DANISH MEDICAL JOURNAL

Original Article

Dan Med J 2023;70(12):A06230412

# A comparison of cover letters written by ChatGPT-4 or humans

Can Deniz Deveci, Jason Joe Baker, Binyamin Sikander & Jacob Rosenberg

Center for Perioperative Optimization, Department of Surgery, Copenhagen University Hospital – Herlev Hospital, Denmark

Dan Med J 2023;70(12):A06230412

## ABSTRACT

**INTRODUCTION.** Artificial intelligence has started to become a part of scientific studies and may help researchers with a wide range of tasks. However, no scientific studies have been published on its ussefulness in writing cover letters for scientific articles. This study aimed to determine whether Generative Pre-Trained Transformer (GPT)-4 is as good as humans in writing cover letters for scientific papers.

**METHODS.** In this randomised non-inferiority study, we included two parallel arms consisting of cover letters written by humans and by GPT-4. Each arm had 18 cover letters, which were assessed by three different blinded assessors. The assessors completed a questionnaire in which they had to assess the cover letters with respect to impression, readability, criteria satisfaction, and degree of detail. Subsequently, we performed readability tests with Lix score and Flesch Kincaid grade level.

**RESULTS.** No significant or relevant difference was found on any parameter. A total of 61% of the blinded assessors guessed correctly as to whether the cover letter was written by GPT-4 or a human. GPT-4 had a higher score according to our objective readability tests. Nevertheless, it performed better than human writing on readability in the subjective assessments.

**CONCLUSION.** We found that GPT-4 was non-inferior at writing cover letters compared to humans. This may be used to streamline cover letters for researchers, providing an equal chance to all researchers for advancement to peer-review.

**FUNDING.** This study received no financial support from external sources.

**TRIAL REGISTRATION.** This study was not registered before the study commenced.

Artificial intelligence (AI) has gained popularity in scientific research, aiding researchers with a wide range of tasks [1]. An example is the Chat Generative Pre-Trained Transformer (ChatGPT) series by OpenAI, which included GPT-3.5 in November 2022 and the subscription-based GPT-4 in March 2023 [2]. Researchers have already used ChatGPT to write essays, summarise literature, draft and improve papers, and write computer code [3, 4]. Nevertheless, no studies have been published on the capabilities of GPT-4 in writing cover letters for scientific journal submissions. In the submission process of an article, a cover letter promotes the manuscript and is often mandatory [5]. The cover letter provides an opportunity to highlight the importance of the study to journal editors. It should include key elements such as the manuscript title, journal name, a concise study description, its importance and its appeal to readers. Moreover, the cover letter should address matters such as the originality of the manuscript, clarify any conflicts of interest and provide contact information [5]. Researchers must invest time in preparing a well-written cover letter as an effective cover letter may tip the balance between immediate rejection or peer review. If GPT-4 may help with this process, researchers would be able to streamline their workflows and optimise their use of time.

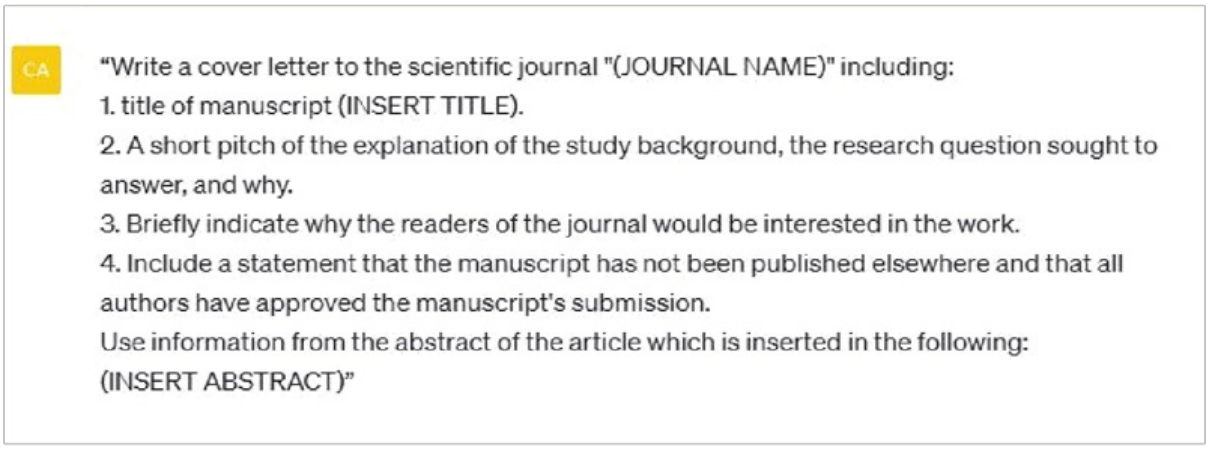
This produced the rationale of a non-inferiority study investigating whether GPT-4 may write cover letters that are just as good as human-written cover letters. Thus, the aim of this non-inferiority randomised study was to determine whether ChatGPT-4 is as good as humans in writing cover letters in terms of impression, readability, criteria satisfaction, degree of detail and preference.

## METHODS

This randomised, blinded non-inferiority study was reported according to the Consolidated Standards of Reporting Trials extension for non-inferiority and equivalence randomized trials guideline (CONSORT non-inferiority) [6] and the Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI) [7]. The study consisted of two parallel groups: one with GPT-4 generated cover letters and one with original cover letters written by human researchers. A total of eight assessors evaluated 4–8 cover letters each, from Marts 23 to April 25, ensuring that each pair of cover letters was assessed three times. No changes to the methods were made after the study had commenced.

Cover letters for previously published articles were used. The blinded assessors were recruited from our local network and needed at least a PhD degree. Assessors did not evaluate cover letters that they had contributed to. Thus, none of the assessors had seen any of the cover letters before evaluating them. The GPT-4-generated cover letters were created by feeding GPT-4 a pretested prompt based on key points for cover letters [5] and the abstracts of the articles (**Figure 1**). When feeding GPT-4 the abstract of the original source, the risk of plagiarism and hallucinations is low since it is given a source to create its content from.

**FIGURE 1** The prompt used for this study.




The generated cover letter was then transferred to a Word document with no adjustments to the content. Occasionally, GPT-4 would write "Dear [Editor's name]". If this occurred, we changed it to "Dear Editor". We conducted a pilot assessment before finalising the prompt for all cover letters. The pilot assessment involved sending one pair of cover letters to an assessor who was not part of the main assessor group.

Our primary intervention was assessed using a questionnaire in which blinded assessors evaluated cover letters on impression, readability, degree of detail and criteria satisfaction. The questionnaire underwent face validation and was reviewed three times by different people [8]. These parameters were assessed using a 1–10 Likert scale, where one represented a bad score and ten a good score. If the assessors rated the degree of detail

below ten, they were required to identify the reasons for their evaluation. They were given three options: too superficial, both lacking and too detailed in some areas, or too detailed. Furthermore, they were asked which cover letter they preferred and which they believed was written by GPT-4 or humans, and to provide an explanation of their choice. Assessors also had the option to select "I cannot decide" for both preference and guess. The option “I cannot decide” for guesses was grouped with incorrect answers in the subsequent data analysis. Our secondary outcome involved objective language analysis of the cover letters using Lix score [9] and Flesch-Kincaid grade level score [10].

Sample size calculation was based on continuous data with a Likert scale of 1–10. The least relevant difference was set to one point on the Likert scale. For the sample size calculation, standard deviation was set to one, alpha was set at 5%, 1-beta at 90%, and d, the non-inferiority limit, was set at one [11]. This produced a desirable sample size of 18 cover letters per group; i.e. a total of 36 cover letters. Randomisation was performed using random.org to determine the order in which human- or GPT-4-written cover letters were presented. The Likert scales were analysed as continuous data, whereas the degree of detail was considered ordinal categorical data. Continuous data were assessed for normality with the Shapiro-Wilk’s test and histograms. Normally distributed data were presented as means with 95% confidence intervals (CI) and analysed with the Student's t-test. Non-normally distributed data were presented as medians with ranges and analysed with the Mann-Whitney U test. The Statistical Package for the Social Sciences (SPSS version 28, IBM, US) was used to perform these tests. Categorical data, such as preference, were compared with the $\chi^2$ test and presented as crude rates. The significance level was set to a p value $\leq 0.05$. Qualitative data were gathered through comments from the questionnaires and grouped with similar answers. The Fleiss' Kappa formula was used for inter-rater agreement to assess subjective bias as there were three different assessors per cover letter [12]. The scales ranged from 0 to 1, where a score of $< 0.2$ was considered poor and $> 0.8$ was considered almost perfect [13]. The Danish Data Protection Agency was consulted on whether permission was necessary but concluded that no permission was needed as our study did not include personal data. Additionally, our study was exempt from ethics committee approval as it did not meet the criteria requiring project permission according to Danish legislation [14].

*Trial registration:* This study was not registered before the study commenced.

## RESULTS

**Table 1** presents the results from the questionnaires regarding assessments of the cover letters and the Lix and Flesch-Kincaid scores. No significant difference was found between human and GPT-4-written cover letters regarding impression, readability, and degree of detail. Human-written cover letters had a significantly higher score on criteria satisfaction based on the assessors’ opinion of necessary points. However, the median difference was only one point on the 1-10 Likert scale, which was predefined as a non-relevant difference. The degree of detail was inadequate for GPT-4 in 48% (n = 26) of the cover letters, whereas 18% (n = 10) were inadequate due to superficiality. In contrast, the degree of detail was inadequate for 52% (n = 28) of the human-written cover letters, of which 24% (n = 13) were due to an excessive level of detail (Table 1). Overall, the results indicated that GPT-4 was non-inferior to humans in writing cover letters.

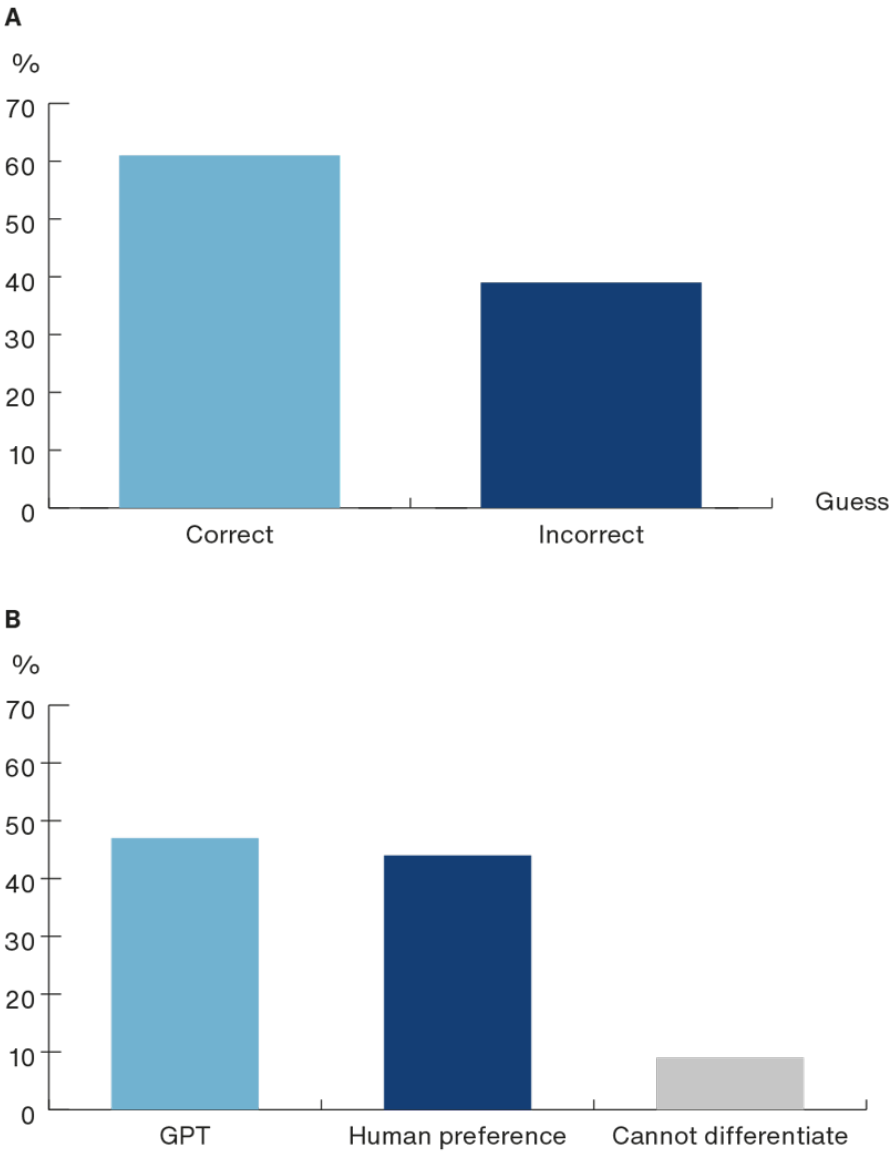
**TABLE 1** Summary of results. The blinded assessors' evaluation of the cover letters and the Lix and Flesch Kincaid grade level.

| | GPT-4 | Human | p value |
|---|---|---|---|
| *Evaluation criteria* | | | |
| Impression, median (range)[a] | 9 (7-10) | 9 (6-10) | 0.681 |
| Readability, median (range)[a] | 10 (5-10) | 9 (6-10) | 0.961 |
| Criteria satisfaction, median (range)[a] | 9 (6-10) | 10 (4-10) | < 0.001 |
| Degree of detail, median (range)[a] | 10 (5-10) | 9 (4-10) | 0.802 |
| Detail level, % (n): | | | |
| "Too superficial" | 18 (10) | 15 (8) | - |
| "Lacking or too detailed in some area" | 17 (9) | 13 (7) | - |
| "Too detailed, overall" | 13 (7) | 24 (13) | - |
| "No changes needed" | 52 (28) | 48 (26) | - |
| *Readability score, mean (± SD)* | | | |
| Lix | 64 (± 6) | 55 (± 3) | < 0.001 |
| Flesch Kincaid grade level | 16 (± 1) | 14 (± 2) | < 0.001 |

GPT = Generative Pre-Trained Transformer; SD = standard deviation.
a) 1-10 scale.

The blinded assessor's guess on whether the cover letters were written by GPT-4 or humans was correct in 61% (n = 33) of the cases and incorrect in 39% (n = 21) (**Figure 2** A). For those who guessed correctly, the reasons included the absence of a phrase about the ICMJE authorship criteria [15]. This is a specific style in the cover letter writing of our research group. Reasons for guessing correctly also included specific phrases and statements used for cover letters by our research group. They also noted that human-written cover letters tended to include more polite phrases. Regarding the question of which cover letter the assessors preferred, 47% (n = 25) preferred GPT-4-written cover letters, whereas 44% (n = 24) preferred human-written cover letters, and 9% (n = 5) had no preference (Figure 2 B). This difference was not significant.

**FIGURE 2** **A**. Blinded assessors' guess as to whether the cover letter was written by Generative Pre-Trained Transformer (GPT)-4 or humans. **B**. Blinded assessors' preferability of cover letters.




The mean Lix score for GPT-4 was nine points higher and the mean Flesch Kincaid grade level was two points higher, both indicating the use of longer sentences and words (Table 1). The Fleiss' Kappa values indicate that the agreement between assessors was poor, which meant that they, in general, had different assessments; impression $\kappa = -0.11$ (95% CI: −0.23-0.01), $p = 0.06$, readability $\kappa = -0.03$ (95% CI: −0.14-0.08), $p = 0.55$, criteria satisfaction $\kappa = 0.09$ (95% CI: −0.03-0.20), $p = 0.15$, and degree of detail $\kappa = 0.10$ (95% CI: −0.01-0.21), $p = 0.07$.

## DISCUSSION

We found that GPT-4 was just as good as humans in writing cover letters regarding the assessed parameters. No difference was found in the preference for cover letters written by GPT-4 or humans. GPT-4 had longer sentences and words. However, no difference was recorded in the subjective assessments of readability.

In short, we detected no significant or relevant difference between the cover letters. The only significantly different parameter was criteria satisfaction, but the difference fell within the definition of a non-relevant difference. Post-hoc testing of our prompt showed that GPT-4 could easily meet the criteria that the assessors indicated were missing. A similar distribution was found between the preference for human-written or GPT-4 cover letters. This supports previous results that GPT-4 may process scientific information, make the scientific research process faster and serve to simplify large portions of scientific information [3]. In terms of differentiation between the two types of cover letters, researchers were able to differentiate only 61% of the cover letters correctly. This was similar to results from another study, which found that researchers were unable to differentiate between GPT and human-written abstracts [4]. The reason why 61% guessed correctly may be that our research group has a specific approach to writing cover letters. In a post-hoc test, we found that this could have been incorporated into our prompt, which would have made it more difficult to differentiate between the cover letters.

For our objective readability assessment, GPT-4 had higher scores measured by both Lix and Flesch Kincaid grade levels. However, these differences may not matter since no difference was found in the subjective readability assessment. Moreover, editors had a high educational level, and higher readability levels are usually acceptable. Furthermore, Lix and Flesch Kincaid determine "readability" through the length of sentences and words included. GPT-4 is a language model that improves the structure and flow of language. According to the formulae for Lix and Flesch Kincaid evaluations, these calculations are commonly used to assess the complexity of written language, but they do not account for the impact of flow and structure on readability. Thus, they may not provide a complete understanding of language complexity and how it affects readers. We could have asked GPT-4 to make the language more readable or simpler, but the blinded assessors' subjective assessment indicated that this was unnecessary.

This study had several strengths. Firstly, it was reported according to CONSORT non-inferiority [6] and CONSORT-AI [7]. Secondly, our study employed a randomised blinded study with both quantitative and qualitative data investigating reliability on a subjective and objective basis. Thirdly, the questionnaire was face-validated three times before enrollment, which ensured that the questionnaire was appropriate and clear for the target group [8]. Fourthly, since inter-rater agreement was poor, it is a strength that every cover letter was assessed three times to reduce subjective bias. Lastly, assessors needed to have a PhD degree or higher, ensuring an assessor group with experience in writing and assessing cover letters. This study also had some limitations. Firstly, we used the first created answer from GPT-4 as GPT-4 generates different responses to the same prompt. This produced a cover letter with no additional refinements. When researchers use this prompt, they may require several versions to get the one they prefer, but we recommend using it as a drafting tool but not necessarily to create the final version. A risk exists of selection bias as the cover letters and assessors were recruited by sampling within our research group. The sample size was calculated before the study commencement regarding the non-inferiority design, but it may reduce the external validity that only 18 pairs of cover letters were included.

We are on the verge of a new era in which AI is becoming a part of our daily lives with growing use in both research and healthcare [2, 3, 16, 17]. AI has also entered the era of randomised controlled trials and the workflow of clinicians [17, 18]. With the growing use of AI and ChatGPT capable of writing articles in seconds, this could indicate that ChatGPT is a beneficial tool for researchers who need to streamline cover letters and save

time, leading to more effective and efficient scientific communication [3, 19, 20]. This study also shows that researchers may boost the equality of their cover letters, leading to equal chances of advancing to the peer-reviewing phase of the editorial process. Thus, AI also seems to be a potential help for young researchers that have no experience in writing cover letters. Future studies could investigate other AI tools and explore whether AI may improve other parts of scientific writing.

## CONCLUSION

GPT-4 written cover letters were as good as cover letters written by human researchers, and this may potentially streamline the process for researchers ensuring equal opportunities for peer-review advancement in the editorial process.

**Correspondence** *Can Deniz Deveci.* E-mail: can.d.deveci@gmail.com

**Accepted** 20 October 2023

**Conflicts of interest** Potential conflicts of interest have been declared. Disclosure forms provided by the authors are available with the article at ugeskriftet.dk/dmj

**Cite this as** Dan Med J 2023;70(12):A06230412



## REFERENCES

1. Mojadeddi ZM, Rosenberg J. The impact of AI and ChatGPT on research reporting. N Z Med J. 2023;136(1575):60-4.
2. OpenAI. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. https://openai.com/product/gpt-4/ (7 May 2023).
3. van Dis EAM, Bollen J, Zuidema W et al. ChatGPT: five priorities for research. Nature. 2023;614(7947):224-6. doi: 10.1038/D41586-023-00288-7.
4. Gao CA, Howard FM, Markov NS et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med. 2023;6(1):75. doi: 10.1038/s41746-023-00819-6.
5. Springer. Cover letters. www.springer.com/gp/authors-editors/authorandreviewertutorials/submitting-to-a-journal-and-peer-review/cover-letters/10285574 (7 May 2023).
6. Piaggio G, Elbourne DR, Pocock SJ et al. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA. 2012;308(24):2594-604. doi: 10.1001/jama.2012.87802.
7. Liu X, Rivera SC, Moher D et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-74. doi: 10.1038/s41591-020-1034-x.
8. Thomas SD, Hathaway DK, Arheart KL. Face validity. West J Nurs Res. 1992;14(1):109-12. doi: 10.1177/019394599201400111.
9. Kwichmann. LIX calculator. https://kwichmann.github.io/my_a2z/Week02/lix/ (6 Jun 2023).
10. Goodcalculators. Flesch Kincaid Calculator. https://goodcalculators.com/flesch-kincaid-calculator/ (6 Jun 2023).
11. Sealed Envelope. Power calculator for continuous outcome non-inferiority trial. www.sealedenvelope.com/power/continuous-noninferior/ (6 Jun 2023).
12. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Res Social Adm Pharm. 2013;9(3):330-8. doi: 10.1016/j.sapharm.2012.04.004.
13. Altman DG. Practical statistics for medical research. Chapman & Hall, 1999.
14. [Danish legislation]. www.regionh.dk/til-fagfolk/forskning-og-innovation/de-regionale-videnskabsetiske-komiteer/sider/hvilke-projekter-skal-jeg-anmelde.aspx (6 Jun 2023).

15. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. ICMJE, 2023. www.icmje.org/icmje-recommendations.pdf (6 Jun 2023).
16. Plana D, Shung DL, Grimshaw AA et al. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open. 2022;5(9):e2233946. doi: 10.1001/jamanetworkopen.2022.33946.
17. Kung TH, Cheatham M, Medenilla A et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198.
18. Microsoft and Epic expand strategic collaboration with integration of Azure OpenAI Service. Microsoft News Center, 2023. https://news.microsoft.com/2023/04/17/microsoft-and-epic-expand-strategic-collaboration-with-integration-of-azure-openai-service/ (6 Jun 2023).
19. Ariyaratne S, Iyengar KP, Nischal N et al. A comparison of ChatGPT-generated articles with human-written articles. Skeletal Radiol. 2023;52(9):1755-8. doi: 10.1007/s00256-023-04340-5.
20. Biswas, S. ChatGPT and the future of medical writing. Radiology. 2023;307/2)):e223312. doi: 10.1148/radiol.223312.