DOI:10.14744/bej.2025.70973
Beyoglu Eye J 2025; 10(3): 175-180

Original Article

Comparative Detector Analysis for the Identification of Academic Articles Synthesized by Artificial Intelligence in the Field of Ophthalmology

Sebnem Kaya Ergen

Department of Ophthalmology, Kocaeli State Hospital, Kocaeli, Türkiye

Abstract

Objectives: The increasing use of large language models, such as ChatGPT, in academic writing has raised significant ethical concerns within the academic community. This study explores the potential challenges posed by the ability of artificial intelligence (AI) to produce realistic, evidence-based academic texts and investigates whether these challenges can be effectively controlled.

Methods: Three original articles in the field of ophthalmology were provided as input to ChatGPT-4o to generate introduction sections. A total of 50 introduction texts were synthesized from 150 original articles. These AI-generated texts were analyzed using AI detectors (GPTZero, Writer, CorrectorApp, and ZeroGPT) and a plagiarism detector. In addition, the ability of AI detectors to differentiate between original and AI-generated texts was evaluated.

Results: There was a statistically significant difference in AI detector probabilities between original and AI-generated texts ($p<0.001$ for all detectors). GPTZero demonstrated a sensitivity of 100% and a specificity of 96% in distinguishing original from AI-generated texts, outperforming all other AI detectors. However, paraphrased AI-generated texts significantly reduced the detection accuracy of GPTZero ($p<0.001$).

Conclusion: ChatGPT-4o demonstrated the ability to synthesize new texts with referenced citations within seconds, capable of bypassing plagiarism detectors. However, AI detectors showed limitations in achieving absolute accuracy and occasionally misclassified original texts. Even with the most accurate AI detectors, a simple paraphrasing method significantly compromised prediction accuracy, highlighting the need for improved detection strategies and ethical oversight.

Keywords: Artificial intelligence detection, Artificial intelligence ethic, ChatGPT-4o, large language models, research policy

Introduction

Large language models (LLMs) are complex neural network-based transformative models that create natural, conversational content that is often difficult to distinguish from human-written text (1). ChatGPT is built on the largest of such models, generative pre-trained transformer-3 (GPT-3), and millions of people started using this tool, which OpenAI (San Francisco, CA, USA) released for free in November 2022 (2,3).

In the last few years, articles that include artificial intelligence (AI) tools such as ChatGPT as authors in their research have been entering the literature (4). Concerns are growing in academia about the misuse of AI chatbots for scientific paper writing, and some reputable scientific journals have reported that they have banned the use of ChatGPT for scientific article writing (5,6). However, there is no universal policy yet.

How to cite this article: Kaya Ergen S. Comparative Detector Analysis for The Identification of Academic Articles Synthesized by Artificial Intelligence in The Field of Ophthalmology. Beyoglu Eye J 2025; 10(3): 175-180.

Address for correspondence: Sebnem Kaya Ergen, MD. Department of Ophthalmology, Kocaeli State Hospital, Kocaeli, Türkiye
Phone: +90 507 207 35 53 **E-mail:** dr.sebnem.kya@gmail.com

Submitted Date: January 08, 2025 **Revised Date:** July 11, 2025 **Accepted Date:** August 15, 2025 **Available Online Date:** September 25, 2025

Beyoglu Eye Training and Research Hospital - Available online at www.beyoglueye.com

Recently, various studies have been published investigating the potential of distinguishing abstracts prepared by AI (7-9). As a result of these studies, it was found that none of the human examinations and AI detector scans were perfect discriminators (7-10). While the functionality of existing AI detectors is just being discussed, OpenAI has released ChatGPT-4 (March 2023) and introduced the PDF upload feature in this version. An important point here is that when AI is prompted to generate a scientific article using references from the literature, it may produce incomplete or fabricated information due to its limited access to external sources and inability to verify content. The PDF upload feature will provide AI with the opportunity to increase the accuracy and credibility of the text it produces. As AI and AI detectors rapidly improve, continuous effort is needed to evaluate their performance.

In this study, it was requested that ChatGPT-4o (Version May 2024), the latest version released to the market, read and analyze three different original articles submitted to the literature with similar titles in the field of ophthalmology and synthesize an introduction with appropriate references. A total of 50 AI-generated texts were scanned in AI detectors, and their functionality was investigated. At the same time, the original texts were scanned by these detectors to investigate the possibility of misidentification.

Besides this study, which test the latest version of ChatGPT, no other study has been found in the literature in which AI chatbots have produced scientific synthesis text from more than one scientific article. The aim of this study is to draw attention to the potential threat of AI-generated texts, whose accuracy has not been examined, to the ophthalmology academy.

Methods

A total of 150 original articles, three each under similar titles, were sourced from PubMed and collected from February 2012 to November 2021 issues of six high-impact open-access journals (Eye and Vision, Investigative Ophthalmology and Visual Science, Retina, Translational Vision Science and Technology, Clinical Ophthalmology, and BMJ Open Ophthalmology). Articles with similar titles have been uploaded to ChatGPT-4o (OpenAI, San Francisco, CA, Version May 2024), the latest version on the market. The prompt given to the model was "Based on the introduction in the three articles provided, please synthesize an introduction for a new article. Cite references that contain actual articles using the Harvard reference style." 50 AI-generated texts were obtained by opening a new session each time. Ethical approval was not required because the study did not involve human participants and used only publicly available data. Sample prompts and AI-generated texts are available in the supplementary material.

The following four AI detectors were tested: GPTZero (https://gptzero.me/), ZeroGPT (https://www.zerogpt.com/), Writer (https://writer.com/ai-content-detector/), and CorrectorApp (https://corrector.app/ai-content-detector/). These programs, which can be used online for free, evaluate the probability of the given text being human or an AI production. To detect plagiarism, the introductions created by ChatGPT were scanned in "Plagiarism Checker" (https://plagiarismdetector.net/), which is a free web scanning plagiarism detection tool. All of these tools are software that offer a percentage probability from 0 to 100 when evaluating the probability of the text being human-written or AI-generated. Increasing value indicates a higher probability of AI production. Finally, 50 AI-generated texts were rewritten in the QuillBot program (https://quillbot.com/paraphrasing-tool), which offers paraphrasing tool features, and these texts were retested by GPTZero.

Statistical Analysis

All statistical analyses were performed using IBM Statistical Package for the Social Sciences, version 25.0 (IBM Corp., Armonk, NY, USA). The convenience of the data to normal distribution was evaluated using the Kolmogorov–Smirnov and Shapiro–Wilk normality tests. Variables were presented as mean±standard deviation and median (interquartile range [IQR]). A comparison of the probabilities given by the detector software for the original and ChatGPT-generated texts was made with a 2-sided Mann–Whitney U test. The Friedman test was used to compare the detectors with each other. Pearson's r effect size value was calculated for the Mann–Whitney U-test. Kendall's W effect size value was calculated for the Friedman test. For two-sided hypothesis testing, statistical significance was set at $p<0.05$.

Online-Only Supplemental Material Description

The article includes online-only supplementary material, and this material provides detailed examples of ChatGPT-generated text based on academic articles in ophthalmology. It includes synthesized introductions and reference citations, demonstrating the methodology and outcomes of AI-assisted content creation in this field.

Results

The likelihood of 100 texts (50 original and 50 ChatGPT-generated) being written by AI was evaluated, and a statistically significant difference was found in the probabilities detected by the AI detectors ($p<0.001$, all). The results were as follows: GPTZero (100% [IQR, 98.0–100%] vs. 2% [IQR, 1.00–3.00%]; $p<0.001$); Writer (15.5% [IQR, 12.0–18%] vs. 0% [IQR, 0.00–3.00%]; $p<0.001$); ZeroGPT (86.72% [IQR, 60.88–100%] vs. 24.28% [IQR, 0.00–66.03%]; $p<0.001$); CorrectorApp (85.91% [IQR, 58.77–100%] vs. 26.44% [IQR, 0.00–71.86%]; $p<0.001$) (Table 1 and Fig. 1).

Table 1. Comparative evaluation of the probabilities found by artificial intelligence text detectors for original texts and ChatGPT-generated texts (%)

AI-detector	Texts	n	Mean±SD	Median (Q1–Q3)	Z	p[a]	r[b]
ZeroGPT	ChatGPT-generated	50	80.11±21.26	86.72 (60.88–100.00)	–5.823	<0.001	0.58
	Original	50	36.50±35.47	24.28 (0.00–66.03)			
GPTZero	ChatGPT-generated	50	99.10±2.08	100.00 (98.00–100.00)	–8.908	<0.001	0.89
	Original	50	3.12±4.25	2.00 (1.00–3.00)			
Corr.App	ChatGPT-generated	50	76.94±24.18	85.91 (58.77–100.00)	–5.235	<0.001	0.52
	Original	50	38.41±35.57	26.44 (0.00–71.86)			
Writer	ChatGPT-generated	50	16.34±10.03	15.50 (12.00–18.00)	–8.397	<0.001	0.84
	Original	50	1.70±3.11	0.00 (0.00–3.00)			

Test statistics: Mann–Whitney U test; [a]Significant P-values (<0.05) are shown in bold; [b]Pearson r effect size value; SD: Standard deviation.

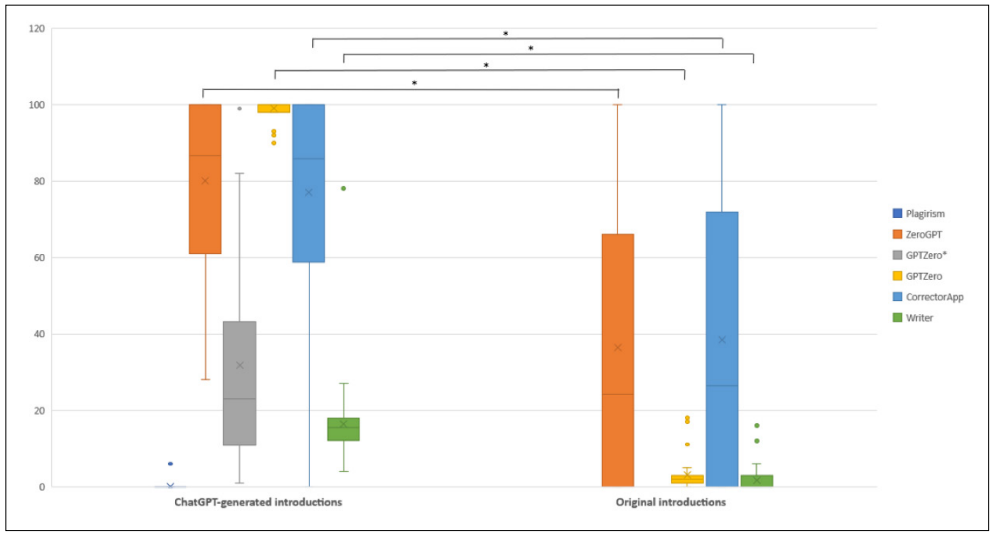

Figure 1. Percentage estimates of artificial intelligence (AI) text detectors giving ChatGPT-generated and original texts as AI generation.

When the effect size values were examined, it was seen that the AI detector that gave the most successful results was GPTZero (r=0.89). The average of the probabilities given by GPTZero for AI-generated texts is 99.10±2.08, whereas for original texts, this average is 3.12±4.25. When the AI-generated texts were paraphrased and evaluated again with GPTZero, it was observed that there was a statistically significant decrease in the probability in the second evaluation (100% [IQR, 98.0–100%] vs. 23% [IQR, 10.75–43.25%]; p<0.001) between the two evaluations of the same texts (Table 2).

Based on effect size, GPTZero was the most accurate detector (r=0.89). Its average probability for AI-generated texts was 99.10±2.08, whereas for original texts it was 3.12±4.25. After paraphrasing the AI-generated texts and re-evaluating them with GPTZero, a significant decrease was observed (100% [IQR, 98.0–100%] vs. 23% [IQR, 10.75–43.25%]; p<0.001) (Table 2).

When the results of AI detectors evaluating AI-generated texts were examined, it was seen that there was a statistically significant difference between them (p>0.05) (Table 2). Mean rank values for the Friedman test were found as follows: GPTZero: 5.63; ZeroGPT: 5.01; CorrectorApp: 4.23; GPT-Zero (post-paraphrase evaluation): 2.77; Writer: 2.35. While the Plagiarism Detector program gave a 0% plagiarized score in 49 of the 50 introductions produced with ChatGPT, it presented a 6% plagiarized percentage for only one text.

The comparison of the probabilities given by AI detectors for original introductions is presented in Table 3. According to these results, it was seen that there was a significant statistical difference between GPTZero and ZeroGPT (p=0.04), between Writer and ZeroGPT (p<0.001), and

Table 2. Evaluation of ChatGPT-generated texts by AI text detectors (%)

AI-detector	Mean±SD	Min-max.	Median (Q1–Q3)	p[q]
ZeroGPT	80.11±21.26	27.95–100.00	86.72 (60.88–100.00)	<0.05[b,t]
GPTZero	99.10±2.08	90.00–100.00	100.00 (98.00–100.00)	<0.05[a,x,y]
Corr.App	76.94±24.18	0.00–100.00	85.91 (58.77–100.00)	<0.05[c,x,z]
Writer	16.34±10.03	4.00–78.00	15.50 (12.00–18.00)	<0.05[a,b,c]
GPTZero*	31.84±25.91	1.00–99.00	23.00 (10.75–43.25)	<0.05[y,z,t]

Test statistics: Friedman test; [q]Significant P-values (<0.05) are shown in bold. GPTZero*: Post-paraphrase evaluation; [a]Between Writer and GPTZero; [b]Between Writer and ZeroGPT; [c]Between Writer and Corr.App; [x]Between Corr.App and GPTZero; [y]Between GPTZero* and GPTZero; [z]Between GPTZero* and Corr.App; [t]Between GPTZero* and ZeroGPT; AI: Artificial intelligence; SD: Standard deviation.

Table 3. Evaluation of ChatGPT-generated texts by AI text detectors (%)

AI-detector	Mean±SD	Min-Max	Median (Q1–Q3)	p[q]
ZeroGPT	36.50±35.47	0.00–100.00	24.28 (0.00–66.03)	<0.05[k,m]
GPTZero	3.12±4.25	0.00–18.00	2.00 (1.00–3.00)	<0.05[k]
Corr.App	38.41±35.57	0.00–100.00	26.44 (0.00–71.86)	<0.05[n]
Writer	1.70±3.11	0.00–16.00	0.00 (0.00–3.00)	<0.05[m,n]

Test statistics: Friedman test; [q]Significant P-values (<0.05) are shown in bold; [k]Between GPTZero and ZeroGPT; [m]Between Writer and ZeroGPT; [n]Between Writer and Corr.App; AI: Artificial intelligence.

between Writer and Corr.App (p<0.001). The mean rank order for the Friedman test was found as follows: ZeroGPT: 3.04; CorrectorApp: 2.93; GPTZero: 2.34; Writer: 1.69.

ROC curves for scores of AI detectors are presented in Figure 2 and Table 4. According to the results, ZeroGPT had an area under the receiver operating characteristic (AUROC) curve of 0.84 for detecting generated introductions. At the optimal cutoff (95.39%), maximizing sensitivity and specificity, ZeroGPT had a sensitivity of 38% and a specificity of 94% in differentiating original versus generated introductions. GPTZero had an AUROC curve of 1.000 for detecting AI-generated texts. At the optimal cutoff (17.5%), maximizing sensitivity and specificity, GPTZero had a sensitivity of 100% and a specificity of 96% in differentiating original versus generated introductions. CorrectorApp had an AUROC curve of 0.802 for detecting generated introductions. At the optimal cutoff maximizing sensitivity and specificity (94%), CorrectorApp had a sensitivity of 32% and a specificity of 94% in differentiating original versus generated introductions. The Writer had an AUROC curve of 0.981 for detecting generated introductions. At the optimal cutoff maximizing sensitivity and specificity (12.5%), Writer had a sensitivity of 68% and a specificity of 98% in differentiating original versus generated introductions.

Discussion

Writing an original academic article requires researchers to spend a long time on processes such as collecting information, analyzing it with critical thinking, and accurately referencing the data. The possibility of using LLM as an author in scientific research has brought up ethical and accuracy problems (11). Although AI facilitates language-related challenges and eases the writing process for researchers, it also raises concerns about the inclusion of unreliable studies or inadequately reviewed systematic reviews into the scientific literature.

In the study, three original scientific articles, different in each new chat, were uploaded to CHATGPT-4o, and it was asked to read and analyze these articles and write an introduction with their references. While 33 of these 50 introductions produced by ChatGPT showed only the uploaded articles in their references, it was seen that in 17 texts, some other references were added to the text. When checked, all of these references were taken from actual articles in the literature that the main articles referenced, with authors and publication titles matching. Even though the references were real articles from the correct year and journal, it was seen that in some places, the sentences that they claimed to be

Table 4. Receiver operating characteristics analysis to predict the probability of texts being AI or human-written

	AUC (% 95)	Cut off	p	Sensitivity (%)	Specificity (%)
ZeroGPT	0.836 (0.756–0.916)	95.39	<0.001	38	94
GPTZero	1.000 (1.000–1.000)	17.5	<0.001	100	96
CorrectorApp	0.802 (0.716–0.889)	93.995	<0.001	32	94
Writer	0.981 (0.955–1.000)	12.5	<0.001	68	98

AI: Artificial intelligence; AUC: Area under the curve.

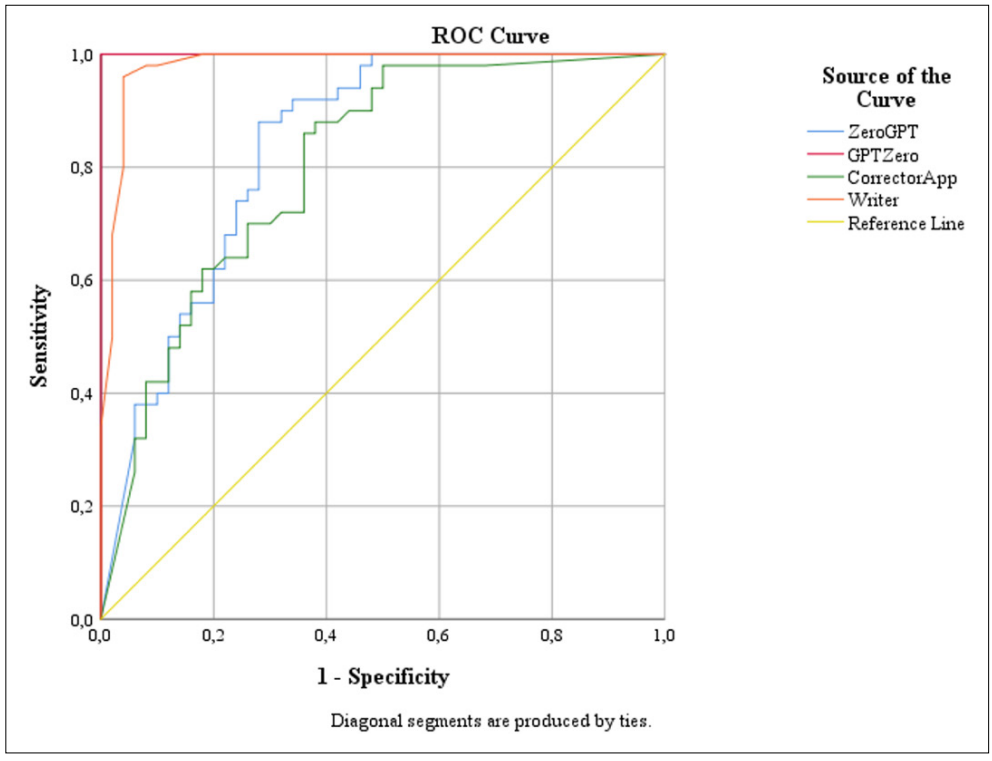

Figure 2. Receiver operating characteristic curve for scores of artificial intelligence text detectors.

relevant were not included in the referred articles, and the content of another article could be referred to a different article. Accepting seemingly evidence-based texts without careful examination could undermine trust in the scientific literature in the future. At this point, the effectiveness of detectors developed to detect texts produced by AI is worth examining. In the study, four AI detectors and a plagiarism detector program were used to detect plagiarism.

The plagiarism detection tool gave a 0% plagiarism score in 49 of 50 texts. Among the AI detectors whose functionality was examined, it can be said that GPTZero predicts AI-generated texts more accurately than all other detectors except ZeroGPT (Table 2). We see that the prediction probability of GPTZero, which works with an accuracy rate of nearly 100%, decreases significantly when the texts are simply paraphrased (QuillBot program) with an effort that takes seconds ($p<0.001$). This situation draws attention to the difficulty of detecting the texts produced when an AI robot is used for fraudulent research.

Another important point here was the possibility that AI detectors could mistakenly show human-written texts as AI production. It has been observed that the writer program, which has the lowest mean rank in detecting AI-generated texts, gives more accurate predictions than CorrectorApp and ZeroGPT in predicting the original texts, but it cannot be said that any tool works with near-perfect accuracy, raising the risk of unfair accusations against authors.

In this study, where the effectiveness of more than one AI detector was compared, the latest version of ChatGPT, 4o, was tested. While the AI-generated texts were initially

evaluated without any modifications, they were subsequently paraphrased to simulate potential human editing, and the impact on AI detector prediction probabilities was assessed. Furthermore, there are several limitations in the study. First, the content of the produced texts was not examined with a critical approach, so how original the content produced by AI was and how balanced and synthesized the subjects were not examined in this article. Therefore, no manual human control was provided to the texts. Another limitation of the study is the possibility that ChatGPT may produce a different response to the same prompt each time. Future studies could be expanded with a larger number of AI texts and plagiarism detectors, and to areas outside the field of ophthalmology.

Conclusion

This study highlights the growing challenges and ethical dilemmas associated with the use of AI, particularly LLMs such as ChatGPT-4o in academic writing. While AI demonstrates impressive capabilities in synthesizing and referencing texts, the inaccuracies in citation content and the ability of simple paraphrasing techniques to bypass even the most advanced AI detectors underscore the limitations of current AI detectors. Furthermore, the risk of AI misclassifying human-written texts and the potential for fraudulent research to infiltrate scientific literature emphasize the urgent need for improved detection systems and rigorous human oversight. As AI continues to advance, researchers and academic institutions must prioritize the development of robust ethical guidelines and reliable tools to ensure the integrity of academic research. Future studies should explore broader applications, test additional AI detectors, and critically evaluate the originality and synthesis quality of AI-generated content.

This article was presented as an oral presentation at the 58th National Congress of the Turkish Ophthalmological Association, Antalya, Türkiye, November 20–24, 2024.

Disclosures

Ethics Committee Approval: Ethical approval was not required because the study did not involve human participants and used only publicly available data.

Conflict of Interest: None declared.

Funding: The authors declare that this study has received no financial support.

Use of AI for Writing Assistance: Not declared.

Peer-review: Externally peer-reviewed.

References

1. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. J Med Syst 2023;47:33. [CrossRef]
2. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021;4:93. [CrossRef]
3. Vallance C. ChatGPT: New AI chatbot has everyone talking to it. BBC News. 2022 Dec 5. Available from: https://www.bbc.com/news/technology-63861322. Accessed Sep 2, 2025.
4. Stokel-Walker C. ChatGPT listed as author on research papers: Many scientists disapprove. Nature 2023;613:620–1. [CrossRef]
5. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023;613:612. [CrossRef]
6. Thorp HH. ChatGPT is fun, but not an author. Science 2023;379:313. [CrossRef]
7. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023;6:75. [CrossRef]
8. Pan ET, Florian-Rodriguez M. Human vs machine: Identifying ChatGPT-generated abstracts in gynecology and urogynecology. Am J Obstet Gynecol 2024;231:276.e1–e10. [CrossRef]
9. Rashidi HH, Fennell BD, Albahra S, Hu B, Gorbett T. The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. J Pathol Inform 2023;14:100342. [CrossRef]
10. Else H. Abstracts written by ChatGPT fool scientists. Nature. 2023 Jan 12. Available from: https://www.nature.com/articles/d41586-023-00056-7. Accessed Sep 2, 2025.
11. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: Ethical challenges for medical publishing. Lancet Digit Health 2023;5:e105–6. [CrossRef]