GPT for RCTs? Using AI to determine adherence to reporting guidelines

Wrightson, J.G. PhD[1], Blazey, P. MSc[2], Moher, D. PhD[3]. Khan, K.M. MD, PhD [1,4], Ardern, C.L. PhD [2,5,6]

Affiliations:

Department of Family Practice, University of British Columbia, BC, Canada

Centre for Aging SMART, University of British Columbia, BC, Canada

Ottawa Hospital Research Institute and University of Ottawa, ON, Canada

School of Kinesiology, University of British Columbia, BC, Canada

Department of Physical Therapy, University of British Columbia, BC, Canada

Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia

Key Points

Question: How accurately can Large Language Models determine adherence to clinical trial reporting guidelines?

Findings: In this cross-sectional study, the GPT-4 Large Langue Model accurately (~90%) determined reporting guidelines adherence in a sample of 113 randomized clinical trial trials. Following fine-tuning, the open-source Llama2 70B model achieved acceptable overall accuracy (~84%).

Meaning: A Large Language Model such as GPT-4 could be used by journals, peer reviewers and authors to quickly and accurately check clinical trial reporting guideline adherence.

Abstract

Importance

Adherence to established reporting guidelines can improve clinical trial reporting standards, but attempts to improve adherence have produced mixed results.

Objective

This exploratory study aimed to determine how accurate Large Language Model generative AI systems (AI-LLM) were for determining reporting guideline compliance in a sample of clinical trial reports.

Design, Setting, and Participants

In this cross-sectional study, the OpenAI GPT-4 and Meta LLama2 AI-LLM were evaluated for their ability to determine reporting guideline adherence in a sample of 113 published sports medicine and exercise science clinical trial reports. For each paper, the GPT-4-Turbo and Llama 2 70B models were prompted to answer a series of nine reporting guideline questions about the text of the article. The GPT-4-Vision model was prompted to answer two additional reporting guideline questions about the participant flow diagram in a subset of articles. The dataset was randomly split (80/20) into a TRAIN and TEST dataset. Hyperparameter and fine-tuning were performed using the TRAIN dataset. The Llama2 model was fine-tuned using the data from the GPT-4-Turbo analysis of the TRAIN dataset.

Main Outcome

Model performance (F1-score, classification accuracy) was assessed using the TEST dataset.

Results

Across all questions about the article text, the GPT-4-Turbo AI-LLM demonstrated acceptable performance (F1-score = 0.89, accuracy[95% CI] = 90%[85-94%]). Accuracy for all reporting guidelines was > 80%. The Llama2 model accuracy was initially poor (F1-score = 0.63, accuracy[95%CI] = 64%[57-71%]), and improved with fine-tuning (F1-score = 0.84, accuracy[95%CI] = 83%[77-88%]). The GPT-4-Vision model accurately identified all participant flow diagrams (accuracy[95% CI] = 100%[89-100%]) but was less accurate at identifying when details were missing from the flow diagram (accuracy[95% CI] = 57%[39-73%]).

Conclusions and Relevance

Both the GPT-4 and fine-tuned Llama2 AI-LLMs showed promise as tools for assessing reporting guideline compliance. Next steps should include developing an efficient, open-source AI-LLM and exploring methods to improve model accuracy.

Introduction

Poor reporting of clinical trials is common,[1] threatens the reliability and credibility of medical research [2] and affects patient care.[3] Improving trial reporting, therefore, is an ethical imperative.[4,5] Using reporting guidelines, such as the CONsolidated Standards of Reporting Trials (CONSORT), can improve trial reporting standards,[6–8] but adherence is often poor.[9] Following recent calls to evaluate the role of Artificial Intelligence (AI) in facilitating editorial and peer review decisions,[10] we assessed how well an AI model could determine reporting guideline adherence in clinical trial reports.

Medical journals often attempt to improve reporting standards by instructing authors to complete and submit reporting guideline checklists with their trial reports.[11] However, author-submitted checklists may not accurately reflect the contents of the report.[12] Other recent attempts to improve reporting standards involve training peer reviewers, authors, or editors with mixed results.[13,14] Using AI, specifically Large Language Model generative AI systems (AI-LLM), to perform these checks might save time and make the editorial process more efficient.[15] An AI-LLM can discern—with 80-90% accuracy—whether the content of computer science manuscripts corresponded to author-submitted submission checklists.[15] Given this success, generative AI systems hold promise for evaluating and improving adherence to reporting guidelines.

This exploratory research aimed to answer the following research question: How accurately can an AI-LLM measure reporting guideline compliance in a sample of clinical trial reports?

Method

This study was an exploratory retrospective data analysis. The study is reported in accordance with the Minimum Information about CLinical Artificial Intelligence Modeling (MI-CLAIM) standards,[16] and a completed MI-CLAIM checklist is available on the Open Science Framework (OSF; see the "Data availability statement" below).

Pilot study

In a pilot study, we tested whether an earlier version of OpenAI's GPT AI-LLM (GPT 3.5, San Francisco, USA) could measure reporting guideline compliance in a sample of Sports Medicine clinical trial reports. Details of the methods and results of that study can be found on the OSF. We used the same data in the present study.

Data

We used a sub-sample of the dataset provided by Schulz et al. (2022).[17] In their systematic review, Schulz and colleagues analyzed the reporting practices, including items from the CONSORT checklist, of 160 peer-reviewed scientific papers published in sports medicine journals in 2020. We chose this dataset because reporting practices in sports medicine are often inadequate[17–21] , and we expected these data would provide a severe test of the model's ability to identify reporting guideline adherence. We extracted text from all papers from the Schulz et al. dataset that were available in a machine-readable format. Details for the data extraction are shown in the R notebooks located on the OSF.

The data for open-access papers (n = 24) were extracted from the PubMed Central database in machine-readable form. The data for the remaining papers were extracted from articles with electronic ('Epub') or PDF files accessible by the study lead author (JW). Papers were removed from analysis if a) the text extraction contained errors or b) the electronic file was inaccessible. The by-journal distribution of papers included in the analysis is shown in Figure 1. Data were split into TRAIN (80% of text-question pairs) and TEST (20%) datasets. The split was stratified across the paper sections (Introduction/Method/Results). The characteristics of the datasets are shown in the Supplementary Materials Table S1. We did not create a validation data set because of the relatively low number of training examples.

Data extraction

The limit on the size of data (the 'context window') submitted to the GPT 3.5 (from the pilot study) and Meta's Llama 2 (Menlo Park, USA) AI models meant that entire papers could not be analyzed. Instead, we followed the example of Liu and Shah[15] and divided each paper into three sections: the Introduction, Method and Results. For each paper, nine pairs of a section of text from the paper (e.g. methods) and a question about the text (text-question pairs) were created to match the reporting guideline items that could be assessed using the AI model (see Table 1 below). For one item (see Question 9, Table 2), we combined the method and results section of the paper. This was necessary because the label provided in the Schulz et al. (2022) data specifically referred to the primary outcome, which may only have been identified in the method section of each paper. This question could not be analyzed using the Llama 2 model because of the smaller context window allowed for this model. The model response was limited to 512 tokens. Only the text-question pair was removed at this stage, and thus, in the final analysis, some papers did not have all the text-question pairs (Supplementary Material Table S1).

Reporting Guideline Items and Data Labelling

The extracted papers were assessed for adherence to eleven reporting guideline items. These items included the nine 'core' Methods and Results items from the CONSORT 2010 guidelines[14], with an additional item from the Introduction section (Table 1, Question 1) and Method section (Table 1, Question 3) sections. The text of each paper was assessed for adherence to nine reporting guideline items (Table 1, Items 1-9). The participant flow figures from a subset of the dataset were assessed for adherence to two further reporting items (Table 1, Items 10-11, see 'Image Analysis' below for details). Each reporting guideline item was assessed using a generative question and answering format, wherein the model was prompted to answer a question, formulated using natural language, about the text/image extracted from each paper. The model was required to summarize the text that was relevant to the question and answer YES or NO to the question. Each question corresponded to a variable from the labelled dataset provided by Schulz et al. (2022). The label ("ground truth") for each question ("YES" or "NO") was extracted from the systematic analysis by Schulz et al. (2022). Details for how each label was transformed into a "YES" or "NO" answer can be found in the raw data spreadsheets on the OSF.

Model Choice and Optimization

We tested three models: OpenAI's GPT-4-Turbo and GPT-4-Vision, and Meta's Llama 2 70B. In our pilot study, we found that an earlier version of OpenAIs AI-LLM (GPT-3.5) could achieve adequate performance in the same task but required significant tuning to do so. The newer GPT-4 models perform

significantly better on similar tasks.[15,22] We also tested the open-source Llama 2 AI-LLM. We included an open-source model because we envisage use cases where closed-source models, such as the GPT-4 family of models, cannot be used to check for reporting guideline adherence (e.g. for papers which contain commercially sensitive results).

For the GPT-4 analysis, we used the hyperparameter settings from our pilot study (Temperature = 0.2 and Top-P = 0.2; see the online materials for details of the hyperparameter tuning steps). At the time of this analysis, we did not have access to fine-tune the GPT-4 model. Instead, model tuning was achieved through iterative "prompt engineering". We initially used the system and user prompts from our pilot study, which were developed using the guidelines provided by OpenAI[23] and included asking the model to adopt a persona (the system prompt), using delimiters to distinguish parts of the input and specifying the steps required to complete the task (the user prompt). We then analyzed the training dataset using these prompts. We subsequently took the first ten examples from the training dataset that the model had incorrectly answered and used the OpenAI ChatGPT application to help us improve the system and user prompts by asking it to improve the language of the prompts to increase the likelihood that the model would respond correctly. The training dataset was rerun with these adjustments, and then a second round of examples was used to improve the prompts further. We used the text provided by ChatGPT for each prompt verbatim. The final system prompt attached to each question pair was as follows:
"You are a health researcher reviewing a scientific article for a peer-reviewed sports medicine journal. You will be supplied with text from the article, a description of a task, and a question (delimited by XML tags). Use only the information provided directly in the text to answer the question. Your answer must be accurate and precise. You must answer each question in steps. Delimit each step. Step 1: Summarize the information in the text that is relevant to the task description. Step 2: Answer the question with either 'YES' or 'NO' only."
ChatGPT suggested providing a task description and a question in the user prompt. For example, the user prompt given to the model with the introduction text was formatted thus:
"Task: Analyze the article focusing specifically on the details of the hypotheses used in the study. Note that acceptable hypotheses are clear, testable propositions or predictions that are specifically formulated for statistical analysis. Your task is to identify whether the hypotheses are clearly stated and if they meet these criteria. If the text only mentions hypotheses in general without clearly stating them or if they are not formulated for statistical analysis, your answer should be 'NO'.
Question: Does the article include the study hypotheses?"

For clarity, only the questions included in the user prompt are shown in Table 1. The final full user prompt (task and question) for each reporting guideline item is shown in the Supplementary Material Table S2.

The Llama 2 model was fine-tuned using the correctly answered questions from the GPT-4 analysis of the text training data. One of the reporting guideline items could not be included in this analysis (Table 2, Question 9) because this question required the text from the method and results to be passed to the model, exceeding the number of tokens allowed in the context window. The tuning was performed, and the model was hosted on the TogetherAI platform (San Francisco, USA). The analysis of the TEST dataset using the base and fine-tuned Llama 2 models was performed using the TogetherAI playground using the platform default hyperparameter values (Temperature and Top-P = 0.7).

Image analysis

During the study, OpenAI made their GPT-4-Vision model available for general use. We used this model to assess adherence to two further research guideline items (Table 2, Questions 10-11) related to the participant flow diagrams in a subset of papers (n = 20). We encoded the papers' participant flow diagram images in base64 and appended the system prompt and item questions to the encoded image using the same 'chat' format API call used to analyze the text. Ten of the papers contained a flow diagram which showed the number of participants randomized into each group (answer "YES" to Table 2, Question 10) but did not include reasons for exclusions and loss to follow-up (answer "NO" to Table 2, Question 11). These papers were identified using the labels from Schulz et al. (2022). Ten of these papers were in the dataset. The remainder had complete flow diagrams (answer "Yes" to both questions).

The PubMed Identifiers (PMID) for the papers included in the image analysis are available in the online materials. Because the number of suitable papers was so low, the included papers were drawn from both the test and training datasets. To ensure the model did not simply label any flow diagram as a participant flow diagram, a sample of 10 flow diagrams which did not show participant flow through a clinical trial was included in the dataset for one question (Table 2, Question 10). The details of these images are in the online materials.

Table 1. about here please

Analysis

Data extraction and analysis were performed using the R (version 4.3.2) and Python (version 3.8.17) programming languages. The primary outcome of this study was the F1-score (%) from the GPT-4 text analysis. The F1-score is the harmonic mean of the model precision (the ratio of true positives to the total number of identified positives) and recall (the ratio of true positives to the actual number of positive cases in the data). The F1-score controls for the expected class imbalances (YES or NO answers) in the dataset. The model classification accuracy (the ratio of true positives to the total number of cases) and associated 95% confidence interval (95%CI[24]) were also calculated.

Results

The breakdown of included papers (n = 113) by publication name is shown in Figure 1.

Figure 1. about here please

Text analysis

Model performance for the text analysis was evaluated in the TEST (20% held back) dataset.

GPT-4 Performance

The confusion matrix for the model performance in the analysis of the text from the TEST dataset is shown in Figure 2. The model performance on each question is shown in Table 2 (Questions 1-9) and in the Supplementary Material Table S3. Across all questions about the article text, the GPT-4-Turbo AI-LLM demonstrated acceptable performance (F1-score = 0.89, accuracy[95% CI] = 90%[85-94%]). Accuracy for the reporting guideline items ranged from 83-100%.

Figure 2. about here please

Llama 2 performance

Llama 2 model performance was evaluated in eight questions. Model accuracy for each question is shown in the Supplementary Material Table S4. The accuracy of the base model was low (F1-score = 0.63, accuracy[95%CI] = 64%[57-71%]). Fine tuning the model improved accuracy (F1-score = 0.84, accuracy[95%CI] = 83%[77-88%]).

Table 2. about here please

Image analysis

The flow diagrams from a subset of the data (n = 20) were assessed for adherence to two further reporting guideline items (Questions 10-11, Table 2). The GPT-4-Vision model accurately identified all flowcharts (accuracy[95% CI] = 100%[89-100%]) but was inaccurate at identifying when details were missing from the flow chart (accuracy[95% CI] = 57%[39-73%]).

Discussion

We wanted to determine how accurately AI-LLMs measured reporting guideline compliance in a sample of clinical trial reports. The OpenAI GPT-4-Turbo AI-LLM achieved ~90% accuracy across nine reporting guideline questions. Using an AI-LLM may help journal editors, peer reviewers, and authors check reporting guideline adherence quickly and accurately, which could reduce editorial workloads and research waste.[25]

Poor reporting of clinical trials negatively impacts downstream evidence synthesis, such as systematic reviews and clinical practice guidelines, and can impact care and public trust in science.[3,4,26,27] Interventions to improve clinical trial reporting guideline adherence have been examined, with inconsistent results.[13] Typically, the interventions target authors' or peer reviewers' behaviour[13,14] but may fail because they increase the already high workload for these groups.[13] Interventions targeting publishers and editors are less common.[14] There are many automated tools that journal publishers and editors use to screen manuscripts for errors and misconduct.[28] Using a trained AI-LLM, publishers could screen submitted publications for adherence to relevant reporting guidelines and flag to authors, peer reviewers and editors where details may be missing. This approach could allow publishers to improve reporting standards without adding to the workload of editors and peer reviewers.

The accuracy of the GPT-4 AI-LLM was slightly higher than the previous version of the model (GPT-3.5 ~86%) from our pilot study and that reported by Liu and Shah.[15] It is also better than the accuracy (<80%) of author-submitted CONSORT checklists reported by Blanco et al. (2019).[12] As expected, the Open-AI model was more accurate than the open-source Llama 2 model. However, fine-tuning the Llama 2 model significantly improved its accuracy. It is likely that any eventual tool that uses an AI-LLM to screen unpublished papers will be one in which the confidentiality of the text and the geolocation where the application is hosted can be controlled. For example, entities that publish clinical trials testing drug

efficacy may not wish, or be permitted, to submit potentially sensitive information to corporations whose servers are across international borders. Open-source models will allow these stakeholders to host and train models in secure environments.

Beyond establishing quantitative information regarding the accuracy of using AI models to assist in efficiently assessing the completeness of reporting clinical trials, it is also important to ask editors about the minimum accuracy thresholds they would accept. For example, would they agree that 70% of a clinical trial report needs to adhere to a reporting guideline checklist before they would consider peer review? When journals establish this information, prospective authors will likely want to use the same tools to ensure their clinical report meets the thresholds.

Developing models that allow users to pose natural language-style instructions and questions about images should expand the possibilities for AI-LLMs to evaluate scientific papers. We tested the model's ability to evaluate participant flow diagrams, but these technologies could also be used to examine the figures provided as evidence of risk of bias in systematic reviews or the contents of forest plots in meta-analyses. In a small subset of data, we found that the GPT-4-vision model could correctly identify all the participant flow diagrams. It was substantially less accurate when extracting information from the flow diagrams. This suggests that either the model will require tuning for it to be used to evaluate the contents of scientific figures and diagrams or that, at present, it is not sophisticated enough to perform this task.

The less-than-perfect accuracy of the AI-LLMs and the tendency of the current generation of AI-LLMs to "hallucinate" content[29] means that human confirmation of compliance is required. However, academics and scientists have long used imperfect automated tools to assess reporting standards (e.g.[30]). Our results suggest a similar role for AI-LLMs in efforts to improve clinical trial reporting. To help more clearly define the efficacy and limitations of AI-LLMs, future research comparing custom, well-trained AI models with author-completed checklists is warranted.

Limitations

The homogenous nature of the clinical trials included in this analysis may limit the generalizability of our findings to other fields. Any future AI-LLM developed to assess reporting guideline adherence will likely need training on (much larger) datasets that include clinical trials from a broad range of disciplines. Given that at least 75 clinical trials are published daily, this task appears achievable. We did not analyze full papers, something that is possible with GPT-4-Turbo because of its larger context window but not for the version of Llama 2 used here (though large context window Llama 2 models have been developed (e.g.

https://tinyurl.com/gpt4rct3). It is likely that analysis of whole papers will be necessary if AI-LLM are to be used to evaluate reporting guidelines compliance. For example, it will be necessary to know what the study hypothesis was when evaluating the statistical methods, results and discussion.

Conclusion

AI-LLMs were sufficiently accurate for assessing reporting guideline compliance in clinical trial reports. However, variations in accuracy across different items indicate that, at present, these tools can assist with, but not replace, human evaluation of reporting standards compliance.

Acknowledgements

All authors have seen and approved the manuscript form. We would like to thank Dr. Nihar Shah for his advice and feedback and the organizers and attendees of the Meta-Research Innovation Center at Stanford's international forum for their constructive feedback on our work.

Competing Interests

The authors declare they have no competing interests.

Funding

This work was supported by a CIHR Research Operating Grant (Scientific Directors) held by Karim Khan. The funder had no role in the design and conduct of the study.

Data availability statement

Code notebooks will be available on the Open Science Framework (relevant links within the text) upon publication. Copyright issues prevent us from sharing the text extracted from the papers used in this analysis; however, details of the steps needed to reproduce the extracted text from open and closed papers can be found within these notebooks.

References

1. Dechartres A, Trinquart L, Atal I, et al. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ*. 2017;357:j2490. doi:10.1136/bmj.j2490

2. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med*. 2010;8(1):24. doi:10.1186/1741-7015-8-24

3. Duff JM, Leather H, Walden EO, LaPlant KD, George TJ. Adequacy of Published Oncology Randomized Controlled Trials to Provide Therapeutic Details Needed for Clinical Application. *JNCI J Natl Cancer Inst*. 2010;102(10):702-705. doi:10.1093/jnci/djq117

4. Chalmers I. Underreporting research is scientific misconduct. *JAMA*. 1990;263(10):1405-1408. doi:10.1001/jama.1990.03440100121018

5. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *The Lancet*. 2014;383(9912):156-165. doi:10.1016/S0140-6736(13)62229-1

6. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med*. 2010;8:18. doi:10.1186/1741-7015-8-18

7. Pandis N, Shamseer L, Kokich VG, Fleming PS, Moher D. Active implementation strategy of CONSORT adherence by a dental specialty journal improved randomized clinical trial reporting. *J Clin Epidemiol*. 2014;67(9):1044-1048. doi:10.1016/j.jclinepi.2014.04.001

8. Hopewell S, Ravaud P, Baron G, Boutron I. Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis. *BMJ*. 2012;344:e4178. doi:10.1136/bmj.e4178

9. Samaan Z, Mbuagbaw L, Kosa D, et al. A systematic scoping review of adherence to reporting guidelines in health care literature. *J Multidiscip Healthc*. 2013;6:169-188. doi:10.2147/JMDH.S43952

10. Ioannidis JPA, Berkwits M, Flanagin A, Bloom T. Peer Review and Scientific Publication at a Crossroads: Call for Research for the 10th International Congress on Peer Review and Scientific Publication. *JAMA*. 2023;330(13):1232-1235. doi:10.1001/jama.2023.17607

11. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane reviewa. *Syst Rev*. 2012;1(1):60. doi:10.1186/2046-4053-1-60

12. Blanco D, Biggane AM, Cobo E, et al. Are CONSORT checklists submitted by authors adequately reflecting what information is actually reported in published papers? *Trials*. 2018;19(1):80. doi:10.1186/s13063-018-2475-0

13. Speich B, Mann E, Schönenberger CM, et al. Reminding Peer Reviewers of Reporting Guideline Items to Improve Completeness in Published Articles: Primary Results of 2 Randomized Trials. *JAMA Netw Open*. 2023;6(6):e2317651. doi:10.1001/jamanetworkopen.2023.17651

14. Blanco D, Altman D, Moher D, Boutron I, Kirkham JJ, Cobo E. Scoping review on interventions to improve adherence to reporting guidelines in health research. *BMJ Open*. 2019;9(5):e026589. doi:10.1136/bmjopen-2018-026589

15. Liu R, Shah NB. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. Published online June 1, 2023. doi:10.48550/arXiv.2306.00622

16. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y

17. Schulz R, Langen G, Prill R, Cassel M, Weissgerber TL. Reporting and transparent research practices in sports medicine and orthopaedic clinical trials: a meta-research study. *BMJ Open*. 2022;12(8). doi:10.1136/bmjopen-2021-059347

18. Caldwell AR, Vigotsky AD, Tenan MS, et al. Moving Sport and Exercise Science Forward: A Call for the Adoption of More Transparent Research Practices. *Sports Med Auckl NZ*. 2020;50(3):449-459. doi:10.1007/s40279-019-01227-1

19. Twomey R, Yingling V, Warne J, et al. The Nature of Our Literature: A Registered Report on the Positive Result Rate and Reporting Practices in Kinesiology. *Commun Kinesiol*. 2021;1(3). doi:10.51224/cik.v1i3.43

20. Murphy J, Mesquida C, Warne J. A Survey on the Attitudes Towards and Perception of Reproducibility and Replicability in Sports and Exercise Science. *Commun Kinesiol*. 2023;1(5). doi:10.51224/cik.2023.53

21. Mesquida C, Murphy J, Lakens D, Warne J. Replication concerns in sports and exercise science: a narrative review of selected methodological issues in the field. *R Soc Open Sci*. 2022;9(12):220946. doi:10.1098/rsos.220946

22. OpenAI. GPT-4 Technical Report. Published online March 27, 2023. doi:10.48550/arXiv.2303.08774

23. OpenAI Platform. Accessed March 1, 2024. https://platform.openai.com

24. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*. 1934;26(4):404-413. doi:10.2307/2331986

25. Moher D, Glasziou P, Chalmers I, et al. Increasing value and reducing waste in biomedical research: who's listening? *The Lancet*. 2016;387(10027):1573-1586. doi:10.1016/S0140-6736(15)00307-4

26. Altman DG. The scandal of poor medical research. *BMJ*. 1994;308(6924):283-284. doi:10.1136/bmj.308.6924.283

27. Heneghan C, Mahtani KR, Goldacre B, Godlee F, Macdonald H, Jarvies D. Evidence based medicine manifesto for better healthcare. *BMJ*. 2017;357:j2973. doi:10.1136/bmj.j2973

28. Bordewijk EM, Li W, van Eekelen R, et al. Methods to assess research misconduct in health-related research: A scoping review. *J Clin Epidemiol*. 2021;136:189-202. doi:10.1016/j.jclinepi.2021.05.012

29. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. *Crit Care*. 2023;27(1):180. doi:10.1186/s13054-023-04473-y

30. Straumsheim C. What Is Detected? Inside Higher Ed. Accessed October 25, 2023. https://www.insidehighered.com/news/2015/07/14/turnitin-faces-new-questions-about-efficacy-plagiarism-detection-software

Tables

Table 1. Number of text-question pairs and adherence for each guideline

#	Paper Section	Reporting Guideline Item	Pairs (n)	Adherence (%)
1	Introduction	The study hypotheses	113	65%
2	Method	The primary outcome	108	50%
3	Method	The sample size	108	67%
4	Method	The eligibility criteria	108	77%
5	Method	The implementation of the randomization sequence	108	58%
6	Method	The randomization methods	108	41%
7	Method	The random allocation sequence & assignment	108	26%
8	Method	Blinding	108	49%
9	Results	Standardized effect sizes and confidence intervals	113	20%
10	Results	The number of participants randomly assigned	20*	100%*
11	Results	The number of participants lost to follow-up	20*	50%*

* Subset of the full dataset; see 'Image analysis' for details

Table 2. Accuracy of the GPT-4 models for each reporting guideline item

#	Question	Accuracy [95%CI]
Text Analysis		
1	Does the article include the study hypotheses?	83% [63-93%]
2	Does the article define the primary outcome or primary endpoint?	89% [73-96%]
3	Does the Method include how the sample size was determined?	90% [71-97%]
4	Does the article define the study's eligibility criteria?	86% [67-95%]
5	Does the article provide a detailed description of the specific techniques used for the practical implementation of the randomization sequence?	94% [72-99%]
6	Does the article include the randomization methods used in the study?	94% [73-99%]
7	Does the article EXPLICITLY detail the individuals or bodies or roles involved in either GENERATING the random allocation sequence, or ENROLLING participants, or ASSIGNING participants to trial groups?	86% [67-95%]
8	Does the article include details about who, if anyone, was blinded to the assignment to interventions?	100% [87-100%]
9	For the PRIMARY outcome, do the results reported in the article include the standardized effect size and confidence interval?	87% [68-95%]
Image Analysis		
10	Does this image contain a CONSORT flow diagram showing, for each group, the number of participants who were randomly assigned, received the intended treatment, and were analysed for the primary outcome?	100% [89-100%]*
11	Does this image display a CONSORT flow diagram that details both the number of participants lost to follow-up and excluded after randomization AND specifies the reasons for these losses in each group?	57% [39-73%]*

* = Analysis performed on a subset (n=20) of the dataset

Figures

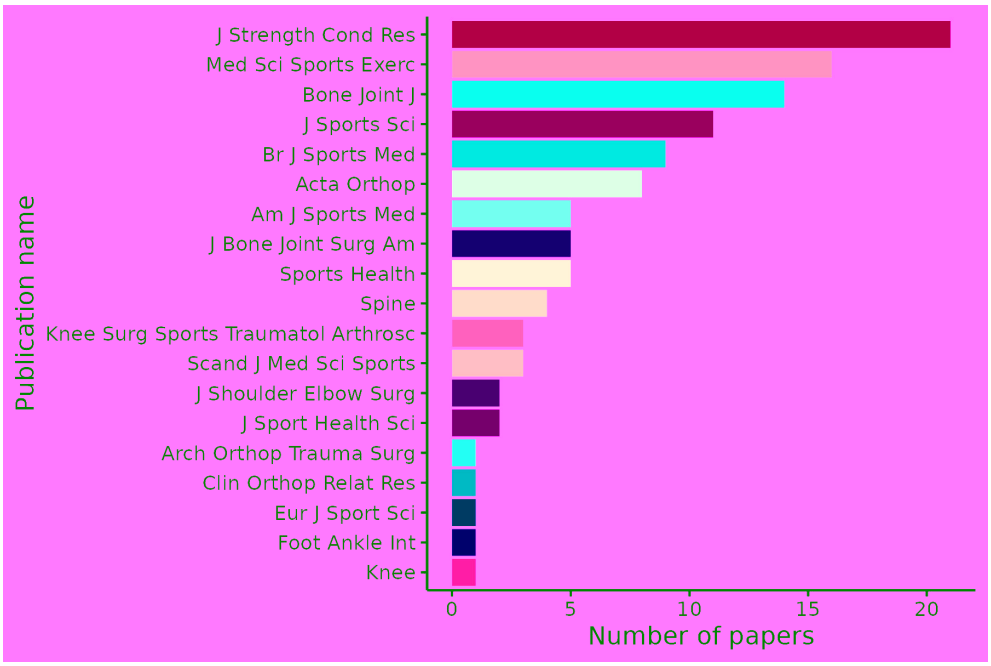

Figure 1. Breakdown of included papers by journal name.

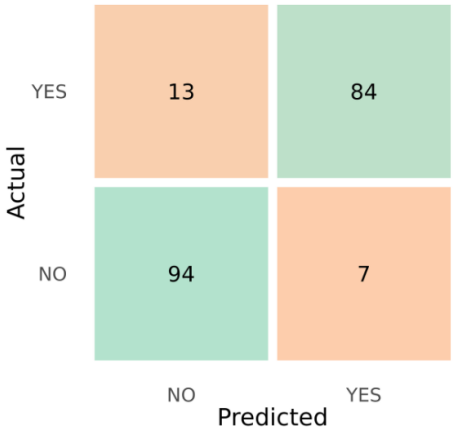

Figure 2. Confusion matrix from the analysis of the TEST dataset. Actual (y-axis) represents the answers given by the GPT-4-turbo model and Predicted (x-axis) represents the labels for each question extracted from Schulz et al. (2022).

Supplementary Material

Table S1: Description of the TRAIN and TEST datasets

	TEST N=198[1]	TRAIN N=84[1]
Papers (n, unique)	96	113
Paper Section (n, %)		
Introduction	23 (12%)	90 (11%)
Method	152 (77%)	604 77%)
Results	23 (12%)	90 (11%)
Questions (n, %)		
Does the article include the study hypotheses?	23 (12%)	90 (11%)
Does the article define the primary outcome or primary endpoint?	28 (21%)	80 (10%)
Does the Method include how the sample size was determined?	21 (11%)	87 (11%)
Does the article define the study's eligibility criteria?	22 (11%)	86 (11%)
Does the article provide a detailed description of the specific techniques used for the practical implementation of the randomization sequence?	16 (8%)	92 (12%)
Does the article include the randomization methods used in the study?	17 (9%)	91 (12%)
Does the article EXPLICITLY detail the individuals or bodies or roles involved in either GENERATING the random allocation sequence, or ENROLLING participants, or ASSIGNING participants to trial groups?	22 (11%)	86 (11%)
Does the article include details about who, if anyone, was blinded to the assignment to interventions?	26 (13%)	82 (10%)
For the PRIMARY outcome, do the results reported in the article include the standardized effect size and confidence interval?	23 (12%)	90 (11%)

[1] = n question-text pairs

Table S2 Full user prompt for each text reporting guideline item

Guideline Item	User prompt
The study hypotheses	"Task: Analyze the article focusing specifically on the details of the hypotheses used in the study. Note that acceptable hypotheses are clear, testable propositions or predictions that are specifically formulated for statistical analysis. Your task is to identify whether the hypotheses are clearly stated and if they meet these criteria. If the text only mentions hypotheses in general without clearly stating them or if they are not formulated for statistical analysis, your answer should be 'NO'. Question: Does the article include the study hypotheses?"
The primary outcome	"Task: The primary outcome or primary endpoint is the pre-specified outcome considered to be of greatest importance to relevant stakeholders and is usually used in the sample size calculation. Question: Does the article define the primary outcome or primary endpoint?"
The sample size	"Task: Analyze the article focusing specifically on the details of how the sample size was determined. Question: Does the Method include how the sample size was determined?"
The eligibility criteria	"Task: Analyze the article for detailed information about the predefined study inclusion and exclusion criteria. Inclusion and exclusion criteria are specific conditions predefined by researchers that determine who can or cannot participate in the study. If the terms 'inclusion criteria' or 'exclusion criteria' are not present in the article, you must answer NO. Question: Does the article define the study's eligibility criteria?"
The implementation of the randomization sequence	"Task: Identify whether the article provides a detailed description of the specific techniques used for the implementation of the randomization sequence. Randomization techniques are practical techniques used to generate randomization sequences or store each person's random assignment. Examples include random number tables or lists, computer software or website-generated randomization methods, sealed envelopes, and coin toss. Question: Does the article provide a detailed description of the specific techniques used for the practical implementation of the randomization sequence?"
The randomization methods	"Task: Analyze the article focusing specifically on the details of the randomization method used in the study. Note that acceptable randomization methods include covariate adaptive randomization, stratified randomization, block randomization, simple randomization, minimization, factorial randomization, and cluster randomization. Your task is to identify whether any of these specific randomization methods are mentioned or described in the text. If the text only mentions randomization in general without specifying any of these methods, your answer should be NO. Question: Does the article include the randomization methods used in the study?"
The random allocation sequence & assignment	"Task: Examine the article to determine if it explicitly details the individuals or bodies or roles involved in either generating the random allocation sequence, enrolling participants, or assigning participants to trial groups in the study.

	Question: Does the article EXPLICITLY detail the individuals or bodies or roles involved in either GENERATING the random allocation sequence, or ENROLLING participants, or ASSIGNING participants to trial groups?"
Blinding	"Task: Analyze the article to determine if it includes details about who was blinded after the assignment to interventions in the trial. The term 'blinding' or 'masking' refers to withholding information about the assigned interventions from people involved in the trial who may potentially be influenced by this knowledge. Examples include no blinding, or participants, care providers, or assessors being blinded. If the article does not include these details, your answer should be 'NO.' Question: Does the article include details about who, if anyone, was blinded to the assignment to interventions?"
Standardized effect sizes and confidence intervals	"Task: Identify the primary outcome and determine if the article reports a standardized effect size and confidence interval for the PRIMARY outcome only. Examples include odds ratios, hazard ratios, risk ratios, risk difference, standardized mean differences, Cohen's d. Question: For the PRIMARY outcome, do the results reported in the article include the standardized effect size and confidence interval?"
The number of participants randomly assigned	"Question 1: Does this image contain a CONSORT flow diagram showing, for each group, the number of participants who were randomly assigned, received the intended treatment, and were analysed for the primary outcome?"
The number of participants lost to follow-up	"Question 2: Does this image display a CONSORT flow diagram that details both the number of participants lost to follow-up and excluded after randomization AND specifies the reasons for these losses in each group?")

Table S3: GPT-4 models performance for each question in the TEST dataset

Question	FN	FP	TN	TP
Does the article include the study hypotheses?	2	2	7	12
Does the article define the primary outcome or primary endpoint?	0	3	10	15
Does the Method include how the sample size was determined?	0	2	8	12
Does the article define the study's eligibility criteria?	3	0	10	12
Does the article provide a detailed description of the specific techniques used for the practical implementation of the randomization sequence?	1	0	6	9
Does the article include the randomization methods used in the study?	1	0	11	5
Does the article EXPLICITLY detail the individuals or bodies or roles involved in either GENERATING the random allocation sequence, or ENROLLING participants, or ASSIGNING participants to trial groups?	3	0	11	8
Does the article include details about who, if anyone, was blinded to the assignment to interventions?	0	0	16	10
For the PRIMARY outcome, do the results reported in the article include the standardized effect size and confidence interval?	3	0	18	2
Does this image contain a CONSORT flow diagram showing, for each group, the number of participants who were randomly assigned, received the intended treatment, and were analysed for the primary outcome?	0	0	20*	10
Does this image display a CONSORT flow diagram that details both the number of participants lost to follow-up and excluded after randomization AND specifies the reasons for these losses in each group?	1	2	8	9

FN; False negative, FP; False positive, TN; True negative, TP; True positive. *; includes 10 flow diagrams not taken from the papers in dataset

Table S4. Llama 2 performance for the assessed reporting guideline items

#	Question	Accuracy [95%CI]
1	Does the article include the study hypotheses?	83% [63-93%]
2	Does the article define the primary outcome or primary endpoint?	93% [77-98%]
3	Does the Method include how the sample size was determined?	86% [65-95%]
4	Does the article define the study's eligibility criteria?	68% [47-84%]
5	Does the article provide a detailed description of the specific techniques used for the practical implementation of the randomization sequence?	87% [62-96%]
6	Does the article include the randomization methods used in the study?	71% [47-87%]
7	Does the article EXPLICITLY detail the individuals or bodies or roles involved in either GENERATING the random allocation sequence, or ENROLLING participants, or ASSIGNING participants to trial groups?	76% [55-89%]
8	Does the article include details about who, if anyone, was blinded to the assignment to interventions?	96% [80-99%]