Croat Med J. 2024;65:93-100
https://doi.org/10.3325/cmj.2024.65.93

Exploring ChatGPT's abilities in medical article writing and peer review

Gültekin Kadi, Mehmet Ali Aslaner

Department of Emergency Medicine, Gazi University Faculty of Medicine, Ankara, Turkey

Aim To evaluate the quality of ChatGPT-generated case reports and assess the ability of ChatGPT to peer review medical articles.

Methods This study was conducted from February to April 2023. First, ChatGPT 3.0 was used to generate 15 case reports, which were then peer-reviewed by expert human reviewers. Second, ChatGPT 4.0 was employed to peer review 15 published short articles.

Results ChatGPT was capable of generating case reports, but these reports exhibited inaccuracies, particularly when it came to referencing. The case reports received mixed ratings from peer reviewers, with 33.3% of professionals recommending rejection. The reports' overall merit score was 4.9 ± 1.8 out of 10. The review capabilities of ChatGPT were weaker than its text generation abilities. The AI as a peer reviewer did not recognize major inconsistencies in articles that had undergone significant content changes.

Conclusion While ChatGPT demonstrated proficiency in generating case reports, there were limitations in terms of consistency and accuracy, especially in referencing.

Received: November 30, 2023

Accepted: March 28, 2024

Correspondence to:
Gültekin Kadi
Department of Emergency Medicine
Gazi University Faculty of Medicine
Emniyet Mahallesi, Yenimahalle
06560 Ankara, Turkey
gultekinkadi@gazi.edu.tr

The rapid advance of artificial intelligence (AI) technologies is positively affecting many fields of life, including health, education, and finance (1-4). AI-based natural text processing software (natural language processing [NLP]) is able to produce logical and consistent material by rapidly processing large quantities of data. This software, used for generating text with entertaining and interesting subject matter, has recently begun affecting academic publishing. The effects of AI-based text processing software are being increasingly felt in areas such as content production, expansion, and condensation. Generative Pre-trained Transformer (ChatGPT) launched by OpenAI in November 2022 is currently the most employed and investigated NLP software in this area (5).

NLP can make the process of publishing scientific research more efficient by accelerating text production and organization processes. In addition, AI-based systems may be used in areas such as rapid article screening and quality control to fast-track the article review process.

However, the use of this technology also may lead to ethical transgressions and raise concerns regarding the originality and quality of the articles (2,6-9). In particular, it may hinder the development of thinking and writing skills among students and researchers and thus restrict critical thinking. For example, it may lead to misinterpretations in situations that require genuine experience and accumulated knowledge, such as peer review process. We are aware of no studies in which articles entirely written by AI software were peer-reviewed by human experts, or in which published articles were peer-reviewed by AI. The aim of this study is to evaluate the quality of ChatGPT-generated case reports by submitting them to peer review by human experts and assess the ability of ChatGPT to peer review medical articles.

MATERIAL AND METHODS

This study was performed between February and April 2023. In the first part of the study, ChatGPT was prompted to generate 15 case reports from scratch, which were subsequently reviewed by emergency medicine professionals. In the second part of the study, ChatGPT was asked to review 15 previously published short articles. For this purpose, we searched Q1 journals in the international Science Citation Index Expanded (SCI-E) in the field of emergency medicine for articles published between 2017 and 2020, and individually examined 645 case reports and short articles. Due to the character (approximately 2000 words for the query and 750 for the output) and visual processing limitations of ChatGPT, 15 case reports and 15 short papers with lower word counts and not including tables and figures were selected. The flowchart of the study procedure is shown in Figure 1.

For the first part of the study, only the titles and keywords of the selected case reports were collected, and the following standard query template (Prompt 1) was entered into ChatGPT 3.0 (Supplemental material 1).

Prompt 1

Topic: "Title of original case report"

Context: A case report to be sent to an academic journal

Sections: Title, abstract, introduction, case, discussion, and references

Requirements: Include in-text citations where appropriate for at least three references from real published original articles

Language: Academic

Tone: Formal

Keywords: "from the original case report"

Total word count: At least 750 words.

When the software reached the maximum word limit and stopped writing, a simple second instruction was given: "*Continue where you left off.*" Fifteen rewritten case reports were sent for peer review to 30 emergency medicine specialists with academic experience, each report to two reviewers (Supplemental Material 2: Peer review form for case report). The evaluation form and case reports were sent by email to the reviewers, who were asked to assess the text entirely in line with a normal review procedure. The reviewers were first asked about their Google Scholar h-index, years of experience in emergency medicine, and the number of manuscripts peer-reviewed last year. Then they answered specific questions about each article section with answers on a three-point scale (good, fair, and poor). They also rated the overall article in terms of originality/novelty, significance of content, quality of presentation, scientific soundness, and interest to the readers. Furthermore, they were asked to recommend a target journal for publishing of the article (low-ranking non-SCI-E, high-ranking non-SCI-E, low ranking SCI-E, high-ranking SCI-E); rate the overall merit of the article on a scale 1-10; and give the final recommendation (reject, major revision, minor revision, accept). Additional space was left for comments. The reviewers were given no information concerning whether the research involved AI. When all the reviewers had completed the forms, a second form was issued, and the participants were asked whether the reports had been written by a human or by AI software. The answers were provided on a three-point scale (human, not sure, AI).

The case reports were further evaluated by AI Text Classifier (https://platform.openai.com/ai-text-classifier)_provided by OpenAI. "Possibly or likely AI-generated" results were classified as AI-generated, and "very unlikely or unlikely" results as human-generated, with the "unclear" option also being available. Finally, the case reports underwent plagiarism check (http://www.ithenticate.com).

In the second part of the study, ChatGPT was asked to review 15 previously published, short research articles. Since ChatGPT version 4.0 was launched at the end of the second part of the study, the review process continued with the new version. No change was made to three of the 15 short articles. In order to determine how the AI would respond in the event of errors, we made some changes to 12/15 articles. In two articles, the title was altered in such a way as to create an inconsistency between it and the article content; in two articles, the methodology section was partly/completely altered; in two articles, numeric data

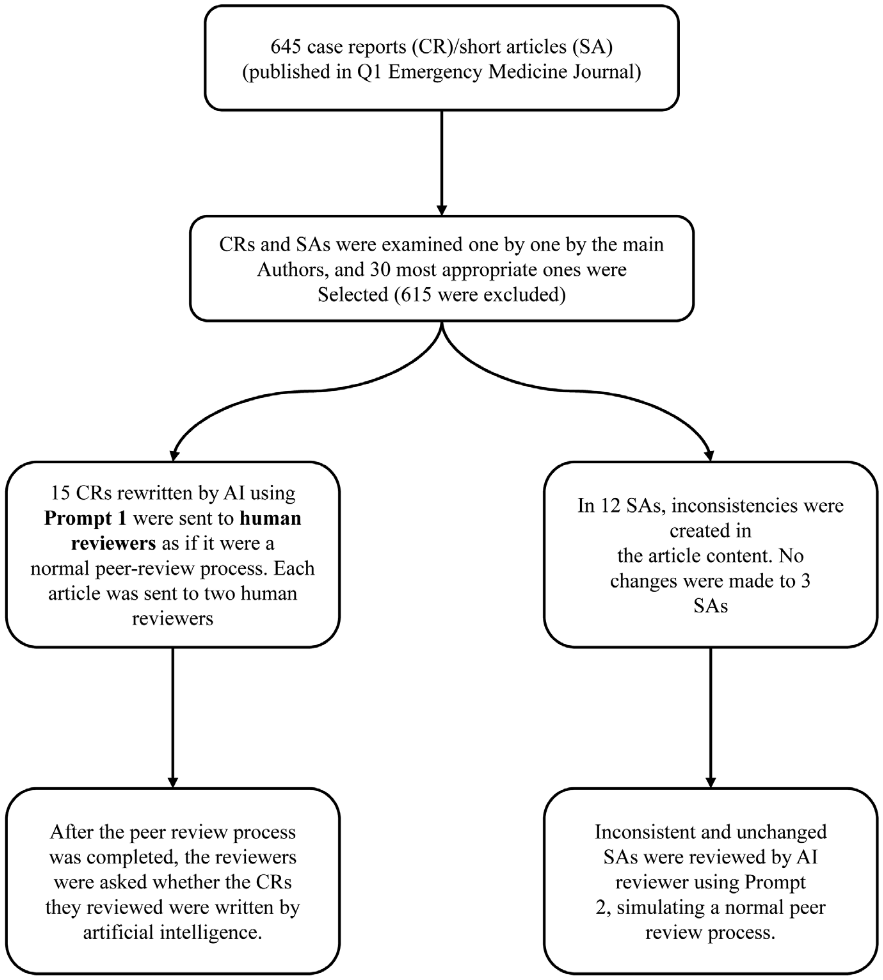
645 case reports (CR)/short articles (SA)
(published in Q1 Emergency Medicine Journal)
CRs and SAs were examined one by one by the main Authors, and 30 most appropriate ones were Selected (615 were excluded)
15 CRs rewritten by AI using **Prompt 1** were sent to **human reviewers** as if it were a normal peer-review process. Each article was sent to two human reviewers
In 12 SAs, inconsistencies were created in the article content. No changes were made to 3 SAs
After the peer review process was completed, the reviewers were asked whether the CRs they reviewed were written by artificial intelligence.
Inconsistent and unchanged SAs were reviewed by AI reviewer using Prompt 2, simulating a normal peer review process.

FIGURE 1. Flowchart of study.

partly or entirely inconsistent with the results were created; in two articles, the references were entirely altered or removed; and in two articles, minor/major grammatical errors were created. The newly produced articles were submitted to the review process with the standard query template (Prompt 2; Supplemental material 3 and Supplemental material 4). The study was approved by the Gazi University Ethics Committee, Ankara (2023-482)

Prompt 2

Review the following brief article for its suitability for publication in an academic journal and give a score of good, fair, or poor for each of the following aspects. For recommendation, give a decision, which can be reject/minor revision/major revision/accept. If the article is not rejected, give a score for the target journal, which can be a low ranking non-SCI-E journal, a high ranking non-SCI-E journal, a low ranking SCI-E journal, or a high ranking SCI-E journal.

Title

Introduction

Methodology

Results and Data analysis

Discussion

Conclusion

References

Originality / Novelty

Significance of Content

Quality of Presentation

Scientific Soundness

Interest to the readers

Ethics

Writing quality

Recommendation

Target journal (if not rejected)

"Title and main text of the manuscript"

Statistical analysis

Continuous variables are expressed as means and standard deviations, while categorical variables are expressed as numbers and frequencies. Differences between the reviewers in terms of years of experience and opinions regarding the source of case reports were assessed with a Kruskal-Wallis test. Statistical analysis was performed with SPSS, version 22 (IBM., Armonk, NY, USA).

RESULTS

Thirty emergency medicine professionals peer-reviewed the case reports. The mean length of time spent by the reviewers in the field of emergency medicine was 11.4 ± 5.6 years, and their mean h-index was 5.4 ± 4.1. The mean number of articles reviewed by them in the previous year was 7 ± 6.7.

Case reports generated by ChatGPT

Each case report written by ChatGPT was evaluated by two reviewers. When individual sections were reviewed, the following percentage of reviewers rated the individual sections as good: 66.7% for titles, 33.3% for the references, 50% for the abstract, 36.4% for the introductions, 7% for the case sections, 20% for the discussion sections, and 50% for the conclusions. Examination of the references produced by ChatGPT showed that 65% were non-existent, and that 87% of the valid references were older than 10 years. The titles of these unreal references were consistent with the case reports, but when accessed using the DOI, a paper on a completely different topic appeared. The reviewers found that the majority of the references that appeared did not actually exist.

Overall, 33.3% of the reviewers rated the case reports as good in terms of quality of presentation, while 43.3% rated them as poor in terms of scientific soundness (Table 1). The mean overall merit of the case reports was 4.9 ± 1.8 out of 10. A third of reviewers recommended rejection as a final decision, and two case reports were rejected by both reviewers. Another third recommended major revisions. In terms of the target journals suggested by the remaining 20 physicians, 11 reviewers recommended low ranking (LR) non-SCI-E, 5 recommended HR non-SCI-E, 3 recommended LR SCI-E, and 1 recommended HR SCI-E. In addition, 13.3% of the reviewers thought that the case reports had been written by AI, and 30% that they were human in origin, while 56.7% were uncertain. Only 20% of the reviewers evaluated the case reports as "of interest to the readers" and 27% rated them as good in terms of ethics.

AI Text Classifier reported 93.3% of the case reports to be written by humans, while for 6.7% it reported to be "unsure." Although the reviewers who recognized the use of AI had more years of experience and greater h-index, the differences were not significant (Table 2). The case reports' mean plagiarism score was 17.47 ± 7.34. Ten reviewers made additional comments regarding the case reports (Supplemental Material 5: Comments). Five of these concerned incorrect references. One referee commented that the study was not appropriate to be published as it did not provide any information about patient consent.

Short articles reviewed by ChatGPT

In the second part of this study, ChatGPT rated three original short articles previously published in a high-impact journal in which no alterations were made as generally good. It recommended major revisions for two out of these articles, and recommended that these could be published in a journal in the LR SCI-E index. It suggested a minor revision to the third, deeming it suitable for publication in an HR non-SCI-E index.

Two articles contained titles that had been altered so as to be inconsistent with the content. ChatGPT detected no discrepancy between the title and content and evaluated the titles as good. In two other articles, the introduction section was altered so as to conflict with the text content. ChatGPT regarded these as generally good and recommended publication in HR non-SCI-E journals following minor revision. Although the software described the two articles subjected to partial and total changes in terms of methodology as generally good, it reported the methodology as fair, and recommended that these could be published in HR non-SCI-E/LR SCI-E journals after correction. Two articles contained minor and major typographical errors. The software described these articles as generally good but the writing as poor, and recommended major revision for both. The software described the two articles whose references had been altered and entirely removed as good in terms of reference evaluation and recommended publication following revision (Table 3).

DISCUSSION

Our results showed that with adequate prompts, ChatGPT can create appropriate introduction, case report, and discussion sections of a scientific case report. However, it also cited unreal and old sources in the references section. The review capacity of NLP was weaker than its ability to produce a written text. The software failed to detect major changes in articles in which the content was significantly altered, it was unable to fully evaluate the integrity of the text in a holistic manner, and even considered the references appropriate in an article with no references at all.

Overall, 66.6% of the reviewers evaluated the titles of the case reports produced by ChatGPT as good, and this was the report section that received the highest score. The title is an important part of an academic text as it ensures a favorable first impression and provides information about the study content (10). On the other hand, when prompting ChatGPT to generate case reports, we used the titles and key words of previously published articles. The software did not create the titles itself. It made some minor changes, but usually employed the original titles. These therefore elicited good scores from the reviewers. This is important as it confirms that the referees' evaluations were appropriate.

The reviewers' evaluations of the introduction section, which contains more original content, showed that the NLP produced monotonous, repetitive texts. Due to the

TABLE 1. Peer review ratings of case reports generated by ChatGPT by human reviewers (N = 30)

	No. of reviewers who rated the article in terms of						
	originality/ novelty	significance of content	quality of presentation	scientific soundness	interest to the readers	ethics	writing quality
Good	7	9	10	2	6	8	9
Fair	20	15	10	5	17	9	14
Poor	3	6	10	13	7	13	7

TABLE 2. The relationship between the reviewers' experience and opinions regarding the source of case reports

Experiences*	Human	Not sure	AI	*P*
Years of experience in emergency medicine	12.44 (7.601)	9.82 (3.877)	15.50 (6245)	0.284
h-index†	6.67 (4.610)	4.12 (2.395)	8.25 (7.365)	0.413
Peer review per year	7.44 (9002)	7.18 (5.919)	5.25 (4.717)	0.655

*Mean (standard deviation).
†Based on Google Scholar.

TABLE 3. Peer review ratings of previously published short articles created by ChatGPT, numbers indicate the count of short articles (N = 15)

	Title	Introduction	Methodology	Results and data analysis	Discussion	Conclusion	References	Originality/ novelty	Significance of content	Quality of presentation	Scientific soundness	Interest to the readers	Ethics	Writing quality
Good	13	12	7	10	14	14	14	7	14	14	10	14	15	13
Fair	2	3	8	5	1	1	1	8	1	0	5	1	0	0
Poor	0	0	0	0	0	0	0	0	0	1	0	0	0	2

NLP's limited ability to produce original and up-to-date texts, the peer reviews resulted in low scores.

Overall, 67% reviewers evaluated the case reports as fair in terms of originality/novelty, 50% in terms of content, 34% in terms of quality of presentation, and 50% in terms of scientific soundness. Referee evaluations in these sections are generally powerful predictors of acceptance or refusal (11). Zhu et al reported that AI can write an in-depth discussion, and that the review article created can represent a useful electronic encyclopedia, but not useful scientific literature (12). On the other hand, Dwivedi et al emphasized that "the absence of originality" detected during their experience of article production using ChatGPT was very important in the context of creative writing, although perhaps unimportant for routine text creation functions (4). The accuracy, originality, and academic integrity of NLP in academic publishing have been described as open to debate and requiring improvement (13).

In order for articles to attract readers' interest, they should be consistent with the literature and up-to-date, and the authors should make their language and style clearly accessible to the user (7). However, the referees in our study stated that the articles written by ChatGPT were monotonous, not up-to-date, and exhibited variations in style and a lack of fluidity. On the other hand, some authors prefer to employ AI, even grant it the status of author (14,15).

Some authors suggested that the use of AI-generated texts in academic publishing will increase the number of poor-quality or plagiarized articles, erode trust among the academic community, and lead to ethical concerns (16,17). Another ethical issue relates to the legal framework, such as determining who is to blame when the AI doctor makes a medical mistake (18).

Salvagno et al stated that human input in meta-analyses and systematic texts requiring complicated writing processes remains essential, but that ChatGPT can be used as an editing tool (19). In the present study, the referees rated the writing quality of the case reports, which require a simpler writing process, as good or fair in 77% of the cases.

The mean overall merit score for all the cases produced by NLP was 4.9 out of 10. Such a low rating explains the high rejection rate by the referees.

The mean plagiarism score in the present study was 17.47 ± 7.34. No consensus and no definite rules exist regarding the acceptable level of plagiarism in academic publications. However, a similarity figure below 15%-20% is usually acceptable to journals. Aydın et al determined a plagiarism rate of 40% in academic papers written by NLP (20,21).

In terms of responses to the question of authorship, four reviewers stated that the texts might have been produced by AI, although 17 thought they could not be sure. The reviewers who recognized AI had more years of experience and a higher h-index, but the difference was not significant, possibly due to small sample size. ChatGPT AI Text Classifier reported that 93.3% of the same case reports could have been drafted by humans. In effect, the NLP stated that almost all the texts it wrote were human in origin. The inconsistency of the results obtained from different AI detection tools reveals the inadequacy of current methods for accurately detecting scientific abstracts created by AI (9). Reviewers correctly identified 68% of abstracts created by AI (22), although these rates may change as technology advances or with the use of different AI detection software (23).

Although 4/15 articles contained obvious errors that should have led to rejection, NLP failed to detect these or report them negatively in the review. It recommended minor revision for half of these articles and major revision for the other half. It also rated the references as "good" in an article (short article 12) from which the references had been entirely removed. This indicates that ChatGPT can comment on a non-existent text section. During the entire peer review, it rated only those articles that contained typos or grammatical errors as poor. This indicates that NLP is unable to grasp the logical construction, continuity, and integrity of the texts, but does detect typographical and grammatical errors. Elsewhere in the literature, NLP software such as ChatGPT has been reported to accelerate the writing process, improve the main lines and details of a paper, and to enhance the writing style (24).

A limitation of the study is that only 15 case reports and 15 short articles were examined in the study. Increasing the sample size and using reports from different disciplines may increase the generalizability of the findings. The AI programs used are developing and updated over time. This and similar studies reveal repetition with different software.

In conclusion, we believe that with its current technological infrastructure, ChatGPT is not capable of producing original and academic text, but that it can be used for text editing. In the future, it may become necessary to use plagiarism software that can detect AI in academic publishing.

Acknowledgments Use the link below to access the articles and all other data used in the study: https://osf.io/hxa6q/?view_only=1012c7135f8b4a4f8826c024b947056c. The authors thank the following peer reviewers who participated in the study: A.Y., MD; A.B.O., MD; A.A., MD; A.B., MD; A.Ç., MD; A.Ş., MD; A.O., MD; A.K., MD; A.K., MD; A.A.Ç., MD; B.Ü., MD; B.İ., MD; Ç.K., MD; E.Ö.İ., MD; E.K., MD; F.B., MD; İ.K., MD; M.İ., MD; M.Y., MD; M.E.Ç., MD; O.D.G., MD; Ö.C., MD; Ö.G.Ç., MD; S.G., MD; S.P., MD; S.İ., MD; T.Y.K., MD; V.A., MD; Y.E.Ö., MD; Ç.Y., MD.

Funding None.

Ethical approval granted by by the Gazi University Ethics Committee, Ankara (2023-482).

Declaration of authorship Both authors conceived and designed the study; both authors acquired the data; both authors analyzed and interpreted the data; drafted the manuscript; both authors critically reviewed the manuscript for important intellectual content; both authors gave approval of the version to be submitted; both authors agree to be accountable for all aspects of the work.

Competing interests All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

References

1 Felmingham C, MacNamara S, Cranwell W, Williams N, Wada M, Adler NR, et al. Improving Skin cancer Management with ARTificial Intelligence (SMARTI): protocol for a preintervention/postintervention trial of an artificial intelligence system used as a diagnostic aid for skin cancer management in a specialist dermatology setting. BMJ Open. 2022;12:e050203. Medline:34983756 doi:10.1136/bmjopen-2021-050203

2 Frye BL. Should using an AI text generator to produce academic writing be plagiarism? Fordham Intell Prop Media & Ent LJ. 2023;33:946.

3 Aydın Ö, Karaarslan E. OpenAI ChatGPT generated literature review: Digital twin in healthcare. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4308687. Accessed: April 10, 2024.

4 Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage. 2023;71:102642. doi:10.1016/j.ijinfomgt.2023.102642

5 AI. ChatGPT 2023 [cited 2023]. Available from: https://chat.openai.com/chat. Accessed: April 10, 2024.

6 Else H. Abstracts written by ChatGPT fool scientists. Nature. 2023;613:423. Medline:36635510 doi:10.1038/d41586-023-00056-7

7 Koteluk O, Wartecki A, Mazurek S, Kołodziejczak I, Mackiewicz A. How do machines learn? artificial intelligence as a new era in medicine. J Pers Med. 2021;11:32. Medline:33430240 doi:10.3390/jpm11010032

8 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence. 2023;6:1169595. Medline:37215063 doi:10.3389/frai.2023.1169595

9 Homolak J. Exploring the adoption of ChatGPT in academic publishing: insights and lessons for scientific writing. Croat Med J. 2023;64:205. Medline:37391919 doi:10.3325/cmj.2023.64.205

10 Gastel B, Day RA. How to write and publish a scientific paper: ABC-CLIO; 2022.

11 Kumar S, Arora H, Ghosal T, Ekbal A. DeepASPeer: towards an aspect-level sentiment controllable framework for decision prediction from academic peer reviews. Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries; Cologne, Germany: Association for Computing Machinery; 2022. p. Article 29.

12 Zhu Y, Han D, Chen S, Zeng F, Wang C. How can chatgpt benefit pharmacy: A case report on review writing. Preprints. 2023, 2023020324.

13 Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. J Multidiscip Healthc. 2023;•••:1513-20. Medline:37274428 doi:10.2147/JMDH.S413470

14 Fessenko D. Can artificial intelligence (re) define creativity? In EthicAI=LABS Project. Sofia: DA LAB Foundation /Goethe-institut Sofia. pp. 34-48.

15 King MR. ChatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cell Mol Bioeng. 2023;16:1-2. Medline:36660590 doi:10.1007/s12195-022-00754-8

16 Curtis N. ChatGPT§. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. Pediatr Infect Dis J. 2023;42:275. Medline:36757192 doi:10.1097/INF.0000000000003852

17 Currie GM, editor. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? Seminars in Nuclear Medicine; 2023: Elsevier.

18 Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J. 2023;64:1. Medline:36864812 doi:10.3325/cmj.2023.64.1

19 Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care. 2023;27:1-5.

20 Khalil M, Er E. Will ChatGPT get you caught? Rethinking of plagiarism detection. arXiv preprint arXiv:230204335. 2023.

21 Marchandot B, Matsushita K, Carmona A, Trimaille A, Morel O. ChatGPT: the next frontier in academic writing for cardiologists or a pandora's box of ethical dilemmas. European Heart Journal Open. 2023;3(2):oead007. Medline:36915398 doi:10.1093/ehjopen/oead007

22 Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to original

abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv. 2022:2022.12.23.521610. doi:10.1101/2022.12.23.521610

23 Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S. Can AI-generated text be reliably detected? arXiv preprint arXiv:230311156. 2023.

24 Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res. 2023;13:1148. Medline:37168339