A Rapid Investigation of Artificial Intelligence Generated Content Footprints in Scholarly Publications

Gengyan Tang (✉ tanggengyan@outlook.com)  
Sichuan Academy of Social Sciences  https://orcid.org/0000-0003-3221-3134

Sarah Elaine Eaton  
University of Calgary  https://orcid.org/0000-0003-0607-6287

Research Article

Keywords: AIGC footprints, research integrity, academic integrity, quality assurance, integrity education, AI-generated content, ChatGPT, publication ethics, artificial intelligence, large language models, LLM

Posted Date: August 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3253789/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

This study reports on a novel phenomenon observed in scholarly publications. Some research articles unrelated to the field of artificial intelligence (AI) generate content (AIGC) containing confusing phrases such as “As an AI language model...”. We conceptualize this phenomenon as “AIGC footprints”. To provide early evidence, we conducted a small-scale sample investigation by collecting 25 articles. We found that the appearance of AIGC footprints corresponds to the time when the public version of ChatGPT was launched. These 25 articles were published by authors from countries in Central Asia, South Asia, and Africa. Among these authors, there were assistant professors \( (n=5) \), Ph.D. researcher \( (n=6) \), as well as Ph.D. and master’s students \( (n=3) \). Single authors \( (n=16) \) and single affiliations \( (n=23) \) were more common. Analysis of the article content revealed that some authors utilized ChatGPT for literature reviews \( (n=11) \) or idea generation \( (n=11) \). Articles with AIGC footprints are widely distributed across various professional fields, such as Communication and Media Studies \( (n=3) \), Cybersecurity \( (n=2) \), Civil Engineering \( (n=2) \), and Agricultural Technology \( (n=2) \). The 25 articles with AIGC footprints were published in 18 different academic journals. Most of the academic journals did not disclose their APCs on their websites \( (n=11) \), nor were they indexed by Web of Science, Scopus, and DOAJ \( (n=17) \). The emergence of AIGC footprints reflects the potential challenges faced by scholarly publishing and higher education in ensuring quality assurance, as well as indicating potential problems in research integrity. We provide several recommendations, including the development of best research practice guidelines in the context of AIGC, integrating transparent use of AIGC into higher education instruction, and fostering ethical leadership.

Introduction

The advent of AI-generated content (AIGC) has propelled scholarly publishing into a new epoch. Academic journals have been quick to introduce editorial policies in response to the challenges posed by research integrity (Tang 2023a). Although some journals explicitly prohibit authors from employing AIGC, others permit its usage under the condition of appropriate disclosure. In fact, AIGC has sparked intense debates not only in the realm of scholarly publishing (Lund et al., 2023), but also within the fields of academic integrity (Lancaster, 2023) and research integrity (Seth, Bulloch, & Hin, 2023), as well as higher education (Farrokhnia et al., 2023). Researchers are concerned that AIGC may pose more complex challenges.

Some literature takes an optimistic view of the emergence of AIGC. For instance, considering artificial intelligence (AI) tools as an opportunity can enhance the efficiency and quality of scholarly writing, thereby allowing researchers to allocate more time to engage in more meaningful research activities (Huang & Tan, 2023). Another perspective suggests that ChatGPT can assist researchers in handling simulated data, thereby accelerating the drafting process of research articles (Macdonald et al., 2023). Conversely, some authors express concerns about AI tools. Some studies initially questioned whether ChatGPT as an author of scholarly publications violates the four authorship criteria of the International Committee of Medical Journal Editors (ICMJE) (Rahimi and Amin, 2023). Following the introduction of
editorial policies that prohibit ChatGPT from being listed as an author in academic journals, (Thorp, 2023), some researchers have become concerned about issues such as cheating (Tlili et al., 2023), plagiarism (Lo, 2023), and the dissemination of incorrect academic information (Salvagno, Taccone, and Gerli, 2023).

This literature provides many insightful early viewpoints, helping the academic community understand the integrity, ethical challenges, and potential opportunities under the backdrop of AIGC. However, we still lack early evidence to understand the impact of AIGC on research. After experiencing an anecdote¹, we report a novel phenomenon within certain scholarly publications unrelated to AIGC (see Appendix 1). These scholarly publications include perplexing phrases such as “As an AI language model...”—a disclaimer common in ChatGPT outputs (Vincent 2023). We conceptualize this phenomenon as “AIGC footprints” to describe the traces left by the author when using AIGC in the article. Its emergence may reflect more challenging issues faced by research integrity, so we conducted a rapid investigation of this phenomenon to provide some early evidence and serve as a reference for the formulation of editorial policies for academic journals and research integrity policies. We aimed to address the following research question: What are the bibliometric characteristics of articles with AIGC footprints and the journals that publish these articles?

Materials and Methods

Sample selection

Firstly, we conducted a literature search using Google Scholar. The search was performed on June 28, 2023. In order to identify the AIGC footprints present throughout the entire text, while simultaneously excluding articles focused on ChatGPT as the subject of study, we devised a retrieval strategy. Our strategy hinged on utilizing the following phrase: “as an AI language model” – ChatGPT, which resulted in obtaining 71 results.

Subsequently, we established exclusion criteria to filter out articles that did not align with the objectives of our research. These criteria are outlined as follows: 1. non-research articles or those not published in academic journals were disregarded as they lack the essential rigor; 2. inaccessible articles, whose content cannot be reviewed, are excluded from consideration; 3. articles that explicitly state the use of ChatGPT are reasons for exclusion; 4. articles written in languages other than English are excluded because we cannot review them.

Finally, we compared each article, scrutinizing its content against our criteria. After the evaluation, 45 articles failed to meet our research objectives and were consequently excluded from further consideration, resulting in acquisition 25 samples (the sample selection process is illustrated in Fig. 1).

Variables and data collection
We referred to previous research (Oermann et al. 2016; Tang and Jia 2022) on questionable journals and developed a data collection form for observing questionable articles with AIGC footprints, including three dimensions: 1. Author attributes; 2. Article attributes; 3. Journal attributes. Within each dimension, we established 3–4 variables, including: A1. Country of origin of first author; A2. Identity of first author; A3. Number of authors; A4. Number of affiliated institutions; A5. Purpose of using AIGC; A6. Publication date; A7. Professional field; A8. Journal's name; A9. Article processing fees (APCs); A10. Number of databases were indexed (see Table 1).

Data collection was conducted from late June 2023 through early July 2023. For the variables A1, A2, A3, A4, A6, A8, and A9, one researcher was assigned journal websites to review, and the second researcher reviewed the results for accuracy. Data collected from the review of each journal were recorded on the form. During the data collection process for variable A6, it came to our attention that articles numbered 18, 19, 20, 21, and 23 (see Appendix 1) displayed publication dates of 2021 on the journal websites. However, upon examination of their DOI numbers, it became evident that these articles were actually published in 2023. To determine their true publication dates, an analysis of the source code of the website was performed, leading us to conclusively establish that these articles were published in June 2023.

We conducted a manual analysis of the AIGC text to extract data related to the variables A5. To illustrate this process, we examined the following excerpt: “As an AI language model, I don’t have direct access to the most recent studies. However, I can provide you with some general information on the role of the digital economy in achieving sustainable development based on commonly discussed themes in the field.” Based on this text, it can be inferred that the author utilized ChatGPT to gather literature pertaining to the importance of the digital economy in achieving sustainable development. Consequently, we categorized this method of data acquisition as “literature review.” To ensure accuracy and consistency in the categorization process, one team member assigned labels while another team member verified them.

Moreover, we employed ChatGPT-4 to extract data pertaining to the variables A7. Since certain articles lacked literature categorization, we utilized ChatGPT-4 for subject classification in academic literature, as it demonstrated satisfactory performance in previous testing (Datta, 2023). We provided the following instruction to ChatGPT-4 for literature classification: “After reviewing the abstract information, categorize it according to the most representative professional field.” Lastly, for the variables A10, we considered three primary databases: Web of Science, Scopus, and DOAJ (Directory of Open Access Journals). We searched for journal names within these databases and cross-referenced the ISSN numbers to determine whether the journals were indexed in the respective databases (Somoza-Fernández, Rodríguez-Gairín, & Urbano, 2016).
### Table 1
Variable name and data source

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Variables (abbreviations)</th>
<th>Data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author attributes</td>
<td>Country of origin of first author (A1)</td>
<td>Journal websites</td>
</tr>
<tr>
<td></td>
<td>Identity of first author (A2)</td>
<td>Journal websites</td>
</tr>
<tr>
<td></td>
<td>Number of authors (A3)</td>
<td>Journal websites</td>
</tr>
<tr>
<td></td>
<td>Number of affiliated institutions (A4)</td>
<td>Journal websites</td>
</tr>
<tr>
<td>Article attributes</td>
<td>Purpose of using AIGC (A5)</td>
<td>The text of AIGC footprints</td>
</tr>
<tr>
<td></td>
<td>Publication date (A6)</td>
<td>Journal websites</td>
</tr>
<tr>
<td></td>
<td>Professional field (A7)</td>
<td>ChatGPT-4</td>
</tr>
<tr>
<td>Journal attributes</td>
<td>Journal’s name (A8)</td>
<td>Journal websites</td>
</tr>
<tr>
<td></td>
<td>APCs (A9)</td>
<td>Journal websites and real-time exchange rate tool</td>
</tr>
<tr>
<td></td>
<td>Number of databases were indexed (A10)</td>
<td>Web of Science, Scopus, DOAJ</td>
</tr>
</tbody>
</table>

**Note:** APCs are converted to USD (United States Dollars) using real-time exchange rate tool ([https://www.xe.com/currencyconverter/](https://www.xe.com/currencyconverter/)) on June 28th.

### Data analysis

The data for all variables was statistically analyzed using counts. Microsoft Excel and Prism 9.0 were utilized as the software for both data analysis and plotting.

### Results

#### Author Attributes

##### Country of origin of first author

The findings indicated that scholarly articles with AIGC footprints are primarily published by authors from five different countries in Asia and Africa (see Fig. 2). Among them, authors from Uzbekistan ($n = 8$) contribute the highest number of articles, followed by India ($n = 7$) and Algeria ($n = 4$).

##### Identity of first author

The identity of the first author exhibits considerable diversity (see Fig. 2). Approximately 24% of the articles describe the first author as a "Ph.D. researcher" ($n = 6$), while 20% of the articles designate them as an "assistant professor" ($n = 5$). Some articles also identify the first author as a “master researcher” ($n$...
a “Ph.D. student” \((n = 2)\), or a “senior teacher” \((n = 2)\). Additionally, the titles of “associate professor” \((n = 1)\) and “independent researcher” \((n = 1)\) are sporadically mentioned. The identity of the first author is not stated in four articles.

**Number of Authors**

We also calculated the number of authors included in each article (see Fig. 2). 64% of the articles were completed by a single author \((n = 16)\), while 24% were collaborations between two authors \((n = 6)\). Only 12% of the articles were collaborations involving three or more authors \((n = 3)\). This might suggest that single authors are the predominant contributors to articles with AIGC footprints.

**Number of affiliations**

Single affiliations overwhelmingly dominate (see Fig. 2). Approximately 88.5% of the articles are authored by individuals affiliated with a single affiliation \((n = 23)\), with only two articles representing collaborations between authors from two affiliations \((n = 2)\). This might indicate that the degree of inter-institutional collaboration in articles with AIGC footprints is relatively low, with the majority of collaborations occurring within a single affiliation.

**Article attributes**

**Purpose of using ChatGPT**

Figure 3 illustrates that 44% of the articles potentially utilize ChatGPT to generate research ideas \((n = 11)\). These ideas may involve requesting ChatGPT to provide arguments or evidence for their studies or asking for appropriate research methods. Similarly, another 44% of the articles use ChatGPT for the purpose of gathering literature and composing literature reviews \((n = 11)\). Furthermore, a small portion of articles request ChatGPT to generate images \((n = 1)\), analyze data \((n = 1)\), or access data from databases \((n = 1)\).

**Publication date**

By the end of 2022, OpenAI released the public version of ChatGPT, and within two months, its monthly active users exceeded 100 million people (Wu et al., 2023). We observed that articles with AIGC footprints roughly align with the timeline of ChatGPT’s launch. Approximately 24% of the articles were published in December 2022, which is almost synchronous with the release of the public version of ChatGPT. In June 2023, the number of articles with AIGC footprints reached its highest value within the statistical interval, totaling 11 articles. Additionally, in April 2023 \((n = 5)\) and May 2023 \((n = 3)\), articles with AIGC footprints were also published.

**Professional field**

Figure 3 demonstrates that 25 articles with AIGC footprints are distributed across 20 different professional fields. Among them, the field with the highest number of published articles is Communication and Media Studies \((n = 3)\), followed by Cybersecurity \((n = 2)\), Civil Engineering \((n = 2)\), and Agricultural Technology \((n = 2)\). Furthermore, articles with AIGC footprints were discovered in professional
fields such as Education Technology \((n = 1)\), Law \((n = 1)\), Literary Studies \((n = 1)\), and Gender Studies \((n = 1)\).

**Journal attributes**

**Journal’s name**

In the investigation, we discovered that 25 articles with AIGC footprints were published in 18 different academic journals. The most prolific publisher was the *International Journal of Modern Agriculture and Environment* \((n = 5)\), followed by the *International Journal of Advances Engineering and Civil Research* \((N = 3)\), and the *European Chemical Bulletin* \((n = 1)\).

**APCs**

According to the *Principles of Transparency and Best Practice in Scholarly Publishing* (The Committee on Publication Ethics 2022), academic journals are expected to disclose their author fees. However, approximately 61.1% of the academic journals \((n = 11)\) that published articles with AIGC footprints did not adhere to this best practice. These academic journals either did not provide any information about APCs on their websites or made their public statements inaccessible. The remaining academic journals charged APCs below 50 USD \((n = 7)\), with three of them setting APCs at 20 USD, and collectively publishing three articles with AIGC footprints. One academic journal, namely *Web of Scholars: Multidimensional Research Journal (MRJ)*, claimed to have zero APCs.

**Number of databases were indexed**

Through database searches, we found that all academic journals \((n = 18)\) that published articles with AIGC footprints were not indexed in Web of Science. Only one academic journal, the *European Chemical Bulletin*, was indexed in both DOAJ and Scopus, while the remaining academic journals \((n = 17)\) were not indexed in either DOAJ or Scopus.

**Discussion**

We have come across certain instances in scholarly publications that have drawn our attention due to the presence of perplexing phrases, such as “As an AI language model...”. This phenomenon has not received much discussion in previous literature. In order to enhance the academic community’s understanding of this novel phenomenon, we conducted a rapid investigation using a small-scale sample, which provided the early evidence. We conceptualized this phenomenon as “AIGC footprints”. Through meticulous examination, of some instances, we discovered that the authors failed to declare the usage of ChatGPT in their articles, and the appearance of this phrase lacked semantic coherence with the surrounding context. This suggests that authors might have copied ChatGPT’s disclaimer when reproducing its generated content, and these oversights went unnoticed and unaddressed by peer reviewers and editors. AIGC footprints represent the remnants of text left by authors who have employed AI tools in their scholarly...
publications, may expose a systematic neglect within the quality control processes of scholarly publishing.

Through database retrieval, we have identified 25 research articles containing this phrase. Subsequently, we investigated the bibliographic characteristics of these 25 articles. We found that articles with AIGC footprints are primarily authored by researchers from Central Asia, South Asia, and Africa. Among these authors, there are both graduate students who may have just begun their research work and assistant or associate professors who have already obtained professional titles. Single authors and single affiliations are relatively common in the articles with AIGC footprints. While further evidence is required to confirm the reliability of these findings, our discoveries reflect that quality assurance issues in higher education and academic publishing may become more challenging within the AIGC context. It is perhaps unexpected that alongside the complex challenges faced by higher education regarding plagiarism, contract fraud, and data falsification, the impact of artificially generated text may pose a threat to scholarly publication rigour. It is impossible to know the motives of authors who use AIGC without declaring its use. Puzzling through these motives presents a possible avenue for further investigation.

We also reviewed the content of articles with AIGC footprints. The appearance of AIGC footprints in scholarly publications aligns with the release time of the public version of ChatGPT. Furthermore, by reviewing the inserted AIGC snippets in the articles, we infer that most authors use ChatGPT for the purpose of gathering literature, composing literature reviews, or providing research ideas. Only a small number of authors employ ChatGPT for image generation or data analysis. AIGC footprints have currently infiltrated various disciplines such as Education, Media studies, Literary studies, Economics, Engineering, and Chemistry. Although it is currently uncertain whether the occurrence of AIGC footprints in scholarly publications is an incidental phenomenon, its presence emphasizes the importance of redefining criteria for academic misconduct. In other words, we may require a series of guidelines to distinguish between different purposes of using AIGC and whether such behavior violates research integrity and publishing ethics.

Lastly, we analyzed the attributes of academic journals publishing articles with AIGC footprints. In our small sample, the academic journals publishing articles with AIGC footprints have relatively low APCs or do not declare any APCs. This finding is consistent with some conclusions from studies on questionable journals (although the research in this field lacks sufficient evidence) (Xia, 2015). Additionally, a few academic journals have published multiple articles with AIGC footprints. Except for one journal indexed in both DOAJ and Scopus, the remaining academic journals are not indexed in Web of Science, DOAJ, or Scopus, indicating their potential deviation from the best practices of scholarly publishing and the potential challenges in ensuring the quality of scholarly publishing. However, more evidence is needed to substantiate this claim. Nevertheless, our investigation of academic journals raises concerns that the quality assurance issues in scholarly publishing may become more severe when quality issues present in predatory questionable journals are compounded by AI generated-text.
Thus, when discussing AIGC and research integrity, our attention should not solely be focused on authors’ questionable conduct, but also on improving scholarly publishing practices and daily research procedures and quality assurance. This approach aims to establish a comprehensive system that enhances the quality, relevance, and dependability of all research endeavors, to avert potential harm (Nature 2019). Urgent action is necessary from both governmental entities and the academic community to establish research practices pertaining to AIGC. Swift integration of these practices into integrity education at universities is imperative. Through sustained engagement in discussions surrounding research integrity and AIGC with students, early-career researchers, and senior faculty members, we can foster their capacity to uphold research integrity, transcending the mere cultivation of virtues. This includes the adherence to best research practices. The transparency of AIGC usage should also be given attention. Although some literature has already discussed this issue (Hosseini, Resnik, and Holmes 2023), and some academic journals have incorporated policies requiring authors to declare their use of AIGC (Hosseini, Rasmussen, and Resnik 2023), there is a need to extend this policy to higher education in order to encourage students to cite and acknowledge their use of AIGC when writing or submitting academic papers (Kumar et al. 2023). This can contribute to fostering their awareness of academic integrity. Furthermore, it is worth considering the inclusion of the challenges to research integrity brought by AIGC footprints into the broad-based institutional priorities (Christensen Hughes and Eaton 2022). We call upon all stakeholders to participate in the efforts to make the use of AIGC more transparent, fostering ethical leadership and ensuring that educators and administrators uphold the highest standards of education and publishing quality, leading students to embrace the forthcoming AIGC epoch.

**Limitations**

Our rapid investigation provides some early evidence regarding AIGC footprints, but the very small sample size may weaken the reliability and generalizability of the research findings. Additionally, we did not focus on those articles that may have erased AIGC footprints before publishing. One possible scenario is that authors, upon careful review of their article’s content prior to publish, may remove AIGC footprints without providing any declaration, leading readers to believe the content is original. Certainly, further research is required to provide more comprehensive and credible evidence. We acknowledge that this study has significant limitations, but we will continue to retrieve research articles with AIGC footprints as their number is likely to be continuously growing.

**Conclusion**

This study reports a novel phenomenon observed in scholarly publications, conceptualized as “AIGC footprints.” AIGC footprints refer to instances where authors have copied text from AI tools without reviewing its content, resulting in the presence of this text in scholarly publications. We investigated 25 articles with AIGC footprints and found that their publication dates align with the release of the public version of ChatGPT. These scholarly publications were authored by researchers from five countries, including both professors and graduate students. The prevalence of single authors and single affiliations...
suggests a possible lack of research collaboration in this type of publication. Upon examining the content of the articles, we found that most authors used ChatGPT for the purpose of gathering literature, composing literature reviews, or generating research ideas. Articles with AIGC footprints were widely distributed across professional fields such as Education, Media studies, Literary studies, Economics, Engineering, and Chemistry. Apart from one academic journal, the academic journals publishing such articles were not indexed by Web of Science, Scopus, and DOAJ. AIGC footprints highlight the quality assurance challenges faced by higher education and scholarly publishing, as well as potential academic integrity issues. We call for the development of best research practice guidelines specifically targeting AIGC, encouragement of students and educators in higher education to use AIGC more transparently, and the exercise of ethical leadership to ensure that AIGCs contribute meaningfully to research.

**Declarations**

**Notes**

1. In late June 2023, while reviewing a fellow student's academic paper, TGY encountered a perplexing phrase, “As an AI language model...”. Recognizing this phrase as a potential “digital watermark,” akin to an unintentionally left trace, TGY became concerned that researchers might inadvertently forget to remove it. Having maintained an interest in questionable journals, TGY suspected that this phrase might appear in some low-quality scholarly publications. Consequently, he employed this phrase as a search term in a database and discovered several articles containing it.

2. In accordance with the draft editorial policy of *Accountability in Research*, we hereby declare the following: “To collect research data, TGY utilized OpenAI ChatGPT-4 on June 29th, 2023, at 8:37 PM UTC+8. The following prompt was used during data collection: ‘After reviewing the abstract information, categorize it according to the most representative professional field.’ The generated text was not copied verbatim but was entered into a form we created based on ChatGPT-4’s responses.”

3. Our initial manuscript submission took place on June 25th, 2023, during which we identified 17 articles containing AIGC footprints. After the manuscript was returned for revisions, we subsequently identified an additional 8 articles. This suggests that the presence of AIGC footprints in articles may continue to increase over time.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Data Availability**

The data that support the findings of this study are available at: https://doi.org/10.5281/zenodo.8227899.

**Funding**
The authors of this study did not receive any funding neither from industry nor from academic institutions.

**Ethical considerations**

This article does not contain any studies with human participants or animals performed by any of the authors.

**References**


Figures
Figure 1

Flow diagram of sample selection
Figure 2

Statistical analysis graphs of variables A1-A4

**Note:** A1. Country of origin of first author; A2. Identity of first author; A3. Number of authors; A4. Number of affiliated institutions.
Figure 3

Statistical analysis graphs of variables A5-A7

Note: A8. Journal’s name; A9. Article processing fees (APCs); A10. Number of databases were indexed.
Figure 4

Statistical analysis graphs of variables A8-A10

**Note:** A5. Purpose of using AIGC; A6 Publication date; A7. Professional field.

**Supplementary Files**
This is a list of supplementary files associated with this preprint. Click to download.

- Appendix1.docx