

ORIGINAL ARTICLE

Open Access



# The great detectives: humans versus AI detectors in catching large language model-generated medical writing

Jae Q. J. Liu<sup>1</sup>, Kelvin T. K. Hui<sup>1</sup>, Fadi Al Zoubi<sup>1</sup>, Zing Z. X. Zhou<sup>1</sup>, Dino Samartzis<sup>2</sup>, Curtis C. H. Yu<sup>1</sup>, Jeremy R. Chang<sup>1</sup> and Arnold Y. L. Wong<sup>1\*</sup>

\*Correspondence:  
arnold.wong@polyu.edu.hk

<sup>1</sup> Department of Rehabilitation Science, The Hong Kong Polytechnic University, Hong Kong, SAR, China

<sup>2</sup> Department of Orthopedic Surgery, Rush University Medical Center, Chicago, IL, USA

## Abstract

**Background:** The application of artificial intelligence (AI) in academic writing has raised concerns regarding accuracy, ethics, and scientific rigour. Some AI content detectors may not accurately identify AI-generated texts, especially those that have undergone paraphrasing. Therefore, there is a pressing need for efficacious approaches or guidelines to govern AI usage in specific disciplines.

**Objective:** Our study aims to compare the accuracy of mainstream AI content detectors and human reviewers in detecting AI-generated rehabilitation-related articles with or without paraphrasing.

**Study design:** This cross-sectional study purposively chose 50 rehabilitation-related articles from four peer-reviewed journals, and then fabricated another 50 articles using ChatGPT. Specifically, ChatGPT was used to generate the introduction, discussion, and conclusion sections based on the original titles, methods, and results. Wordtune was then used to rephrase the ChatGPT-generated articles. Six common AI content detectors (Originality.ai, Turnitin, ZeroGPT, GPTZero, Content at Scale, and GPT-2 Output Detector) were employed to identify AI content for the original, ChatGPT-generated and AI-rephrased articles. Four human reviewers (two student reviewers and two professorial reviewers) were recruited to differentiate between the original articles and AI-rephrased articles, which were expected to be more difficult to detect. They were instructed to give reasons for their judgements.

**Results:** Originality.ai correctly detected 100% of ChatGPT-generated and AI-rephrased texts. ZeroGPT accurately detected 96% of ChatGPT-generated and 88% of AI-rephrased articles. The areas under the receiver operating characteristic curve (AUROC) of ZeroGPT were 0.98 for identifying human-written and AI articles. Turnitin showed a 0% misclassification rate for human-written articles, although it only identified 30% of AI-rephrased articles. Professorial reviewers accurately discriminated at least 96% of AI-rephrased articles, but they misclassified 12% of human-written articles as AI-generated. On average, students only identified 76% of AI-rephrased articles. Reviewers identified AI-rephrased articles based on 'incoherent content' (34.36%), followed by 'grammatical errors' (20.26%), and 'insufficient evidence' (16.15%).



**Conclusions and relevance:** This study directly compared the accuracy of advanced AI detectors and human reviewers in detecting AI-generated medical writing after paraphrasing. Our findings demonstrate that specific detectors and experienced reviewers can accurately identify articles generated by Large Language Models, even after paraphrasing. The rationale employed by our reviewers in their assessments can inform future evaluation strategies for monitoring AI usage in medical education or publications. AI content detectors may be incorporated as an additional screening tool in the peer-review process of academic journals.

**Keywords:** Artificial intelligence, ChatGPT, Paraphrasing tools, Generative AI, Academic integrity, AI content detectors, Peer review, Perplexity scores, Scientific rigour, Accuracy

## Introduction

Chat Generative Pre-trained Transformer (ChatGPT; OpenAI, USA) is a popular and responsive chatbot that has surpassed other Large Language Models (LLMs) in terms of usage (ChatGPT Statistics 2023). Being trained with 175 billion parameters, ChatGPT has demonstrated its capabilities in the field of medicine and digital health (OpenAI 2023). It has been reported to be able to solve higher-order reasoning questions in pathology (Sinha 2023). Currently, ChatGPT has been used in generating discharge summaries (Patel & Lam 2023), aiding in diagnosis (Mehnen et al. 2023), and providing health information to patients with cancer (Hopkins et al. 2023). Currently, ChatGPT has become a valuable writing assistant, especially in medical writing (Imran & Almusharaf 2023).

However, scientists did not support granting ChatGPT authorship in academic publishing because it could not be held accountable for the ethics of the content (Stokel-Walker 2023). Its tendency to generate plausible but non-rigorous or misleading content has raised doubts about the reliability of its outputs (Sallam 2023; Manohar & Prasad 2023). This poses a risk of disseminating unsubstantiated information. Therefore, scholars have been exploring ways to detect AI-generated content to uphold academic integrity, although there are conflicting perspectives on the utilization of detectors in academic publishing. Previous research found that 14 existing AI detection tools exhibited an average accuracy of less than 80% (Weber-Wulff et al. 2023). However, the availability of paraphrasing tools further complicates the detection of LLM-generated texts. Some AI content detectors were ineffective in identifying paraphrased texts (Anderson et al. 2023; Weber-Wulff et al. 2023). Moreover, some detectors may misclassify human-written articles, which can undermine the credibility of academic publications (Liang et al. 2023; Sadasivan et al. 2023).

Nevertheless, there have been advancements in AI content detectors. Turnitin and Originality.ai have shown excellent accuracy in discriminating between AI-generated and human-written essays in various academic disciplines (e.g., social sciences, natural sciences, and humanities) (Walters 2023). However, their effectiveness in detecting paraphrased academic articles remains uncertain. Importantly, the accuracy of universal AI detectors has shown inconsistencies across studies in different domains (Gao et al. 2023; Anderson et al. 2023; Walters 2023). Therefore, continuous efforts are necessary to identify detectors that can achieve near-perfect accuracy, especially in the detection of medical texts, which is of particular concern to the academic community.

In addition to using AI detectors to help identify AI-generated articles, it is crucial to assess the ability of human reviewers to detect AI-generated formal academic articles. A study found that four peer reviewers only achieved an average accuracy of 68% in identifying ChatGPT-generated biomedical abstracts (Gao et al. 2023). However, this study had limitations because the reviewers only assessed abstracts instead of full-text articles, and their assessments were limited to a binary choice of 'yes' or 'no' without providing any justifications for their decisions. The reported moderate accuracy is inadequate for informing new editorial policy regarding AI usage. To establish effective regulations for supervising AI usage in journal publishing, it is necessary to continuously explore the accuracy of experienced human reviewers and to understand the patterns and stylistic features of AI-generated content. This can help researchers, educators, and editors develop discipline-specific guidelines to effectively supervise AI usage in academic publishing.

Against this background, the current study aimed to (1) compare the accuracy of several common AI content detectors and human reviewers with different levels of research training in detecting AI-generated academic articles with or without paraphrasing; and (2) understand the rationale by human reviewers for determining AI-generated content.

## Methods

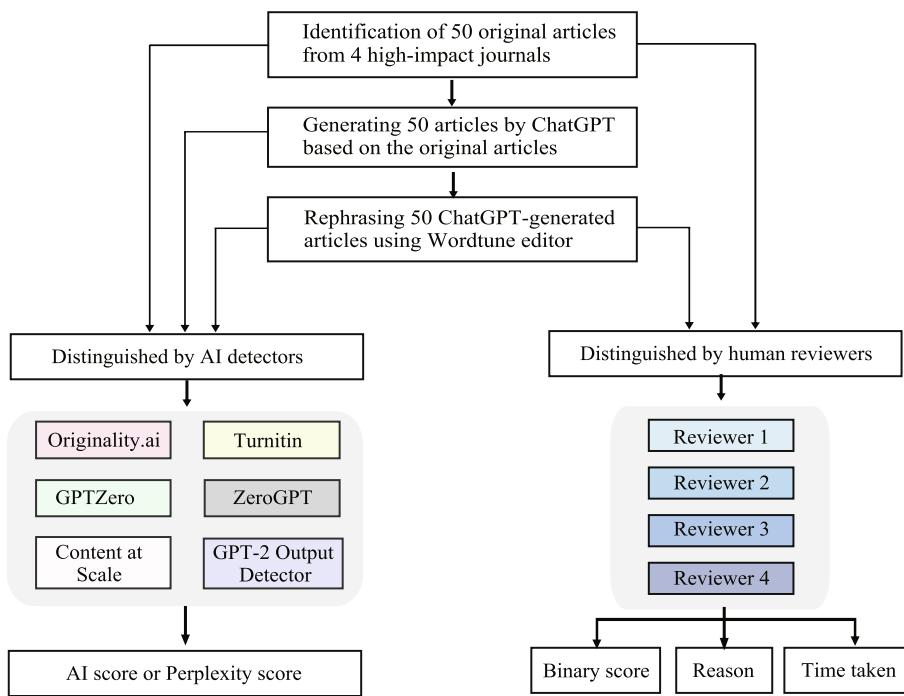
The current study was approved by the Institutional Review Board of a university. This study consisted of four stages: (1) identifying 50 published peer-reviewed papers from four high-impact journals; (2) generating artificial papers using ChatGPT; (3) rephrasing the ChatGPT-generated papers using a paraphrasing tool called Wordtune; and (4) employing six AI content detectors to distinguish between the original papers, ChatGPT-generated papers, and AI-rephrased papers. To determine human reviewers' ability to discern between the original papers and AI-rephrased papers, four reviewers reviewed and assessed these two types of papers (Fig. 1).

### Identifying peer-reviewed papers

As this study was conducted by researchers involved in rehabilitation sciences, only rehabilitation-related publications were considered. A researcher searched on PubMed in June 2023 using a search strategy involving: ("Neurological Rehabilitation"[Mesh]) OR ("Cardiac Rehabilitation"[Mesh]) OR ("Pulmonary Rehabilitation" [Mesh]) OR ("Exercise Therapy"[Mesh]) OR ("Physical Therapy"[Mesh]) OR ("Activities of Daily Living"[Mesh]) OR ("Self Care"[Mesh]) OR ("Self-Management"[Mesh]). English rehabilitation-related articles published between June 2013 and June 2023 in one of four high-impact journals (*Nature*, *The Lancet*, *JAMA*, and *British Medical Journal [BMJ]*) were eligible for inclusion. Fifty articles were included and categorized into four categories (musculoskeletal, cardiopulmonary, neurology, and pediatric) (Appendix 1).

### Generating academic articles using ChatGPT

ChatGPT (GPT-3.5-Turbo, OpenAI, USA) was used to generate the introduction, discussion, and conclusion sections of fabricated articles in July 2023. Specifically, before starting a conversation with ChatGPT, we gave the instruction "*Considering yourself as an academic writer*" to put it into a specific role. After that, we entered "*Please write a*



**Fig. 1** An outline of the study

*convincing scientific introduction on the topic of [original topic] with some citations in the text* into GPT-3.5-Turbo to generate the ‘Introduction’ section. The ‘Discussion’ section was generated by the request “*Please critically discuss the methods and results below: [original method] and [original result], Please include citations in the text*”. For the ‘Conclusions’ section, we instructed ChatGPT to create a summary of the generated discussion section with reference to the original title. Collectively, each ChatGPT-generated article comprised fabricated introduction, discussion, and conclusions sections, alongside the original methods and results sections.

#### Rephrasing ChatGPT-generated articles using a paraphrasing tool

Wordtune (AI21 Labs, Tel Aviv, Israel) (Wordtune 2023), a widely used AI-powered writing assistant, was applied to paraphrase 50 ChatGPT-generated articles, specifically targeting the introduction, discussion, and conclusion sections, to enhance their authenticity.

#### Identification of AI-generated articles

##### Using AI content detectors

Six AI content detectors, which have been widely used (Walters 2023; Crothers 2023; Top 10 AI Detector Tools 2023), were applied to identify texts generated by AI language models in August 2023. They classified a given paper as “human-written” or “AI-generated”, with a confidence level reported as an AI score [% ‘confidence in predicting that the content was produced by an AI tool’] or a perplexity score [randomness or particularity of the text]. A lower perplexity score indicates that the text has relatively few random elements and is more likely to be written by generative AI (GPTZero 2023). The 50

original articles, 50 ChatGPT-generated articles, and 50 AI-rephrased articles were evaluated for authenticity by two paid (Originality.ai, Originality. AI Inc., Ontario, Canada; and Turnitin's AI writing detection, Turnitin LLC, CA, USA) and four free AI content detectors (ZeroGPT, Munchberg, Germany; GPTZero, NJ, USA; Content at Scale, AZ, USA; and GPT-2 Output Detector, CA, USA). The authentic methods and results sections were not entered into the AI content detectors. Since the GPT-2 Output Detector has a restriction of 510 tokens per attempt, each article was divided into several parts for input, and the overall AI score of the article was calculated based on the mean score of all parts.

#### Peer reviews by human reviewers

Four blinded reviewers with backgrounds in physiotherapy and varying levels of research training (two college student reviewers and two professorial reviewers) were recruited to review and discern articles. To minimize the risk of recall bias, a researcher randomly assigned the 50 original articles and 50 AI-rephrased articles (ChatGPT-generated articles after rephrasing) to two electronic folders by a coin toss. If an original article was placed in Folder 1, the corresponding AI-rephrased article was assigned to Folder 2. Reviewers were instructed to review all the papers in Folder 1 first and then wait for at least 7 days before reviewing papers in Folder 2. This approach would reduce the reviewers' risk of remembering the details of the original papers and AI-rephrased articles on the same topic (Fisher & Radvansky 2018).

The four reviewers were instructed to use an online Google form (Appendix 2) to make their decision and provide reasons behind their decision. Reviewers were instructed to enter the article number on the Google form before reviewing the article. Once the reviewers had gathered sufficient information/confidence to make the decision, they would give a binary response ("AI-rephrased" or "human-written"). Additionally, they should select their top three reasons for their decision from a list of options (i.e., coherence creativity, evidence-based, grammatical errors, and vocabulary diversity) (Walters 2019; Lee 2022). The definitions of these reasons (Appendix 3) were explained to the reviewers beforehand. If they could not find the best answers, they could enter additional responses. When the reviewer submitted the form, the total duration was automatically recorded by the system.

#### Statistical analysis

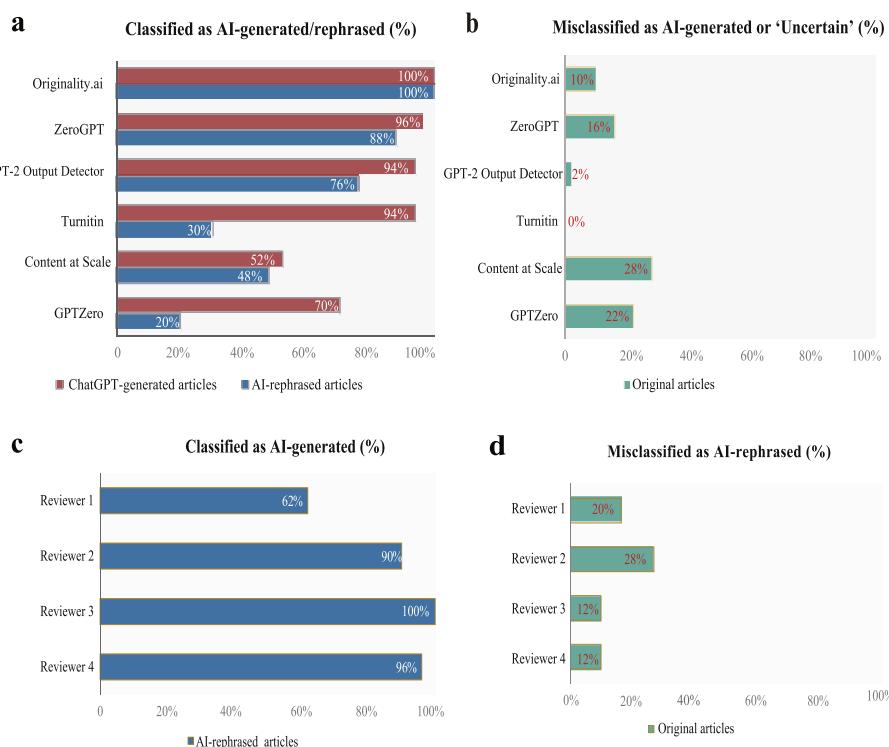
Descriptive analyses were reported when appropriate. Shapiro-Wilk tests were used to evaluate the normality of the data, while Levene's tests were employed to assess the homogeneity of variance. Logarithmic transformation was applied to the data related to 'time taken' to achieve the normal distribution. Separate two-way repeated measures analysis of variance (ANOVA) with post-hoc comparisons were conducted to evaluate the effect of detectors and AI usage on AI scores, and the effect of reviewers and AI usage on the time taken. Separate paired t-tests with Bonferroni correction were applied for pairwise comparisons. The GPTZero Perplexity scores were compared among groups of articles using one-way repeated ANOVA. Subsequently, separate paired t-tests with Bonferroni correction were used for pairwise comparisons. Receiver operating characteristic (ROC) curves were generated to determine cutoff values for the

highest sensitivity and specificity in detecting AI articles by AI content detectors. The area under the ROC curve (AUROC) was also calculated. Inter-rater agreement was calculated using Fleiss's kappa, and Cohen's kappa with Bonferroni correction was used for multiple comparisons. The significance level was set at  $p < 0.05$ . All statistical analyses were performed using SPSS (version 26; SPSS Inc., Chicago, IL, USA).

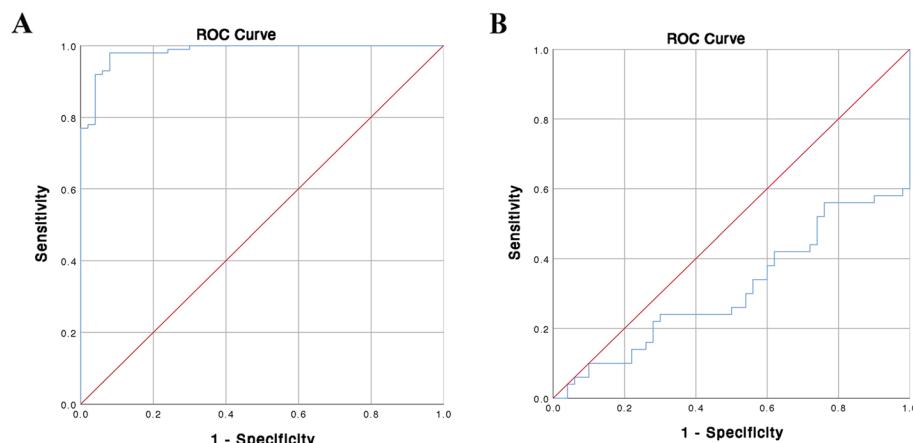
## Results

### The accuracy of AI detectors in identifying AI articles

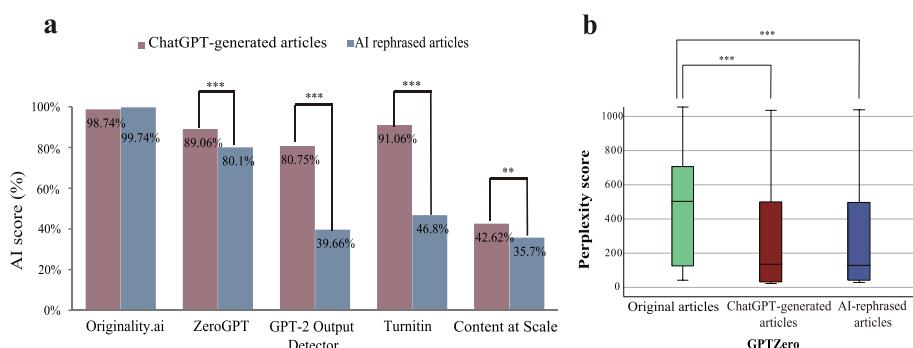
The accuracy of AI content detectors in identifying AI-generated articles is shown in Fig. 2a and b. Notably, Originality.ai demonstrated perfect accuracy (100%) in detecting both ChatGPT-generated and AI-rephrased articles. ZeroGPT showed near-perfect accuracy (96%) in identifying ChatGPT-generated articles. The optimal ZeroGPT cut-off value for distinguishing between original and AI articles (ChatGPT-generated and AI-rephrased) was 42.45% (Fig. 3a), with a sensitivity of 98% and a specificity of 92%. The GPT-2 Output Detector achieved an accuracy of 96% in identifying ChatGPT-generated articles based on an AI score cutoff value of 1.46%, as suggested by previous research (Gao et al. 2023). Likewise, Turnitin showed near-perfect accuracy (94%) in discerning ChatGPT-generated articles but only correctly



**Fig. 2** The accuracy of artificial intelligence (AI) content detectors and human reviewers in identifying large language model (LLM)-generated texts. **a** The accuracy of six AI content detectors in identifying AI-generated articles; **b** the percentage of misclassification of human-written articles as AI-generated ones by detectors; **c** the accuracy of four human reviewers (reviewers 1 and 2 were college students, while reviewers 3 and 4 were professorial reviewers) in identifying AI-rephrased articles; and **d** the percentage of misclassifying human-written articles as AI-rephrased ones by reviewers



**Fig. 3** The receiver operating characteristic (ROC) curve and the area under the ROC (AUROC) of artificial intelligence (AI) content detectors. **a** The ROC curve and AUROC of ZeroGPT for discriminating between original and AI articles, with the AUROC of 0.98; **b** the ROC curve and AUROC of GPTZero for discriminating between original and AI articles, with the AUROC of 0.312



**Fig. 4** Artificial intelligence (AI)-generated articles demonstrated reduced AI scores after rephrasing. **a** The mean AI scores of 50 ChatGPT-generated articles before and after rephrasing; **b** ChatGPT-generated articles demonstrated lower Perplexity scores computed by GPTZero as compared to original articles although increased after rephrasing; \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\* $p < 0.001$

discerned 30% of AI-rephrased articles. GPTZero and Content at Scale only correctly identified 70 and 52% of ChatGPT-generated papers, respectively. While Turnitin did not misclassify any original articles, Content at Scale and GPTZero incorrectly classified 28 and 22% of the original articles, respectively. AI scores, or perplexity scores, in response to the original, ChatGPT-generated, and AI-rephrased articles from each AI content detector are shown in Appendix 4. The classification of responses from each AI content detector is shown in Appendix 5.

All AI content detectors, except Originality.ai, gave rephrased articles lower scores as compared to the corresponding ChatGPT-generated articles (Fig. 4a). Likewise, GPTZero demonstrated that the perplexity scores of ChatGPT-generated ( $p < 0.001$ ) and AI-rephrased ( $p < 0.001$ ) texts were significantly lower than those of the original articles (Fig. 4b). The ROC curve of GPTZero perplexity scores for identifying original articles and AI articles showed that the respective AUROC were 0.312 (Fig. 3b).

### The accuracy of reviewers in identifying AI-rephrased articles

The median time spent by the four reviewers to distinguish original and AI-rephrased articles was 5 minutes (min) 45 seconds (s) (interquartile range [IQR] 3 min 42 s, 9 min 7 s). The median time taken by each reviewer to distinguish original and AI-rephrased articles is shown in Appendix 6. The two professorial reviewers demonstrated extremely high accuracy (96 and 100%) in discerning AI-rephrased articles, although both misclassified 12% of human-written articles as AI-rephrased (Fig. 2c and d, and Table 1). Although three original articles were misclassified as AI-rephrased by both professorial reviewers, they were correctly identified by Originality and ZeroGPT. The common reasons for an article to be classified as AI-rephrased by reviewers included ‘incoherence’ (34.36%), ‘grammatical errors’ (20.26%), ‘insufficient evidence-based claims’ (16.15%), vocabulary diversity (11.79%), creativity (6.15%), ‘misuse of abbreviations’ (5.87%), ‘writing style’ (2.71%), ‘vague expression’ (1.81%), and ‘conflicting data’ (0.9%). Nevertheless, 12 of the 50 original articles were wrongly considered AI-rephrased by two or more reviewers. Most of these misclassified articles were deemed to be incoherent and/or lack vocabulary diversity. The frequency of the primary reason given by each reviewer and the frequency of the reasons given by four reviewers for identifying AI-rephrased articles are shown in Fig. 5a and b, respectively.

Regarding the inter-rater agreement between two professorial reviewers, there was near-perfect agreement in the binary responses, with  $\kappa=0.819$  (95% confidence interval [CI] 0.705, 0.933,  $p<0.05$ ), as well as fair agreements in the primary and second reasons, with  $\kappa=0.211$  (95% CI 0.011, 0.411,  $p<0.05$ ) and  $\kappa=0.216$  (95% CI 0.024, 0.408,  $p<0.05$ ), respectively.

### “Plagiarized” scores of ChatGPT-generated or AI-rephrased articles

Turnitin results showed that the content of ChatGPT-generated and AI-rephrased articles had significantly lower ‘plagiarized’ scores ( $39.22\%\pm10.6$  and  $23.16\%\pm8.54\%$ , respectively) than the original articles ( $99.06\%\pm1.27\%$ ).

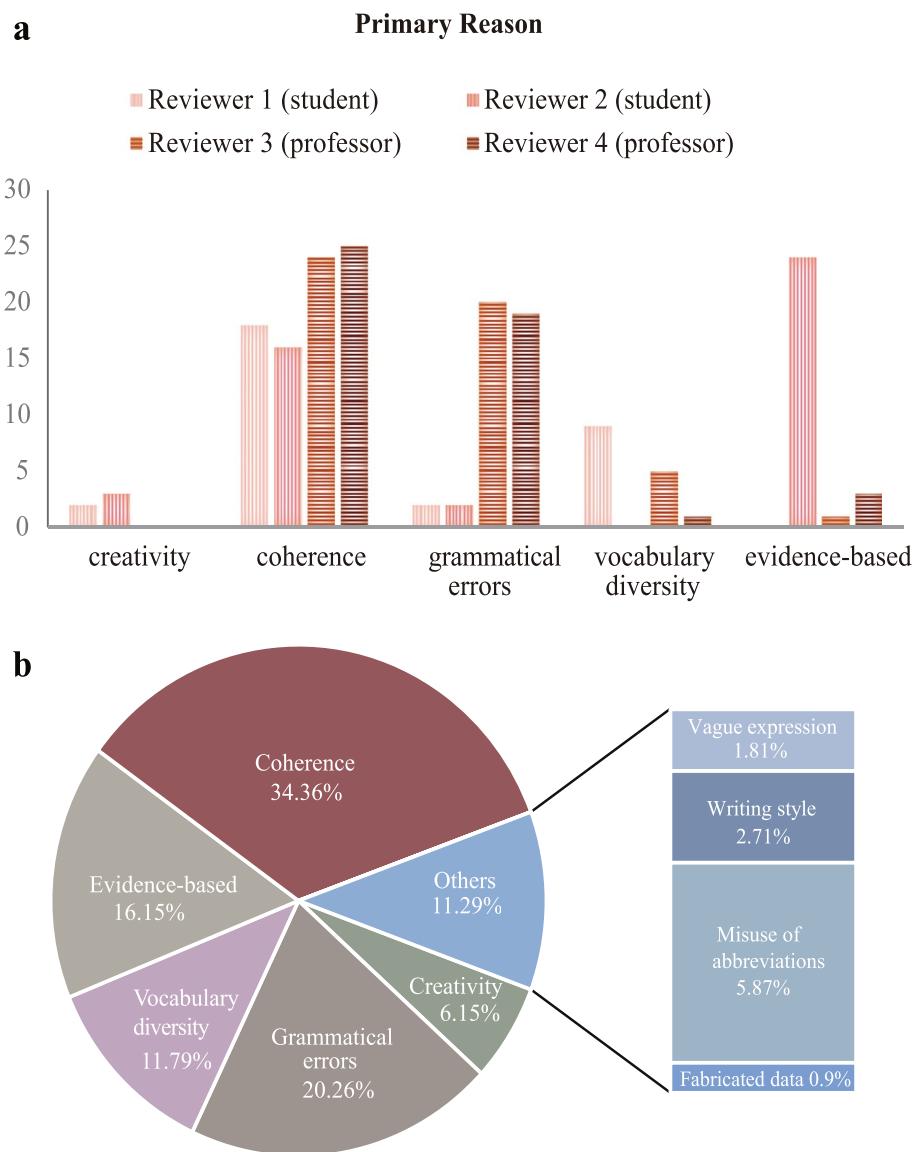
### Likelihood of ChatGPT being used in original articles after the launch of GPT-3.5-Turbo

No significant differences were found in the AI scores or perplexity scores calculated by the six AI content detectors ( $p>0.05$ ), or in the binary responses evaluated by reviewers ( $p>0.05$ ), when comparing the included original papers published before and after November 2022 (the release of ChatGPT).

**Table 1** Peer reviewers’ decisions on whether articles were original (i.e., human-written) or fabricated (i.e., artificial intelligence-generated articles after paraphrasing)

|                                   | Truth      |    |            |  | Truth                             |            |            |    |
|-----------------------------------|------------|----|------------|--|-----------------------------------|------------|------------|----|
|                                   | Original   |    | Fabricated |  | Original                          |            | Fabricated |    |
| Reviewer estimate<br>(reviewer 1) | Original   | 40 | 19         |  | Reviewer estimate<br>(reviewer 2) | Original   | 36         | 5  |
|                                   | Fabricated | 10 | 31         |  |                                   | Fabricated | 14         | 45 |
| Reviewer estimate<br>(reviewer 3) | Original   | 44 | 0          |  | Reviewer estimate<br>(reviewer 4) | Original   | 44         | 2  |
|                                   | Fabricated | 6  | 50         |  |                                   | Fabricated | 6          | 48 |

Reviewers 1 and 2 were college students, reviewers 3 and 4 were professorial reviewers



**Fig. 5** **A** The frequency of the primary reason for artificial intelligence (AI)-rephrased articles being identified by each reviewer. **B** The relative frequency of each reason for AI-rephrased articles being identified (based on the top three reasons given by the four reviewers)

## Discussion

Our study found that Originality.ai and ZeroGPT accurately detected AI-generated texts, regardless of whether they were rephrased or not. Additionally, Turnitin did not misclassify human-written articles. While professorial reviewers were generally able to discern AI-rephrased articles from human-written ones, they might misinterpret some human-written articles as AI-generated due to incoherent content and varied vocabulary. Conversely, AI-rephrased articles are more likely to go unnoticed by student reviewers.

### **What is the performance of generative AI in academic writing?**

Lee et al found that sentences written by GPT-3 tended to generate fewer grammatical or spelling errors than human writers (Lee 2022). However, ChatGPT may not necessarily minimize grammatical mistakes. In our study, reviewers identified ‘grammatical errors’ as the second most common reason for classifying an article as AI-rephrased. Our reviewers also noted that generative AI was more likely to inappropriately use medical terminologies or abbreviations, and even generate fabricated data. These might lead to a detrimental impact on academic dissemination. Collectively, generative AI is less likely to successfully create credible academic articles without the development of discipline-specific LLMs.

### **Can generative AI generate creative and in-depth thoughts?**

Prior research reported that ChatGPT correctly answered 42.0 to 67.6% of questions in medical licensing examinations conducted in China, Taiwan, and the USA (Zong 2023; Wang 2023; Gilson 2023). However, our reviewers discovered that AI-generated articles offered only superficial discussion without substantial supporting evidence. Further, redundancy was observed in the content of AI-generated articles. Unless future advancements in generative AI can improve the interpretation of evidence-based content and incorporate in-depth and insightful discussion, its utility may be limited to serving as an information source for academic works.

### **Who can be deceived by ChatGPT? How can we address it?**

ChatGPT is capable of creating realistic and intelligent-sounding text, including convincing data and references (Ariyaratne et al. 2023). Yeadon et al, found that ChatGPT-generated physics essays were graded as first-class essays in a writing assessment at Durham University (Yeadon et al. 2023). We found that AI-generated content had a relatively low plagiarism rate. These factors may encourage the potential misuse of AI technology for generating written assignments and the dissemination of misinformation among students. In a current survey, Welding (2023) reported that 50% of 1000 college students admitted to using AI tools to help complete assignments or exams. However, in our study, college student reviewers only correctly identified an average of 76% of AI-rephrased articles. Notably, our professorial reviewers found fabricated data in two AI-generated articles, while the student reviewers were unaware of this issue, which highlights the possibility of AI-generated content deceiving junior researchers and impacting their learning. In short, the inherent limitations of ChatGPT as reported by experienced reviewers may help research students understand some key points in critically appraising academic articles and be more competent in detecting AI-generated articles.

### **Which detectors are recommended for use?**

Our study revealed that Originality.ai emerged as the most sensitive and accurate platform for detecting AI-generated (including paraphrased) content, although it requires a subscription fee. ZeroGPT is an excellent free tool that exhibits a high level of sensitivity and specificity for detecting AI articles when the AI score threshold is set at 42.45%. These findings could help monitor the AI use in academic publishing and education, to

promisingly tackle ethical challenges posed by the iteration of AI technologies. Additionally, Turnitin, a widely used platform in educational institutions or scientific journals, displayed perfect accuracy in detecting human-written articles and near-perfect accuracy in detecting ChatGPT-generated content but was proved susceptible to deception when confronted with AI-rephrased articles. This raises concerns among educators regarding the potential for students to evade Turnitin AI detection by using an AI rephrasing editor. As generative AI technologies continue to evolve, educators and researchers should regularly conduct similar studies to identify the most suitable AI content detectors.

AI content detectors employ different predictive algorithms. Some publicly available detectors use perplexity scores and related concepts for identifying AI-generated writing. However, we found that the AUROC curve of GPTZero perplexity scores in identifying AI articles performed worse than chance. As such, the effectiveness of utilizing perplexity-based methods as the machine learning algorithm for developing an AI content detector remains debatable.

As with any novel technology, some merits and demerits require continuous improvement and development. Currently, AI content detectors have been developed as general-purpose tools to analyze text features, primarily based on the randomness of word choice and sentence lengths (Prillaman 2023). While technical issues such as algorithms, model tuning, and development are beyond the scope of this study, we have provided empirical evidence that offers potential directions for future advancements in AI content detectors. One such area that requires further exploration and investigation is the development of AI content detectors trained using discipline-specific LLMs.

#### **Should authors be concerned about their manuscripts being misinterpreted?**

While AI-rephrasing tools may help non-native English writers and less experienced researchers prepare better academic articles, AI technologies may pose challenges to academic publishing and education. Previous research has suggested that AI content detectors may penalize non-native English writers with limited linguistic expressions due to simplified wording (Liang et al. 2023). However, scientific writing emphasizes precision and accurate expression of scientific evidence, often favouring succinctness over vocabulary diversity or complex sentence structures (Scholar Hangout 2023). This raises concerns about the potential misclassification of human-written academic papers as AI-generated, which could have negative implications for authors' academic reputations. However, our results indicate that experienced reviewers are unlikely to misclassify human-written manuscripts as AI-generated if the articles present logical arguments, provide sufficient evidence-based support, and offer in-depth discussions. Therefore, authors should consider these factors when preparing their manuscripts to minimize the risk of misinterpretation.

Our study revealed that both AI content detectors and human reviewers occasionally misclassified certain original articles as AI-generated. However, it is noteworthy that no human-written articles were misclassified by both AI-content detectors and the two professorial reviewers simultaneously. Therefore, to minimize the risk of misclassifying human-written articles as AI-generated, editors of peer-reviewed journals may consider implementing a screening process that involves a reliable, albeit imperfect, AI-content

detector in conjunction with the traditional peer-review process, which includes at least two reviewers. If both the AI content detectors and the peer reviewers consistently label a manuscript as AI-generated, the authors should be given the opportunity to appeal the decision. The editor-in-chief and one member of the editorial board can then evaluate the appeal and make a final decision.

### Limitations

This study had several limitations. Firstly, the ChatGPT-3.5 version was used to fabricate articles given its popularity. Future studies should investigate the performance of upgraded LLMs. Secondly, although our analyses revealed no significant differences in the proportion of original papers classified as AI-written before and after November 2022 (the release of ChatGPT), we cannot guarantee that all original papers were not assisted by generative AI in their writing process. Future studies should consider including papers published before this date to validate our findings. Thirdly, although an excellent inter-rater agreement in the binary score was found between the two professorial reviewers, our results need to be interpreted with caution given the small number of reviewers and the lack of consistency between the two student reviewers. Future studies should address these limitations and expand our methodology to include other disciplines/industries with more reviewers to enhance the generalizability of our findings and facilitate the development of strategies for detecting AI-generated content in various fields.

### Conclusions

This is the first study to directly compare the accuracy of advanced AI detectors and human reviewers in detecting AI-generated medical writing after paraphrasing. Our findings substantiate that the established peer-reviewed system can effectively mitigate the risk of publishing AI-generated academic articles. However, certain AI content detectors (i.e., Originality.ai and ZeroGPT) can be used to help editors or reviewers with the initial screening of AI-generated articles, upholding academic integrity in scientific publishing. It is noteworthy that the current version of ChatGPT is inadequate to generate rigorous scientific articles and carries the risk of fabricating data and misusing medical abbreviations. Continuous development of machine-learning strategies to improve AI detection accuracy in the health sciences field is essential. This study provides empirical evidence and valuable insights for future research on the validation and development of effective detection tools. It highlights the importance of implementing proper supervision and regulation of AI usage in medical writing and publishing. This ensures that relevant stakeholders can responsibly harness AI technologies while maintaining scientific rigour.

### Abbreviations

|         |  |
|---------|--|
| AI      | Artificial intelligence                          |
| LLM     | Large language model                             |
| ChatGPT | Chat Generative Pre-trained Transformer          |
| ROC     | Receiver Operating Characteristic                |
| AUROC   | Area under the Receiver Operating Characteristic |

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s40979-024-00155-6>.

### Supplementary Material 1.

#### Acknowledgements

Not applicable.

#### Authors' contributions

Jae QJ Liu, Kelvin TK Hui and Arnold YL Wong conceptualized the study; Fadi Al Zoubi, Zing ZX Zhou, Curtis CH Yu, and Arnold YL Wong acquired the data; Jae QJ Liu and Kelvin TK Hui curated the data; Jae QJ Liu and Jeremy R Chang analyzed the data; Arnold YL Wong was responsible for funding acquisition and project supervision; Jae QJ Liu drafted the original manuscript; Arnold YL Wong and Dino Samartzis edited the manuscript.

#### Funding

The current study was supported by the GP Batteries Industrial Safety Trust Fund (R-ZDDR).

#### Availability of data and materials

The data and materials used in the manuscript are available upon reasonable request to the corresponding author.

## Declarations

#### Competing interests

All authors declare no conflicts of interest.

Received: 27 December 2023 Accepted: 13 March 2024

Published online: 20 May 2024

#### References

- Anderson N, Belavy DL, Perle SM, Hendricks S, Hespanhol L, Verhagen E, Memon AR (2023) AI did not write this manuscript, or did it? Can we trick the AI text detector into generating texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation. *BMJ Open Sport Exerc Med* 9(1):e001568
- Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R (2023) A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiol* 52:1755–1758
- ChatGPT Statistics, 2023, Detailed Insights On Users. <https://www.demandsage.com/chatgpt-statistics/> Accessed 08 Nov 2023
- Crothers E, Japkowicz N, Viktor HL (2023) Machine-generated text: a comprehensive survey of threat models and detection methods. *IEEE Access*
- Fisher JS, Radvansky GA (2018) Patterns of forgetting. *J Mem Lang* 102:130–141
- Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, Pearson AT (2023) Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med* 6:75
- Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D (2023) How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 9:e45312
- GPTZero, 2023, How do I interpret burstiness or perplexity? <https://support.gptzero.me/hc/en-us/articles/15130070230551-How-do-I-interpret-burstiness-or-perplexity>. Accessed August 20 2023
- Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ (2023) Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr* 7:pkad010
- Imran M, Almusharrif N (2023) Analyzing the role of ChatGPT as a writing assistant at higher education level: a systematic review of the literature. *Contemp Educ Technol* 15:ep464
- Lee M, Liang P, Yang Q (2022) Coauthor: designing a human-ai collaborative writing dataset for exploring language model capabilities. In: CHI Conference on Human Factors in Computing Systems, 1–19 ACM, April 2022
- Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J (2023) GPT detectors are biased against non-native English writers. *Patterns (N Y)* 4(7):100779
- Manohar N, Prasad SS (2023) Use of ChatGPT in academic publishing: a rare case of seronegative systemic lupus erythematosus in a patient with HIV infection. *Cureus* 15(2):e34616
- Mehnen L, Gruarin S, Vasileva M, Knapp B (2023) ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases medRxiv. <https://doi.org/10.1101/2023.04.20.23288859>
- OpenAI, 2023, Introducing ChatGPT:<https://openai.com/blog/chatgpt>. Accessed 30 Dec 2023
- Patel SB, Lam K (2023) ChatGPT: the future of discharge summaries? *Lancet Digital Health* 5:e107–e108
- Prillaman M (2023) ChatGPT detector catches AI-generated papers with unprecedented accuracy. *Nature*. <https://doi.org/10.1038/d41586-023-03479-4> Accessed 31 Dec 2023
- Sadasivan V, Kumar A, Balasubramanian S, Wang W, Feizi S (2023) Can AI-generated text be reliably detected? arXiv e-prints: 2303.11156
- Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare MDPI* 8:87:1

- Scholar Hangout, 2023, <https://www.manuscriptedit.com/scholar-hangout/maintaining-accuracy-in-academic-writing/>. Accessed September 10 2023
- Sinha RK, Deb Roy A, Kumar N, Mondal H (2023) Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus* 15(2):e35237
- Stokel-Walker C (2023) ChatGPT listed as author on research papers: many scientists disapprove. *Nature* 613(7945):620–621
- Top 10 AI Detector Tools, 2023, You Should Use. <https://www.eweek.com/artificial-intelligence/ai-detector-software/#chart>. Accessed August 2023
- Walters WH (2023) The effectiveness of software designed to detect AI-generated writing: a comparison of 16 AI text detectors. *Open Information Science* 7:20220158
- Wang Y-M, Shen H-W, Chen T-J (2023) Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc* 10:1097
- Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, Šigut P, Waddington L (2023) Testing of detection tools for AI-generated text. *Int J Educ Integrity* 19(1):26
- Welding L (2023) Half of college students say using AI on schoolwork is cheating or plagiarism. Best Colleges Wordtune. 2023, <https://app.wordtune.com/>. Accessed 16 July 2023
- Yeadon W, Iryang O-O, Mizouri A, Peach A, Testrow CP (2023) The death of the short-form physics essay in the coming AI revolution. *Phys Educ* 58:035027
- Zong H, Li J, Wu E, Wu R, Lu J, Shen B (2023) Performance of ChatGPT on Chinese National Medical Licensing Examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. *medRxiv*:2023.2007.2009.23292415

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.