since the release of GPT-4, the proportion of synthetic text in the public digital commons may have increased dramatically, thereby potentially creating a feedback loop of misinformation (Hanley & Durumeric, 2024). Another line of work examines identified cases of misconduct. A bibliometric study covering retractions in the fields of AI and machine learning from 1974 to 2024 reports that, among 3456 retracted papers, a substantial proportion involved undisclosed use of AI, AI-generated text misrepresented

as original content, and irreproducible findings resulting from model hallucinations (Nag et al., 2025). More concerningly, randomly generated content has emerged as a major driver of retractions: in 2024 alone, over 2300 such cases were recorded—far exceeding the mere three cases documented in 2010 (Lei et al., 2024). Meanwhile, the academic community's response at the governance level remains largely confined to normative discussions. Bibliometric analyses have traced a steady rise in references to "AI ethics" and "AI governance." However, a significant gap persists between the theoretical discourse on AI regulation and the empirical reality of AI-driven retractions across different disciplines (Ganjavi et al., 2024). Moreover, to the best of our knowledge, no existing study has integrated multidimensional evidence to examine—within the context of AI penetration across open online platforms—the distribution of AI-related academic misconduct governance research across disciplines, nor the systematic divergence between this distribution and the empirically observed pressures of misconduct as reflected in retraction data.

This study develops a multi-scalar analytical framework that integrates evidence across three levels. First, drawing on retraction databases, it characterizes the temporal evolution, typological structure, and "retraction lag" associated with AI-related academic misconduct. Second, through modeling and analysis of long-term corpora derived from Common Crawl, it assesses the penetration trends of AIGC across the open web. Third, employing bibliometric methods, it systematically examines the thematic structure and disciplinary distribution of research on AI-related academic governance.

On this basis, we introduce the Normalized Coverage Index (NCI) to quantify the relationship between governance attention and empirically observed misconduct pressure. Specifically, the NCI measures the relative proportion between governance attention (operationalized as publication share) and the empirical pressure inferred from AI-related retractions (used as a proxy for detected misconduct risk). An NCI value close to 1 indicates approximate proportional alignment between the two distributions. Values substantially below 1 suggest potential under-coverage (i.e., governance attention lags behind observed retraction pressure), whereas values substantially above 1 indicate over-coverage (i.e., governance attention exceeds observed pressure). This proportional benchmark serves as a heuristic reference point for identifying macro-level structural mismatches in cross-disciplinary resource allocation. It draws on principles from research integrity studies, emphasizing that governance efforts should respond to empirically observable risks.

By integrating evidence from the corpus level, behavioral level, and governance level, this study aims to provide a diagnostic framework for the "synthetic contamination," enabling the identification of divergences between risk distribution and governance configuration within the academic system. In doing so, it offers an empirical basis for subsequent resource allocation and policy design.

Specifically, this study addresses the following research questions:

RQ1: What are the temporal dynamics and disciplinary distribution patterns of AI-related academic misconduct? How does its retraction lag differ from that of traditional forms of misconduct?

RQ2: What are the long-term penetration trends of AIGC in open web corpora?

RQ3: To what extent does governance research on AI-related academic misconduct across disciplines align with the empirically observed misconduct pressures reflected in retraction data?

2. Materials and Methods

2.1. Detection of the Severity of AI-Related Academic Misconduct

To quantify the severity and characteristics of AI-related academic misconduct, this study utilized the Retraction Watch Database, an academic integrity monitoring database

that systematically collects retracted papers from peer-reviewed journals worldwide along with the stated reasons for retraction (Kwee & Kwee, 2023). We extracted all retraction records from the database and, based on the standardized classification of retraction reasons provided therein, identified cases related to GenAI, which we consolidated into the following five categories: AI Content Generation & Abuse; Data & Result Fabrication; Plagiarism & Duplicate Publication; Authorship & Process Violations; Article Quality & Citation Issues. The mapping rules between the original retraction reasons and the five categories are provided in Supplementary Table S1.

Annual and cumulative retraction counts for each category were calculated through frequency statistics and compared with the distributional characteristics of overall retractions, in order to characterize the temporal evolution trends of AI-related retraction types. Further analyses were conducted by discipline, journal, and country/region to identify categories with high retraction frequencies.

In addition, "retraction lag" (defined as the time interval between publication and retraction, measured in years) was constructed as a continuous variable for comparative analysis. First, the time difference was calculated using the OriginalPaperDate and RetractionDate fields from the original database. Based on retraction-reason keywords, the sample was divided into two groups: AI-related and Traditional. Gaussian kernel density estimation (Gaussian KDE, bandwidth = 0.3) was applied to visualize the continuous distribution patterns, complemented by boxplots to present medians and interquartile ranges.

Given the evident right-skewed distribution of retraction lag, the nonparametric Mann–Whitney U test was used to compare median differences between the two groups, and the Kolmogorov–Smirnov (KS) test was employed to compare differences in the overall distribution functions. Cohen's *d* was calculated as an effect size measure.

A further stratified analysis was conducted at the level of AI subcategories. First, cases were grouped according to the five predefined types of AI-related retractions. For each group, the mean, median, standard deviation, and sample size were calculated.

Boxplots were used to illustrate distributional differences across categories. In addition, a forest plot–style comparative framework was constructed using the group means and their 95% confidence intervals (based on standard error estimates under the t-distribution). The mean of the Traditional retraction group was used as the reference line for horizontal comparison.

2.2. Detection of AIGC Penetration

To assess the temporal evolution of potential AIGC in open web corpora, we developed an "AI-likeness" measurement framework for large-scale web text, aimed at systematically characterizing changes in linguistic generation features.

The data were sourced from the Common Crawl web archive, which periodically crawls publicly accessible web pages and organizes them into sequential crawl batches. We analyzed web corpora from 2013 to 2026, a period during which the data format and sampling procedures were relatively stable, allowing for cross-year comparability (Jazbec et al., 2021). Data extracted as of 30 January 2026. To balance computational feasibility and sample representativeness, analysis samples were constructed from the downloaded annual web text files, with reproducibility ensured through a standardized random selection process and a fixed random seed (seed = 42).

For training data construction, English web text from 2013–2021 was used as human-written negative samples, employing a cross-year balanced sampling strategy to control annual sample sizes and avoid distributional biases caused by unequal data volumes across years. Positive samples of AI-generated text were obtained from the HC3 dataset. Candidate fields were extracted from its JSONL structure, filtered according to unified

preprocessing rules, and compiled into a class-balanced binary training dataset to support subsequent model learning.

Given the prevalence of structural noise and template content in web corpora, we implemented a hierarchical text-cleaning pipeline. First, WARC/ARC format text was split at the record level, with metadata fields (e.g., WARC headers) removed. Common web templates—including navigation bars, comment sections, interactive prompts, and copyright statements—were then identified and removed using rule-based patterns. Text was further normalized by standardizing whitespace, removing URLs and email addresses, filtering non-English characters, and converting to lowercase. Subsequently, text segments were filtered using quality constraints, including word count, sentence count, proportion of English characters, lexical diversity, repetition rate, and numeric content, to ensure consistency in length and linguistic structure. Residual template patterns were further detected and removed based on template matching frequency and sentence structure, mitigating structural noise in model inputs.

Modeling employed an ensemble learning framework that integrates word-level and character-level features. Specifically, TF-IDF representations were used to extract word n-grams (1–2 g) and character n-grams (3–5 g), which were each input into a logistic regression classifier with class weight balancing. The two models independently learned semantic-level and formal-level generative features, and their predicted probabilities were combined via a weighted average (0.6 for word-level, 0.4 for character-level) to generate a unified AI-likeness score, reducing biases inherent to single-feature representations.

Model performance was evaluated on train-test splits using five-fold cross-validation to obtain robust ROC-AUC estimates. Additional metrics—including accuracy, precision, recall, F1 score, and ROC-AUC—were reported to comprehensively assess model performance across multiple discriminative dimensions. The trained ensemble model was subsequently applied to web text from 2013–2026. For each file, sentence-level segmentation was performed, and text segments were constructed according to length and structural constraints. AI-likeness probabilities were computed for each segment and aggregated at the file and annual levels to derive yearly mean AI-likeness scores and distributional statistics. All data processing and modeling procedures were implemented in Python 3.11.

2.3. Bibliometric Analysis of AI-Related Academic Misconduct Governance

To systematically examine the research progress and developmental trends in the field of academic misconduct governance in the context of AI, this study constructs a multi-source data integration and retrieval framework. Data were retrieved from three international literature databases: Web of Science Core Collection (WoSCC), Scopus, and Google Scholar. The search period was limited to 1 January 2020 to 30 January 2026, covering the critical stage of rapid development of generative AI and the associated discussions on academic integrity governance.

WoSCC is characterized by rigorous journal selection standards and highly structured citation data, making it suitable for high-quality bibliometric analysis (F. Wu et al., 2022). Scopus offers advantages in interdisciplinary coverage breadth and the timely inclusion of emerging research topics (Kanmodi et al., 2022). Given that the iterative pace of generative AI technologies far exceeds that of traditional academic publishing cycles, reliance solely on conventional authoritative databases may lead to the omission of emerging governance discussions due to publication lag. Therefore, this study further incorporates Google Scholar as an additional retrieval source, with the aim of capturing open-access (OA) journal publications and early-stage scholarly debates that may not yet be indexed in WoSCC or Scopus (Villatte et al., 2020). The combined use of these three databases enhances the comprehensiveness and representativeness of the sample.

The literature search was completed on 30 January 2026. A predefined search strategy was applied to the Title, Abstract, and Author Keywords fields (Title–Abstract–Keywords). The search strategy was constructed around four core concepts: AI technologies; the academic domain; misconduct behaviors; and governance and regulation. Keywords corresponding to these four dimensions were combined using the Boolean operator "AND" to ensure that the retrieved records simultaneously addressed AI, academic misconduct issues, and governance mechanisms, thereby accurately focusing on the interdisciplinary field of "AI-related academic misconduct governance."

A quality control procedure was implemented during the data preparation stage. The dataset was refined by excluding non-original studies and non-English publications (including editorial materials, book chapters and letters), as well as removing irrelevant records that did not involve AI technologies or their application contexts, or did not focus on academic misconduct. Ultimately, only English-language journal articles and review papers were retained for analysis. Detailed search strategies for each database are provided in Supplementary Table S2, and the PRISMA flow diagram is presented in Figure 1.

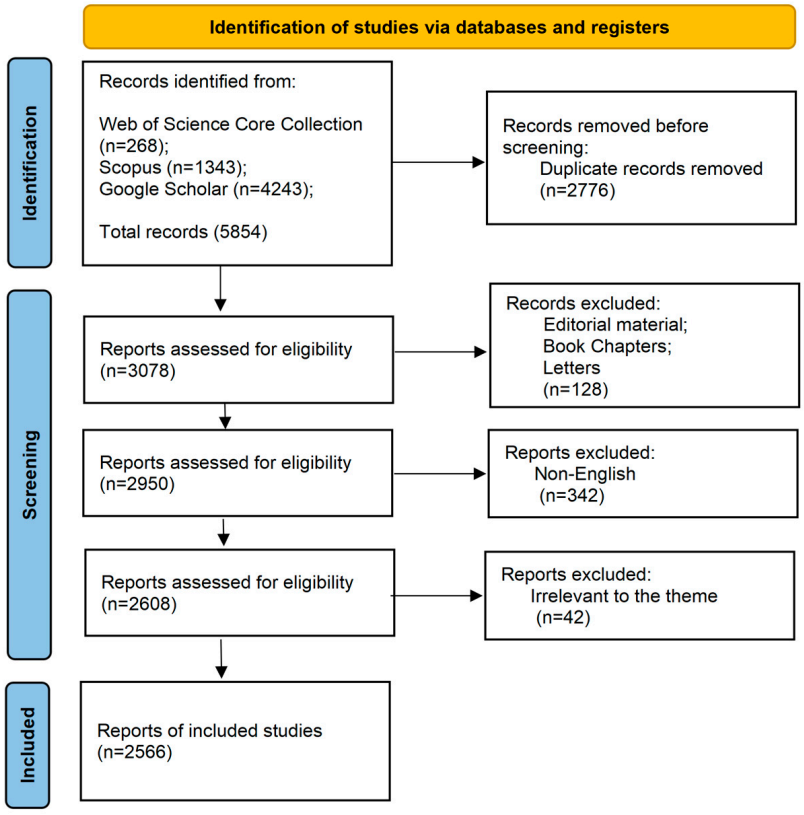


Figure 1. PRISMA Flow Diagram of Literature Identification, Screening, and Inclusion.

During the data analysis stage, synonymous terms were consolidated (e.g., "artificial intelligence" and "artificial intelligence (AI)" were treated as equivalent expressions). Python 3.11 was used to generate a keyword word cloud, visualizing the top 70 most frequently occurring terms. Keyword co-occurrence mapping and clustering analysis were conducted using VOSviewer (version 1.6.20) to delineate the conceptual structure of research on AI-related academic misconduct governance. Author keywords extracted from the bibliometric dataset were used as the unit of analysis. A co-occurrence network was constructed by linking pairs of keywords that appeared together within the same publication. Each pair of keywords co-occurring in a document was connected, with edge weights defined by their total co-occurrence frequency across the corpus, thereby forming

a keyword–keyword co-occurrence matrix. Following prior studies (Pattnaik, 2023), a minimum co-occurrence threshold ($\geq$5) was applied, and isolated nodes were removed during network construction. Based on VOSviewer's built-in visualization and mapping techniques, the network was clustered into distinct thematic groups, with different colors representing different clusters. Node size is proportional to keyword frequency, edge thickness reflects co-occurrence strength, and the spatial distance between nodes indicates the closeness of their relationships.

In addition, the productivity of high-output journals, countries, and publishers was visualized. The annual publication trends of research on AI-related academic misconduct governance were compared, along with changes in its annual proportion within the broader field of AI-related research.

Further keyword burst detection was conducted using the Python NumPy library. First, the "Year" and "Author Keywords" fields were extracted. Records with missing publication years were removed, and the year variable was converted into an integer format. A two-dimensional "keyword–year" data table was then constructed to generate the annual frequency matrix of keywords. Meanwhile, the total annual keyword occurrences were calculated as the yearly baseline. To ensure statistical stability, only keywords with a total frequency of $\geq$5 were retained for subsequent analysis.

Keyword burst detection was performed using the burst detection algorithm based on a hidden-state automaton proposed by Kleinberg's Burst Detection Algorithm (Xu et al., 2020). Specifically, the annual frequency of each keyword was treated as a time series. Within a two-state model framework (normal state/burst state), emission probabilities were constructed based on the binomial distribution, and the optimal state sequence was determined using the Viterbi dynamic programming algorithm. The model parameters were set in accordance with standard practice in Kleinberg's burst detection method, with the state transition base fixed at s = 2.0 and the penalty coefficient at γ = 1.0 to regulate transition costs. Sensitivity analyses using alternative parameter settings (s = 1.5–2.5; γ = 0.5–2.0) yielded highly consistent results (Spearman's $\varrho > 0.9$), confirming the robustness of the findings. For keywords with detected burst intervals, the burst strength (burst score) was calculated. The overall strength metric was computed as: Strength = burst score $\times$ ln(1 + total frequency), where the bubble area in the visualization is proportional to the annual frequency of each keyword. The top 13 keywords were then selected based on burst strength, and their burst time intervals, annual frequencies, and burst intensities were visualized accordingly.

2.4. Subject-Level Normalized Coverage Analysis

To assess the alignment between governance attention and empirically observed misconduct pressure across disciplines, we constructed the Normalized Coverage Index (NCI). This index operationalizes cross-disciplinary structural imbalance by comparing two distributions: (i) the share of AI governance research output within a given discipline, and (ii) the proportion of AI-related retractions attributed to that discipline. As a ratio-based metric, the NCI captures relative discrepancies in attention allocation rather than absolute magnitudes. Retraction share is used as an empirical proxy for observed misconduct pressure, while acknowledging that detection and reporting practices may influence observed retraction patterns.

In this study, an NCI value of 1 indicates proportional alignment between governance attention and retraction pressure. Values below 1 suggest governance attention lags behind the observed pressure, whereas values above 1 indicate governance attention exceeds the observed pressure. Accordingly, the NCI serves as a heuristic benchmark for identifying

macro-level structural imbalances in cross-disciplinary attention allocation, consistent with research integrity principles emphasizing responsiveness to empirically observable risks.

Two datasets were employed: (i) AI governance publications extracted from bibliometric sources and pre-classified according to Web of Science (WoS) subject categories, and (ii) AI-related retractions retrieved from the Retraction Watch database.

To ensure cross-dataset comparability, we implemented a rule-based keyword matching algorithm to harmonize original subject labels into 13 aggregated super-disciplinary categories. Classification reliability was assessed through double-coding of 200 randomly sampled records (100 per dataset), yielding high agreement (Cohen's $\kappa = 0.81$). Remaining discrepancies were resolved through consensus. Approximately 3% of records exhibited multiple or ambiguous subject labels; these were disambiguated using a majority-voting approach. Robustness checks using fractional counting produced highly consistent NCI rankings (Spearman's $\varrho > 0.9$). The mapping scheme was further reviewed by domain experts to ensure semantic consistency and classification validity across data sources.

For each discipline i, governance share and retraction share were calculated as proportions within their respective datasets. Governance share represents the proportion of AI governance publications assigned to discipline i, while retraction share denotes the proportion of AI-related retractions attributed to the same discipline.

The NCI is then defined as

$$\mathrm{NCI}_i = \frac{\text{Governance Share}_i}{\text{Retraction Share}_i + \varepsilon} \quad (1)$$

where $\varepsilon = 0.01$ is a regularization constant introduced to avoid division by zero in disciplines with zero retractions. Sensitivity analyses using ε values of 0.005, 0.01, and 0.02 yielded highly stable rankings (Spearman's $\varrho > 0.95$).

To assess the robustness of NCI estimates under classification uncertainty, we conducted bootstrap resampling (2000 iterations, with replacement) at the record level for both datasets. For each discipline, 95% bootstrap confidence intervals of NCI estimates were computed. Resampling was performed independently for governance and retraction datasets, with consistent denominator definitions and filtering rules applied to both point estimates and resampled data, ensuring internal consistency.

Results were visualized using a three-column heatmap (governance share, retraction share, NCI). A diverging color scale centered at NCI = 1 (TwoSlope normalization) was applied to preserve symmetry around the equilibrium threshold and avoid perceptual bias in skewed distributions.

3. Results

3.1. The Degree of AIGC Penetration and the Severity of AI-Related Academic Misconduct

Figure 2 systematically presents the long-term evolution of global retraction counts from 1990 to 2026 and the phased changes in risks associated with GenAI.

Figure 2A illustrates the annual trends of the five retraction reasons most strongly associated with AI, including: (1) AI content generation and abuse, (2) data and results fabrication, (3) plagiarism and duplicate publication, (4) authorship and peer-review violations, and (5) article quality and citation issues. These trends are compared with the annual retraction patterns of traditional categories in the Retraction Watch Database. Since 2022, retractions related to AIGC-generated content have exhibited a sudden surge, with a growth rate significantly exceeds the historical average and clearly diverges from the patterns observed in traditional retraction categories. This indicates that the new risks introduced by the integration of generative models into scientific writing processes are rapidly becoming explicit. Figure 2B presents the cumulative changes in retraction counts

across categories from 1990 to 2026, more clearly revealing a steep upward trend in the cumulative retraction structure beginning in 2020.

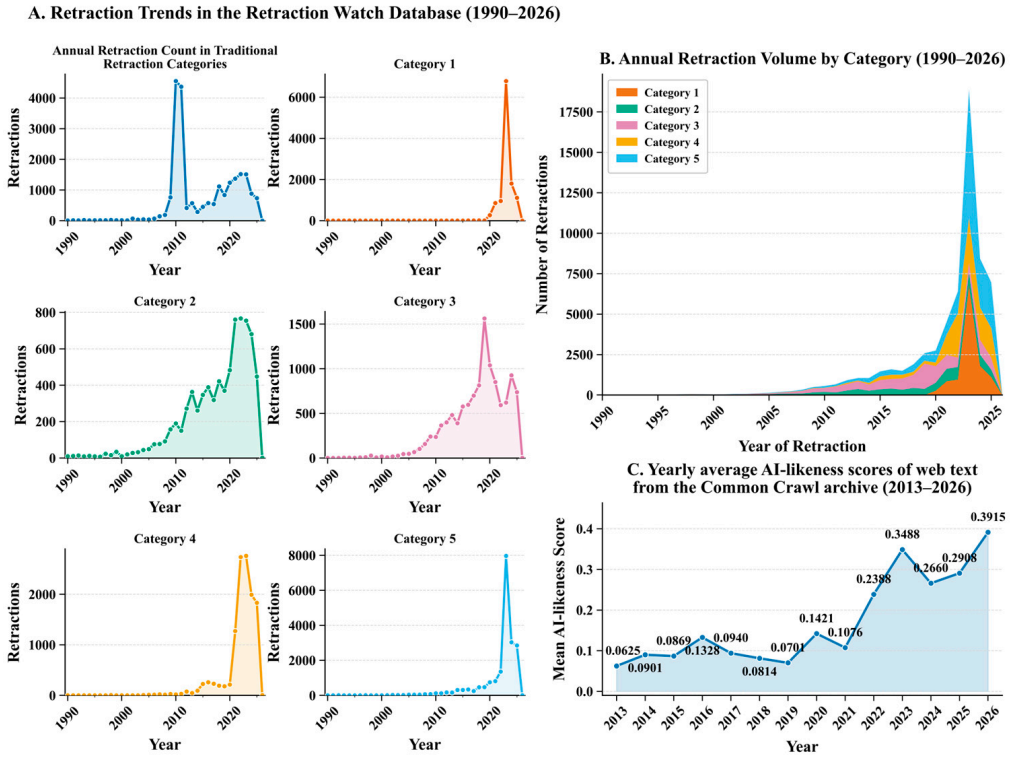


Figure 2. Temporal Dynamics of AIGC Penetration and Severity of AI-Related Academic Misconduct.

Figure 2C shows the trend in the yearly average AI detection score (yearly average AI likeness score) for web texts from the Common Crawl corpus during 2013–2026. Overall, the indicator remained at a low and relatively stable level during 2013–2021, followed by a marked increase starting in 2022. It then persisted at a substantially higher range between 2022 and 2026 (0.2660–0.3915), indicating a significant and sustained rise in the penetration of AIGC in web text corpora.

Model evaluation of the AI-likeness detection framework demonstrates consistently high discriminative performance across different methods. The word-level TF-IDF logistic regression model achieved an accuracy of 0.9918, an F1 score of 0.9953, and an ROC-AUC of 0.9991 (CV: 0.9991 $\pm$ 0.0004). The character-level TF-IDF logistic regression model yielded an accuracy of 0.9961, an F1 score of 0.9978, and an ROC-AUC of 0.9997 (CV: 0.9997 $\pm$ 0.0002). The ensemble model achieved an accuracy of 0.9947, an F1 score of 0.9969, and an ROC-AUC of 0.9995. In the test set, the ensemble model exhibited a low misclassification rate of only 0.53%.

Figure 3 shows that AI-related retractions are primarily concentrated in the fields of technology and computer science, with Technology (5241) and Computer Science (3844) ranking first and second, respectively, followed by disciplines such as Data Science and Genetics.

At the journal level, the *Journal of Intelligent & Fuzzy Systems* (1041) has the highest number of retractions, followed by *Security and Communication Networks* (798), *BioMed Research International* (788), and *Journal of Physics: Conference Series* (782). The clustering of retractions in specific journals may reflect venue-specific editorial or review process vulnerabilities.

In terms of countries/regions, China (11,952) significantly exceeds other countries, ranking first, followed by India (1508), with Saudi Arabia (610) and Pakistan (410) among the subsequent contributors. Overall, the distribution exhibits clear disciplinary and geo-

graphic concentration patterns. It should be noted that the country-level results reported above are based on absolute counts and should be interpreted as indicators of distributional patterns rather than normalized rates of misconduct. These data should not be overinterpreted as a direct reflection of relative levels of research integrity across countries.

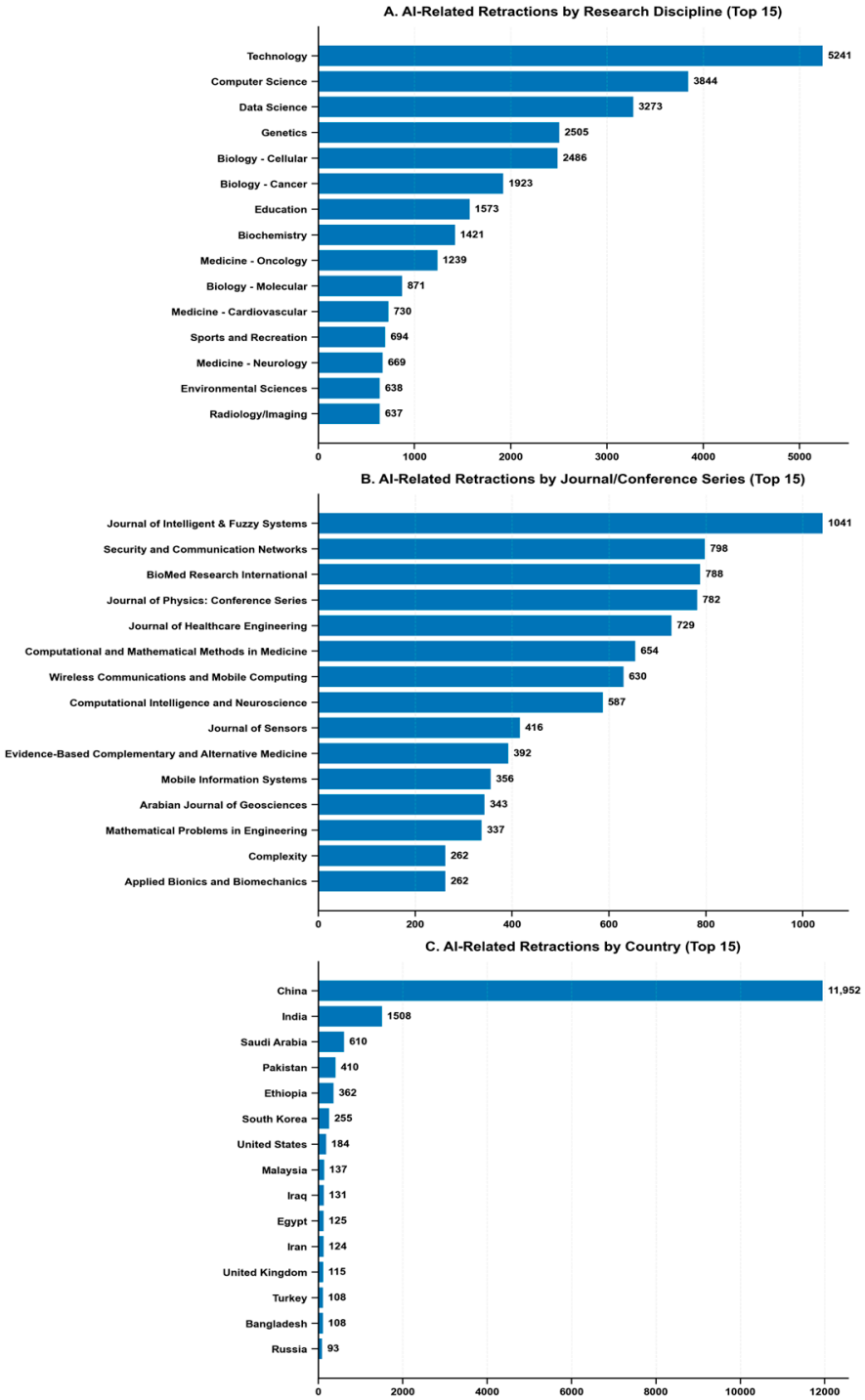


Figure 3. Disciplinary, Journal, and Geographic Distribution of AI-Related Retractions.

Figure 4 compares the distributional characteristics of retraction delays between AI-related misconduct and traditional academic misconduct. The results show that the mean retraction delay for AI-related papers is 2.77 years, with a median of 1.62 years, both significantly higher than those for traditional misconduct (mean = 1.91 years; median = 0.24 years). Kernel density estimates and boxplots consistently indicate a right-

ward shift and a longer tail for AI-related retractions. Both the Mann–Whitney U test and the Kolmogorov–Smirnov (KS) test reach a high level of statistical significance ($p < 0.001$), suggesting systematic differences between the two groups in terms of both median values and overall distributional patterns. Cohen's $d = 0.25$ indicates a small to medium effect size. In terms of temporal trends, retraction delays for AI-related papers have remained at a relatively high and fluctuating level in recent years, overall reflecting a significant lag in the academic governance system's response to AIGC.

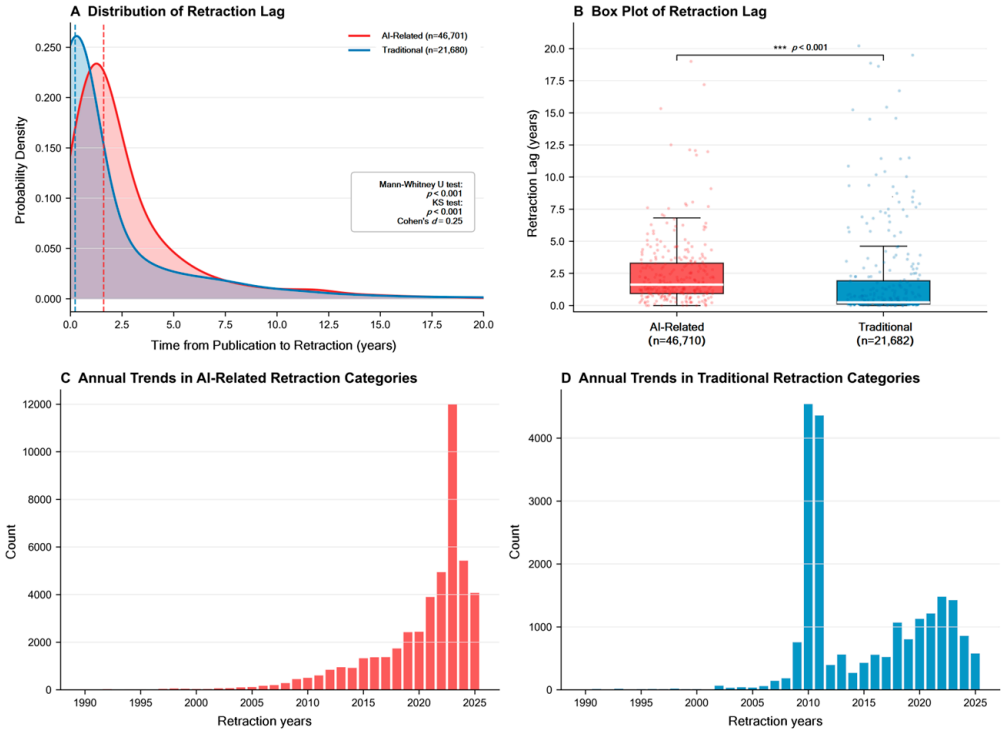


Figure 4. Comparative Retraction Lag Analysis: AI-Related Misconduct Categories vs. Traditional Academic Misconduct Categories.

Figure 5 presents a stratified comparison of retraction lag across AI-related misconduct categories. At the subcategory level, substantial heterogeneity was observed. Category 2 (Data & Result Fabrication) showed the longest delay (mean = 4.472 years; median = 2.916; SD = 4.663), indicating that fabrication-related cases remain the most time-consuming to detect and retract. Category 3 (Plagiarism & Duplicate Publication) also demonstrated prolonged delays (mean = 3.390 years; median = 2.138), exceeding both the overall AI-related average and traditional misconduct. In contrast, Category 1 (AI Content Generation & Abuse) displayed a comparatively shorter lag (mean = 1.964 years; median = 1.525), approaching the traditional group's mean but maintaining a higher median, suggesting more consistent detection timelines. Category 4 (Authorship & Process Violations) and Category 5 (Article Quality & Citation Issues) exhibited the shortest delays among AI-related categories (means = 1.831 and 1.779 years, respectively), though their medians remained above that of traditional misconduct.

Forest-plot comparison of mean retraction lags with 95% confidence intervals (Figure 5B) yielded the same result. These findings indicate that not all AI-related misconduct exerts equivalent governance pressure. Instead, detection latency varies markedly by violation type, with fabrication- and plagiarism-related cases generating the greatest temporal governance burden. Collectively, the results underscore a structural lag in institutional response to AI-mediated misconduct, particularly for cases involving substantive data manipulation rather than surface-level textual generation. This pattern suggests that

AI-assisted fabrication involving underlying data and core arguments may be associated with longer detection timelines; however, this interpretation is inferred from retraction lag patterns rather than directly observed detection mechanisms and therefore should be treated with caution.

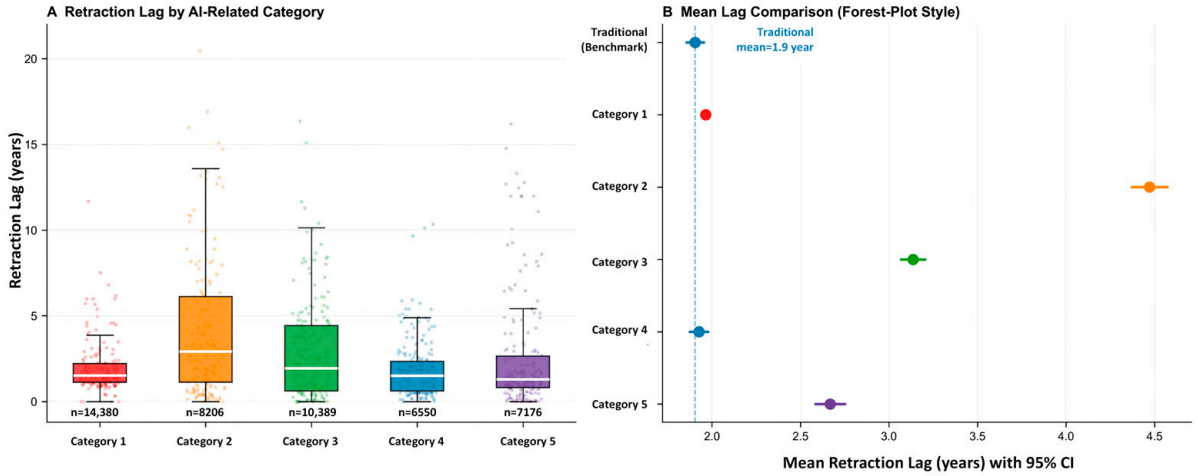


Figure 5. Category-Specific Retraction Lag: Forest Plot Comparison Between AI-Related and Traditional Misconduct.

3.2. Analysis of the Current State of Governance of AI-Related Academic Misconduct

Figure 6 systematically presents the structural characteristics of research on the governance of AI-related academic misconduct across three dimensions: journal/conference series, national distribution, and annual publication trends.

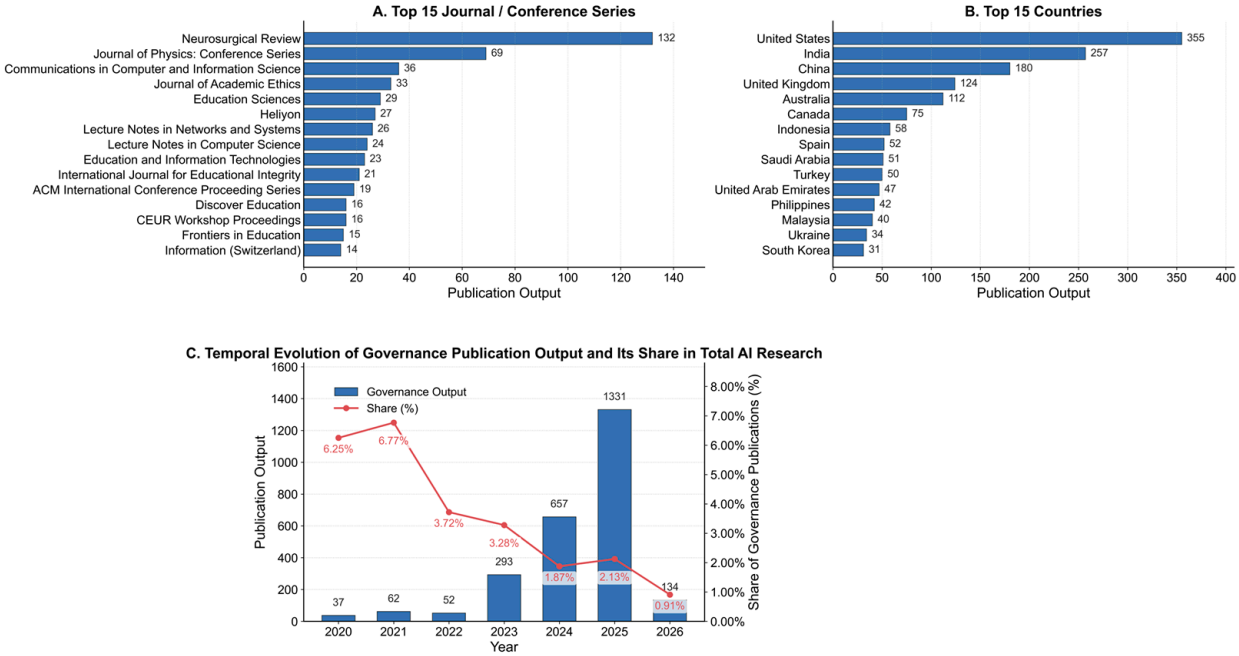


Figure 6. Global Landscape and Temporal Trends of Governance Research on AI-Related Academic Misconduct.

The journal/conference series distribution (Figure 6A) indicates that relevant studies are primarily concentrated in interdisciplinary outlets spanning research ethics, information science, and higher education management. This reflects the inherently interdisciplinary nature of governance issues. However, overall publication dispersion remains high: the

top 15 venues account for only 19% of total publications (500/2566). This suggests that the field is still in an integrative stage, with no clear Matthew effect in which outputs are concentrated in a small number of core journals.

At the national level (Figure 6B), a pronounced geographic concentration is observed. High-output countries are mainly those with large research production capacities and active AI technology applications, such as the United States, India, and China. The top 15 countries collectively contribute 59% of the total publications (1508/2566). This pattern indicates a structural coupling between governance-related research output, national scientific capacity, and the level of AI development. At the same time, such disparities may also reflect differences in institutional capacity across countries to identify, document, and address AI-related academic integrity risks.

The annual trend (Figure 6C) shows a marked acceleration in publication output since 2020. In particular, following the rapid diffusion of generative AI technologies in 2022, the field entered a phase of steep growth, indicating an increase in scholarly attention to AI-induced academic integrity challenges. However, when examining the proportion of governance-related publications relative to the overall volume of AI research, this share has continued to decline over time. This suggests that attention to academic governance is not keeping pace with the broader expansion of AI research, and that the governance gap is widening year by year.

Overall, Figure 6 indicates that research on the governance of AI-related academic misconduct is undergoing rapid growth. Nevertheless, its trajectory remains misaligned with the pace of technological diffusion, and both disciplinary structure and global participation exhibit stage-specific imbalances.

Figure 7 presents the 70 most frequently occurring keywords and their relative proportions in research on the governance of AI-related academic misconduct from 2020 to 2026 via Python 3.11, while Supplementary Table S3 lists the specific frequencies of each keyword. Overall, the keyword distribution exhibits a pronounced technology-oriented pattern and is highly concentrated on generative AI and issues of academic integrity.

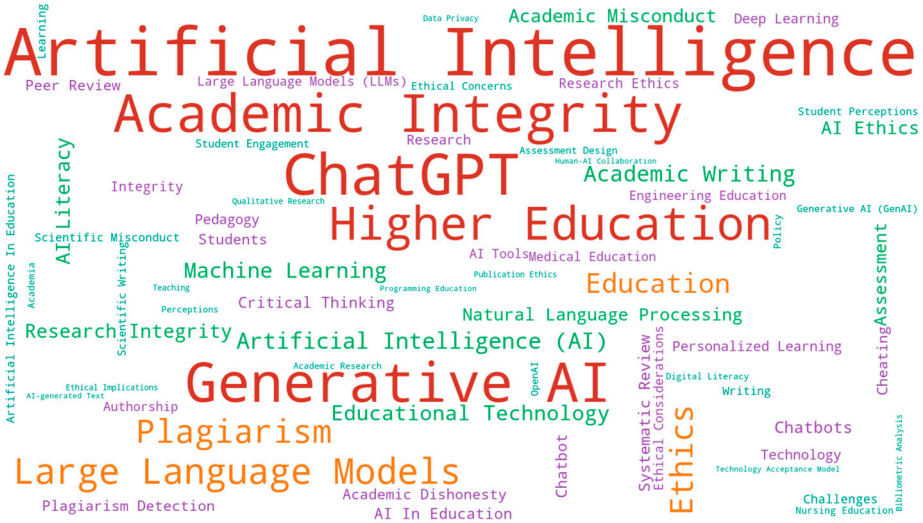


Figure 7. Keyword Frequency Distribution in Governance Research on AI-Related Academic Misconduct (Word Cloud Visualization).

"Artificial Intelligence" appears 716 times (15.65%), ranking first, followed by "ChatGPT" (520, 11.36%) and "Generative AI" (478, 10.44%). Keywords directly related to academic norms, such as "Academic Integrity" (439, 9.59%) and "Higher Education" (333, 7.28%), rank fourth and fifth, respectively. Together, the top five keywords account for

54.32% of the total, indicating a strong concentration of research on generative AI technologies and academic integrity issues in higher education.

From a technological perspective, core technical terms such as "Large Language Models" (180, 3.93%), "Machine Learning" (62, 1.35%), "Natural Language Processing" (41, 0.90%), and "Deep Learning" (23, 0.50%) appear consistently. In addition, terms like "Chatbots" (35, 0.76%), "AI Tools" (21, 0.46%), and "OpenAI" (15, 0.33%) reflect attention to specific application forms and the broader tool ecosystem.

At the level of academic misconduct and governance, "Plagiarism" (147, 3.21%) occupies a central position and, together with "Research Integrity" (52, 1.14%), "Academic Misconduct" (48, 1.05%), and "Academic Dishonesty" (25, 0.55%), forms the main normative framework. Meanwhile, "Plagiarism Detection" (27, 0.59%), "Peer Review" (27, 0.59%), "Authorship" (21, 0.46%), and "Publication Ethics" (12, 0.26%) correspond to specific governance mechanisms and practical approaches.

Overall, the keyword distribution demonstrates a clear head-concentration structure: the top ten keywords each account for more than approximately 1.5%, while frequencies drop rapidly below 1% after the twentieth-ranked keyword, forming a typical long-tail distribution. This suggests that the research field is highly concentrated in its core themes while also exhibiting a certain degree of diversification.

The VOSviewer co-occurrence analysis results (Figure 8) show that, under the specified threshold criteria, a total of 234 keyword nodes were included, forming 15 clusters. The network contains 3081 co-occurrence links, with a total link strength of 8599. Network-level metrics reveal a moderate density (0.112) and a modularity score of 0.68, suggesting relatively well-defined cluster boundaries. The average clustering coefficient is 0.47, indicating a moderate level of intra-cluster cohesion. These results indicate relatively dense and structured associations among keywords within the research field. The thematic divisions exhibit clear modular characteristics, while the overall network demonstrates strong connectivity, reflecting a high level of knowledge integration and cross-thematic interaction.

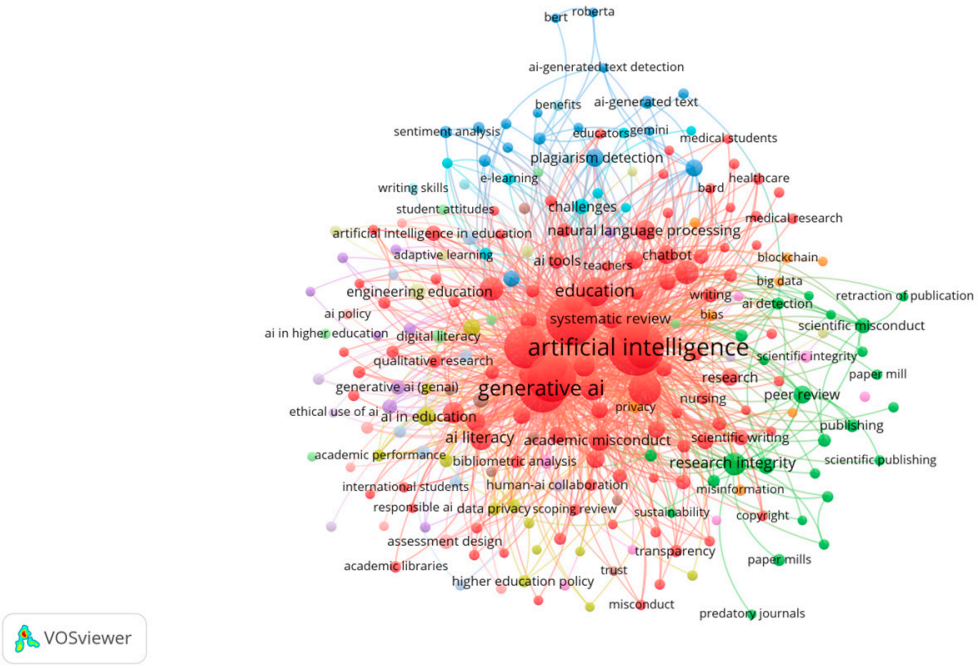


Figure 8. Keyword Co-Occurrence Network in Governance Research on AI-Related Academic Misconduct.

As shown in Figure 9, the top 13 high-intensity keywords identified using Kleinberg's burst detection algorithm exhibit a clear temporal clustering pattern, which can be divided into three sequential phases.



Figure 9. Burst Detection Analysis of Emerging Keywords in Governance Research on AI-Related Academic Misconduct.

The first phase (2020–2022) is characterized by bursts related to "Machine Learning" and "COVID-19." "COVID-19" demonstrates an exceptionally high burst strength during 2021–2022 (291.54), indicating a short-term concentration of research attention in the pandemic context. "Machine Learning" shows a longer burst duration from 2020 to approximately 2023, with moderate intensity (18.53), suggesting sustained baseline relevance during the early period.

The second phase (2022–2024) is marked by the emergence of LLM–related terms. "Language Models" exhibits a strong burst in 2022–2023 (99.13). Beginning in 2022, multiple related keywords—including "OpenAI," "LLM," "ChatGPT," "Chatbot," and "Chatbots"—entered burst status, forming a temporally overlapping cluster. Among them, "OpenAI" shows the highest burst strength in 2023–2024 (297.79). We note that "OpenAI" represents a corporate entity rather than a technology class; its high burst strength reflects scholarly attention to the organization responsible for ChatGPT, which may conflate technology-focused and entity-focused discourse. This stage reflects a shift in keyword prominence from general machine learning to generative AI–related terminology.

The third phase (2024 onward) shows bursts associated with research practice and evaluation. "Academic Research" (57.78) and "Systematic Review" (48.02) emerge around 2024, followed by "Critical Thinking" around 2026 (39.49). These later bursts indicate increasing attention to research processes and evaluative dimensions in the observed literature.

Figure 10 illustrates the variation across 13 major disciplinary categories along three dimensions: governance article share, AI-related retraction share, and the NCI. Supplementary Table S4 reports the 95% bootstrap confidence intervals for the NCI estimates. Retractions reflect detection capacity as much as underlying misconduct. Overall, substantial heterogeneity is observed across disciplines in terms of governance attention relative to retraction pressure.

In several disciplines, the NCI is markedly below 1, indicating insufficient governance attention relative to the burden of AI-related retractions. Among them, Chemistry exhibits the lowest NCI at 0.04 (95% CI: 0.00–0.08), reflecting the most pronounced attention mismatch. Similarly, Physics, Mathematics & Statistics shows a significantly low NCI of 0.11 (95% CI: 0.04–0.22). In addition, Economics, Business & Management (NCI = 0.31, 95% CI: 0.24–0.38), Environmental, Earth & Agricultural Sciences (NCI = 0.33, 95% CI: 0.23–0.44), and Life Sciences & Biology (NCI = 0.34, 95% CI: 0.30–0.37) also demonstrate clearly insufficient governance attention relative to retraction risks.

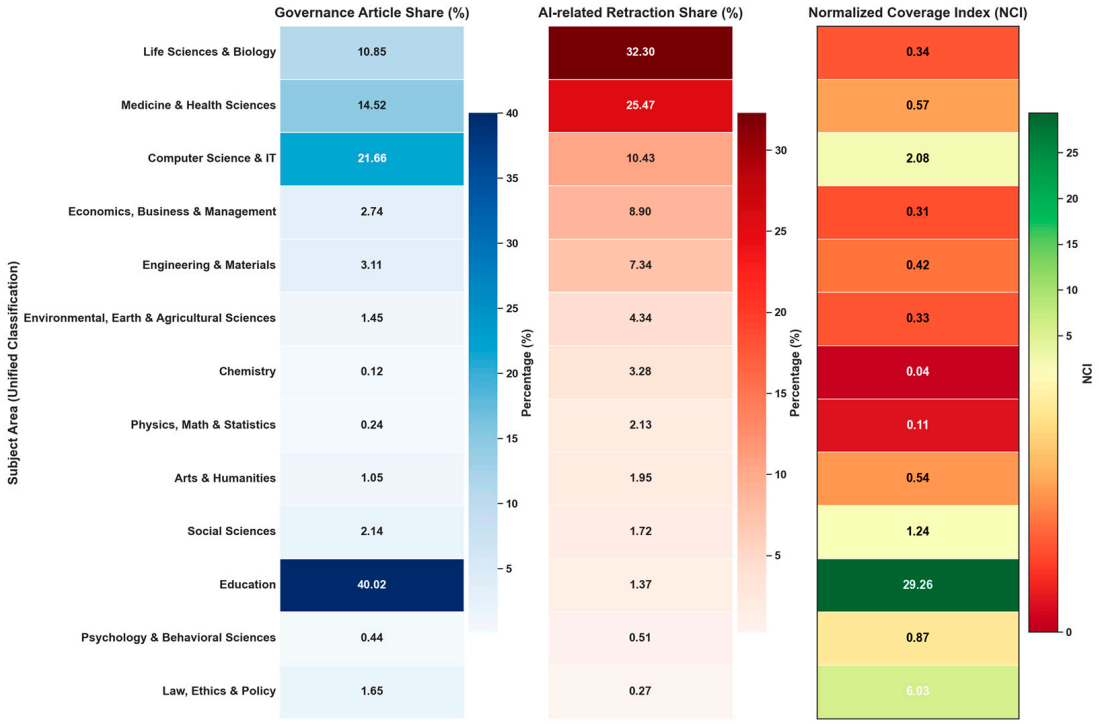


Figure 10. Disciplinary Mismatch Between Governance Research Attention Intensity and Retraction Pressure in AI-Related Academic Misconduct.

Likewise, although Medicine & Health Sciences and Engineering & Materials account for relatively large shares in both governance research and retraction cases, their NCIs remain modest at 0.57 (95% CI: 0.52–0.63) and 0.42 (95% CI: 0.33–0.52), respectively. This suggests that governance attention in these fields still lags behind their corresponding retraction burdens.

In contrast, several disciplines exhibit NCIs substantially greater than 1, indicating an overconcentration of governance research relative to retraction pressure. Computer Science & IT has an NCI of 2.08 (95% CI: 1.92–2.25), Social Sciences 1.24 (95% CI: 0.92–1.61), and Law, Ethics & Policy 6.03 (95% CI: 4.16–8.42). Notably, Education shows an exceptionally high NCI of 29.26 (95% CI: 26.68–32.14). Meanwhile, Psychology & Behavioral Sciences has an NCI of 0.87 (95% CI: 0.38–1.45), approximating a balance between governance attention and retraction risk. Arts & Humanities (NCI = 0.54, 95% CI: 0.33–0.76), although below 1, is associated with relatively high uncertainty.

Taken together, the disciplines with the lowest NCIs—namely Chemistry; Physics, Mathematics & Statistics; Economics, Business & Management; Environmental, Earth & Agricultural Sciences; and Life Sciences & Biology—represent the most critical governance gaps. These findings suggest that the current distribution of research efforts on AI-related academic misconduct governance is misaligned with the actual distribution of AI-related retraction risks across disciplines, revealing a pronounced disciplinary imbalance.

4. Discussion

This study develops a multi-scale integrative framework combining retraction data, large-scale web corpus analysis, and bibliometric mapping of AI-related academic misconduct governance. Based on this framework, we identify a structural mismatch between governance research attention and empirically observed misconduct pressure. This mismatch indicates a potential governance deficit and highlights the growing risk of "synthetic contamination" in the academic ecosystem. By addressing three research questions, we provide an empirical basis for policy intervention and resource reallocation.

4.1. AI-Related Academic Misconduct and Governance Misalignment

Through longitudinal mining of retraction databases, large-scale penetration detection of online textual corpora, and bibliometric mapping of governance literature, this study provides a systematic assessment of synthetic contamination in academic knowledge production. Four key patterns emerge from the multi-scale analysis.

First, AI-related academic misconduct has shown a marked increase in recent years, especially after 2022. Retractions associated with AIGC and data fabrication have risen rapidly, with growth rates clearly exceeding those of traditional forms of academic misconduct (Figure 2). At the same time, the proportion of AI-like textual content in open web corpora has exhibited a parallel upward trend. Analysis based on Common Crawl indicates that, since 2022, the average AI-likeness score of web text has increased significantly and remained at a relatively elevated level (approximately 0.27–0.39 during 2022–2026; see Figure 2C). This temporal shift coincides closely with the widespread adoption of LLMs such as ChatGPT.

Given that web-based text constitutes a major source of training data for contemporary AI systems, the increasing prevalence of AI-like content may introduce potential risks, bringing concerns about "model collapse" from a primarily theoretical discussion toward a more practical consideration. The accumulation of synthetic data may affect model output quality, reinforce existing biases, and alter information feedback dynamics, thereby posing potential challenges to the stability of the digital knowledge ecosystem (Shumailov et al., 2024). In addition, the integration of AI technologies may be associated with more difficult-to-detect forms of misconduct, such as AIGC misuse and hallucinated findings (Lin et al., 2026). It should be noted, however, that these associations require further validation through rigorous causal analyses.

Second, from a temporal perspective, this study indirectly characterizes the detection lag associated with AI-related misconduct based on retraction data. The mean retraction lag for AI-related papers is 2.77 years, with a median of 1.62 years, both of which are higher than those observed for traditional misconduct (mean = 1.91 years; median = 0.24 years). Among subcategories, data fabrication (mean = 4.47 years) and plagiarism (mean = 3.39 years) exhibit the longest delays (Figures 4 and 5). These differences are statistically significant, indicating systematic variation in the time required to identify and correct different types of misconduct.

This finding has important practical implications, as longer detection lags may allow problematic research outputs to circulate within academic networks and influence subsequent knowledge production. In particular, AI-assisted fabrication involving underlying data and core arguments may be more difficult to detect due to its linguistically plausible presentation. Existing governance mechanisms—primarily relying on text similarity detection and peer review—may face limitations when dealing with content that is semantically coherent but factually inaccurate.

It is important to emphasize that this study includes only retracted publications and does not capture cases currently under investigation or instances of misconduct that have

not yet been detected. Therefore, the observed retraction lag should be interpreted as a conservative estimate of the true detection timeline. This further underscores the need for developing proactive detection mechanisms and continuous monitoring systems.

Third, at the level of governance research, Bibliometric analysis indicates a distinctly reactive pattern. Although the volume of related studies has grown rapidly, their relative share within the broader body of AI research has declined (Figure 6). Keyword analysis (Figures 7–9) shows that current research themes are highly concentrated on general topics such as "ChatGPT," "academic integrity," and "plagiarism," while comparatively limited attention has been paid to contexts involving "hallucinated data" or AI-assisted fabrication of results in experimentally intensive disciplines.

This "academic attention mismatch" suggests that the distribution of governance research across disciplines does not fully align with their empirically observed risk exposure. Governance-related research resources appear to be disproportionately concentrated in the social sciences and education, while receiving comparatively less attention in certain core STEM fields. This distributional pattern may be associated with differences in issue visibility, normative traditions, and the ways problems are articulated within disciplines, and may in turn produce asymmetrical impacts on academic credibility across fields.

It is noted that AI-related retractions exhibit clear disciplinary and geographic concentration, with a notable clustering in technical and computer science fields, as well as in regions characterized by high research output intensity (Figure 3). However, this concentration should be interpreted with caution. Absolute retraction counts are not normalized by total publication volume, and countries such as China, which lead global scientific output, would be expected to exhibit higher absolute numbers of retractions irrespective of misconduct rates. Therefore, these patterns do not necessarily imply a higher prevalence of AI-related misconduct.

Instead, the observed distribution may reflect a combination of structural factors. First, higher publication volume increases the exposure surface for both misconduct and its detection. Second, disciplinary differences in publication practices—such as conference-driven outputs in computer science—may influence both the likelihood of problematic submissions and the speed of post-publication scrutiny. Third, editorial and peer-review vulnerabilities in certain journal ecosystems, particularly those with high throughput or weaker quality control mechanisms, may contribute to the observed clustering.

Taken together, these findings suggest that the concentration of AI-related retractions is likely shaped by an interplay between research scale, detection capacity, and editorial governance, rather than representing a direct proxy for underlying misconduct rates. Future research should incorporate normalized indicators (e.g., retractions per 10,000 publications) to disentangle these effects and provide a more accurate assessment of relative risk across disciplines and regions.

Furthermore, based on the discipline-level NCI analysis (Figure 10), we identify substantial variation in the distribution of governance research attention across different disciplines. Fields such as education (NCI = 29.26), computer science & IT (NCI = 2.08), and law, ethics & policy (NCI = 6.03) receive governance research attention far exceeding their empirical retraction shares. In contrast, foundational and applied science domains—including chemistry (NCI = 0.04), physics, mathematics & statistics (NCI = 0.11), economics, business & management (NCI = 0.31), environmental, earth & agricultural sciences (NCI = 0.33), and life sciences & biology (NCI = 0.34)—exhibit clear underrepresentation in governance research relative to their observed risk exposure.

It should be noted that the NCI captures the relative relationship between governance attention and observed risk, rather than providing a direct measure of governance capacity or effectiveness. These disciplinary differences are more likely to reflect the mechanisms

through which governance knowledge is produced and disseminated within the academic system. Discussions surrounding ChatGPT are highly concentrated in educational and normative domains, creating a degree of path dependency (Jiang et al., 2024). In contrast, in the basic sciences, related issues are more often framed as methodological challenges rather than governance concerns, thereby reducing their visibility in the governance literature. Additionally, disparities in issue visibility within public discourse and research funding priorities (Kay et al., 2024) may further reinforce this distributional pattern.

4.2. Research Implications and Future Directions

This study demonstrates that the proliferation of AI-generated synthetic content is not merely an isolated technical issue, but also poses a significant challenge to the academic knowledge production system. Most current governance measures remain focused on behavioral norm declarations (Prifti & Fosch-Villaronga, 2024), the enhancement of detection tools (Perkins et al., 2023), and case-based corrective actions (Pudasaini et al., 2024), which are essentially "peripheral patchwork" responses. However, when synthetic content has become recursively embedded in data sources, publication processes, and reputation allocation systems, reliance solely on ex post identification and punitive mechanisms may be insufficient to fundamentally mitigate the dual risks of training data degradation and erosion of academic trust (Rhodes & Linnenluecke, 2025).

The NCI provides an empirical lens for identifying the relative divergence between governance research attention and observed risk across disciplines. This indicator can serve as a supplementary tool for tracking the evolving relationship between the distribution of governance research and risk signals over time. For disciplines with NCI < 0.5 (where governance attention is markedly lower relative to retraction pressure), priority may be given to developing field-specific detection guidelines, establishing discipline-sensitive review protocols, and increasing investment in research integrity infrastructure.

Secondly, future governance may shift from "violation identification" toward "infrastructure reconstruction." At the knowledge production level, the institutional function of provenance and attribution mechanisms requires re-examination. The current author disclosure system primarily serves ethical transparency purposes and has not been systematically integrated into academic evaluation structures (Emanuele, 2025). To prevent knowledge production from entering a cycle of extreme homogenization and recursive reinforcement, it is necessary to establish manually curated and quality-validated high-fidelity corpora in key disciplinary domains. The core function of such "seed corpora" is to preserve a stable reference benchmark for model training and knowledge accumulation, thereby preventing the continuous decline of human-authored corpora in both proportional representation and epistemic authority. This approach essentially introduces an external anchoring mechanism to counteract potential endogenous drift in training data. It can be led by large public research institutions and the national library system. High-fidelity seed corpora should possess three fundamental attributes: (1) provenance-aware sourcing, operationalized through structured data lineage recording and verifiable metadata infrastructures. In practice, this involves attaching persistent identifiers (e.g., DOIs or dataset-level IDs), maintaining version-controlled data logs, and embedding machine-readable provenance metadata (e.g., W3C PROV standards) that document the origin, transformation history, and ownership of each data element. Such mechanisms enable downstream users and auditors to reconstruct the full lifecycle of the corpus and verify its epistemic reliability. In institutional contexts, this can be implemented via repository-based governance (e.g., curated academic data repositories or national library systems), where ingestion, validation, and update processes are subject to formal audit protocols; (2) measurable quality, assessed through metrics such as factual consistency scores, expert-evaluated coherence, and citation-based

validation; and (3) curated evolution, meaning that while the corpus may evolve over time, its modification and development are subject to strict governance and control (Nature Machine Intelligence, 2024).

From a broader perspective, the proliferation of AI-related academic misconduct reflects the insufficiency of the academic system's capacity to identify and evaluate contributions made by non-human intelligence. Traditional quality-control infrastructures are built upon the assumption that "the author is the cognitive agent," yet the integration of GenAI has destabilized this premise. Research indicates that scholars in practice often struggle to clearly delineate the boundary between their own contributions and AIGC, directly undermining research integrity and transparency (Y. Wu et al., 2025). Therefore, future academic norms must develop a refined taxonomy of contributions that no longer relies on the generalized category of "authorship," but instead explicitly distinguishes among dimensions such as conceptual design, prompt engineering, data validation, result interpretation, and ethical review, thereby providing a clearer basis for accountability (Aburass & Abu Rumman, 2024).

On this foundation, reconstructing the mechanisms through which academic trust is generated becomes crucial. Trust should no longer be anchored solely in the identity of an individual author, but rather shift toward the "process transparency" of the entire knowledge production workflow. This implies the establishment of mandatory, standardized AI usage audit trails that comprehensively document the models employed, the prompts entered, intermediate outputs generated, and key points of human intervention. Such traceability would enable reviewers and readers to assess the extent to which AI influenced the final conclusions and to determine whether human researchers fulfilled adequate supervisory and validation responsibilities. At the level of system implementation, it can be jointly promoted by the education ministries of various countries and universities to incorporate the above-mentioned contribution classification system into the postgraduate training and research integrity courses.

Institutional-level reform must seek a delicate balance between open innovation and clearly defined boundaries of responsibility. On the one hand, outright prohibition or excessive restriction of AI use would suppress its substantial potential to accelerate scientific discovery and promote interdisciplinary integration (Alnaimat et al., 2025). On the other hand, a laissez-faire approach may lead to the homogenization and superficialization of academic output, while exacerbating systemic inequities arising from inherent model biases (Hatos, 2025).

Accordingly, cross-disciplinary collaboration is urgently needed to formulate a dynamically evolving set of "meta-norms." These norms should recognize that AI is not merely a passive tool, but an agentic "cognitive mediator" whose intervention fundamentally reshapes the pathways of knowledge production (Bisenbaev, 2026). Institutional design should encourage responsible innovation, for example, by establishing dedicated ethical review procedures to evaluate high-risk AI-assisted research, and by developing continuous educational programs to cultivate researchers' capacity for "meaningful human control." Such measures would ensure that scholars are able to critically scrutinize AI outputs and avoid falling into automation bias (Mezzadri, 2025).

4.3. Limitations

Although this study constructs a multi-scale diagnostic framework through the integration of multi-source data, several limitations remain. First, with regard to the detection of AIGC, the "AI-likeness" model employed in this study is trained on existing corpora and may be influenced by ongoing model evolution, thereby introducing a risk of misclas-

sification. Accordingly, the results should be interpreted as indicative of trends rather than precise estimates of the true proportion of generated content.

Second, the Common Crawl sample, while substantial, remains limited relative to the full scale of the underlying corpus, and the heterogeneity of web content in terms of domain and structure may affect the generalizability of the findings. Future research could improve robustness through larger-scale or stratified sampling strategies.

Third, retraction data, used here as a proxy for misconduct, may be affected by disciplinary differences in detection capacity, reporting mechanisms, and temporal windows, thereby introducing bias. As such, the analysis reflects "observable risk" rather than the full extent of misconduct. In particular, disciplines with weaker post-publication scrutiny—such as certain experimental or foundational sciences—may exhibit systematically lower observed retraction rates, even when underlying risks are comparable or higher. This creates a potential underestimation of true governance needs in fields showing low NCI values, and introduces a form of detection-dependency circularity: observed misalignment between governance attention and retraction pressure is partly shaped by the very detection infrastructure that governance research aims to strengthen. Sensitivity analyses using alternative proxies, such as normalized retraction rates per 10,000 publications or survey-based misconduct self-reports, would help mitigate this limitation in future work.

5. Conclusions

This study provides the first empirical demonstration that the governance of AI-related academic misconduct is systematically misaligned with the actual distribution of retraction risk across disciplines. By integrating open web corpus analysis, retraction dynamics, and bibliometric evidence, we reveal a clear underrepresentation of governance research attention in the basic sciences relative to their risk exposure (NCI < 0.5), whereas fields such as education receive disproportionate attention (NCI > 29). This disparity reflects the current distributional pattern of governance research within the academic system.

The Normalized Coverage Index (NCI) offers a straightforward yet insightful diagnostic tool for mapping structural mismatches in scholarly attention to AI-related academic misconduct governance. Quantifying the relative alignment between governance research output and observed retraction pressure across disciplines can usefully inform targeted resource allocation and policy design. However, the NCI primarily captures distributions of attention rather than governance effectiveness per se, and its interpretation must be contextualized within specific disciplinary cultures, publication practices, and post-publication scrutiny regimes. Future research could further explore the evolutionary pathways of governance mechanisms across disciplines through qualitative analysis and longitudinal tracking, as well as evaluate the actual effectiveness of related interventions. In addition, cross-national comparative studies would help reveal variations in governance models across different research ecosystems.

Ultimately, safeguarding scientific knowledge in the age of generative AI requires not more generic ethics declarations, but a fundamental reallocation of our scarcest resource—scholarly attention. By diagnosing its current misalignment, this study provides an empirical basis to inform future efforts toward this critical task.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/publications14020027/s1. Tables S1: Classification scheme and mapping rules for ai-related retraction categories in retraction watch database; Table S2: Systematic search strategies for web of science core collection, scopus, and google scholar; Table S3: Descriptive statistics of keyword frequencies in governance research on ai-related academic misconduct (2020–2026); Table S4: Disciplinary normalized coverage index (nci) estimates with bootstrap confidence intervals.

Author Contributions: Conceptualization, H.M.; methodology, H.M. and Z.G.; validation, F.Z. and Z.G.; formal analysis, F.Z. and Z.G.; investigation, H.M., Z.G. and F.Z.; data curation, Z.G.; writing—original draft preparation, H.M.; writing—review and editing, H.M., Z.G. and F.Z.; supervision, H.M.; project administration, H.M.; correspondence, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Project of the National Social Science Fund of China, Grant No. 25BGJ095.

Data Availability Statement: The data supporting this study are derived from a combination of publicly accessible and licensed sources, including Common Crawl, Google Scholar, Scopus, Web of Science Core Collection, and the Retraction Watch Database. Processed data, classification rules, and analysis code can be made available from the corresponding author (mao@upc.edu.cn) upon reasonable request, subject to the licensing restrictions of the original data providers.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
GenAI	Generative Artificial Intelligence
AIGC	AI-Generated Content
LLM	Large Language Model
LLMs	Large Language Models
NCI	Normalized Coverage Index
WoS	Web of Science
WoSCC	Web of Science Core Collection
COPE	Committee on Publication Ethics
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
KDE	Kernel Density Estimation
KS	Kolmogorov–Smirnov
COVID-19	Coronavirus Disease 2019

References

Aburass, S., & Abu Rumman, M. (2024). Authenticity in authorship: The writer's integrity framework for verifying human-generated text. *Ethics and Information Technology*, *26*(3), 62. [CrossRef]

Alnaimat, F., AlSamhori, A. R. F., Hamdan, O., Seiil, B., & Qumar, A. B. (2025). Perspectives of artificial intelligence use for in-house ethics checks of journal submissions. *Journal of Korean Medical Science*, *40*, e170. [CrossRef]

Benítez, T. M., Xu, Y., Boudreau, J. D., Kow, A. W. C., Bello, F., Van Phuoc, L., Wang, X., Sun, X., Leung, G. K., Lan, Y., Wang, Y., Cheng, D., Tham, Y. C., Wong, T. Y., & Chung, K. C. (2024). Harnessing the potential of large language models in medical education: Promise and pitfalls. *Journal of the American Medical Informatics Association*, *31*, 776–783. [CrossRef]

Bisenbaev, A. K. (2026). Scientific artificial intelligence: From a procedural toolkit to cognitive coauthorship. *Philosophies*, *11*, 12. [CrossRef]

Emanuele, E. (2025). Duplicate submission, zero consequences: A reviewer's first-person case study. *Cureus*, *17*(12), e99518. [CrossRef]

Ganjavi, C., Eppler, M. B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G. S., Gill, I. S., & Cacciamani, G. E. (2024). Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: Bibliometric analysis. *BMJ*, *384*, e077192. [CrossRef]

Hanley, H. W. A., & Durumeric, Z. (2024). Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*(1), 542–556. [CrossRef]

Hatos, A. (2025). Between innovation and ethical challenges: The impact of artificial intelligence in social science research. *Sociologie Romaneasca*, *23*, 121–139. [CrossRef]

Jazbec, M., Pàsztor, B., Faltings, F., Antulov-Fantulin, N., & Kolm, P. N. (2021). On the impact of publicly available news and information transfer to financial markets. *Royal Society Open Science*, *8*(7), 202321. [CrossRef]

Jiang, Y., Xie, L., Lin, G., & Mo, F. (2024). Widen the debate: What is the academic community's perception on ChatGPT? *Education and Information Technologies*, *29*, 20181–20200. [CrossRef]

Kanmodi, K. K., Nwafor, J. N., Salami, A. A., Egbedina, E. A., Nnyanzi, L. A., Ojo, T. O., Duckworth, R. M., & Zohoori, F. V. (2022). A Scopus-based bibliometric analysis of global research contributions on milk fluoridation. *International Journal of Environmental Research and Public Health*, *19*, 8233. [CrossRef]

Kay, J., Kasirzadeh, A., & Mohamed, S. (2024). Epistemic injustice in generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 684–697. [CrossRef]

Kendall, G. (2024). When using artificial intelligence tools in scientific publications authors should include the prompts and the generated text as part of the submission. *Journal of Academic Ethics*, *23*, 639–647. [CrossRef]

Kocak, Z. (2024). Publication ethics in the era of artificial intelligence. *Journal of Korean Medical Science*, *39*, e249. [CrossRef]

Kwee, R. M., & Kwee, T. C. (2023). Retracted publications in medical imaging literature: An analysis using the Retraction Watch database. *Academic Radiology*, *30*, 1148–1152. [CrossRef] [PubMed]

Lei, F., Du, L., Dong, M., & Liu, X. (2024). Global retractions due to randomly generated content: Characterization and trends. *Scientometrics*, *129*, 7943–7958. [CrossRef]

Lin, A., Chen, Z., Jiang, A., Tang, B., Qi, C., Zhu, L., Mou, W., Gan, W., Zeng, D., Xiao, M., Chu, G., Peng, S., Wong, H. Z. H., Zhang, L., Zhang, H., Deng, X., Cheng, Q., Zhang, J., & Luo, P. (2026). Navigating academic integrity in biomedical research: The impact of large language models on current practices and future directions. *International Journal of Surgery (London, England)*, *112*(2), 4418–4433. [CrossRef]

Mezzadri, D. (2025). The paradox of ethical AI-assisted research. *Journal of Academic Ethics*, *23*, 2653–2667. [CrossRef]

Mortlock, R., & Lucas, C. (2024). Generative artificial intelligence (Gen-AI) in pharmacy education: Utilization and implications for academic integrity: A scoping review. *Exploratory Research in Clinical and Social Pharmacy*, *15*, 100481. [CrossRef]

Nag, S. N., Roy, A., & Sudhier, K. (2025). Global perspectives on retracted papers in artificial intelligence and machine learning: A bibliometric study. *Global Knowledge, Memory and Communication*. Advanced online publication. [CrossRef]

Nature Machine Intelligence. (2024). Pick your AI poison. *Nature Machine Intelligence*, *6*, 1119. [CrossRef]

Ong, J. C. L., Chang, S. Y., William, W., Butte, A. J., Shah, N. H., Chew, L. S. T., Liu, N., Doshi-Velez, F., Lu, W., Savulescu, J., & Ting, D. S. W. (2024). Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, *6*, e428–e432. [CrossRef]

Pattnaik, M. (2023). Healthcare management and COVID-19: Data-driven bibliometric analytics. *OPSEARCH*, *60*(1), 234–255. [CrossRef] [PubMed Central]

Pellegrina, D., & Helmy, M. (2025). AI for scientific integrity: Detecting ethical breaches, errors, and misconduct in manuscripts. *Frontiers in Artificial Intelligence*, *8*, 1644098. [CrossRef]

Perkins, M., Roe, J., Postma, D., McGaughran, J., & Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*, *22*, 89–113. [CrossRef]

Prifti, K., & Fosch-Villaronga, E. (2024). Towards experimental standardization for AI governance in the EU. *Computer Law & Security Review*, *52*, 105959. [CrossRef]

Pudasaini, S., Miralles-Pechuán, L., Lillis, D., & Llorens Salvador, M. (2024). Survey on AI-generated plagiarism detection: The impact of large language models on academic integrity. *Journal of Academic Ethics*, *23*, 1137–1170. [CrossRef]

Rhodes, C., & Linnenluecke, M. K. (2025). The junkification of research. *Organization*. Advanced online publication. [CrossRef]

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*, 755–759. [CrossRef]

Sridharan, K., & Sivaramakrishnan, G. (2026). Artificial intelligence in the retraction spotlight: Trends, causes and consequences of withdrawn AI literature through a systematic bibliometric review. *Frontiers in Research Metrics and Analytics*, *10*, 1737168. [CrossRef] [PubMed]

Villatte, G., Marcheix, P., Antoni, M., Devos, P., Descamps, S., Boisgard, S., & Erivan, R. (2020). Do bibliometric findings differ between Medline, Google Scholar and Web of Science? Bibliometry of publications after oral presentation to the 2013 and 2014 French Society of Arthroscopy (SFA) congresses. *Orthopaedics & Traumatology: Surgery & Research*, *106*, 1469–1473. [CrossRef]

Wu, F., Gao, J., Kang, J., Wang, X., Niu, Q., Liu, J., & Zhang, L. (2022). Knowledge mapping of exosomes in autoimmune diseases: A bibliometric analysis (2002–2021). *Frontiers in Immunology*, *13*, 939433. [CrossRef] [PubMed]

Wu, Y., Lu, X., & Lin, C. (2025). AI, originality, and attribution: Researchers' perspectives on distinguishing contributions. *Accountability in Research*, *33*(3), 2536817. [CrossRef] [PubMed]

Xu, S., Xu, D., Wen, L., Zhu, C., Yang, Y., Han, S., & Guan, P. (2020). Integrating unified medical language system and Kleinberg's burst detection algorithm into research topics of medications for post-traumatic stress disorder. *Drug Design, Development and Therapy*, *14*, 3899–3913. [CrossRef] [PubMed]

Yao, M., Wei, Y., & Liu, H. (2025). AI practices and ethical concerns: An analysis of undeclared uses of AI in published research articles. *Ethics & Behavior*. Advanced online publication. [CrossRef]

Zhang, G., Xu, Z., Jin, Q., Chen, F., Fang, Y., Liu, Y., Rousseau, J. F., Xu, Z., Lu, Z., Weng, C., & Peng, Y. (2025). Leveraging long context in retrieval augmented language models for medical question answering. *NPJ Digital Medicine*, *8*, 239. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.