

# 1 GPT for RCTs?: Using AI to determine 2 adherence to reporting guidelines 3

4 Wrightson, J.G.<sup>1</sup>, Blazey, P.<sup>2</sup>, Moher, D<sup>3</sup>. Khan, K.M.<sup>1,4</sup>, Ardern, C.L.<sup>2,5,6</sup>

5  
6 Affiliations:

- 7 1. Department of Family Practice, University of British Columbia, BC, Canada
- 8 2. Centre for Aging SMART, University of British Columbia, BC, Canada
- 9 3. Ottawa Hospital Research Institute and University of Ottawa, ON, Canada
- 10 4. School of Kinesiology, University of British Columbia, BC, Canada
- 11 5. Department of Physical Therapy, University of British Columbia, BC, Canada
- 12 6. Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia

13  
14 Corresponding author:

15 Dr. James Wrightson  
16 317 - 2194 Health Sciences Mall  
17 Vancouver, B.C. V6T 1Z3  
18 Canada  
19 [j.wrightson@ubc.ca](mailto:j.wrightson@ubc.ca)  
20

21 This article is a preprint and has not been peer-reviewed. It reports new medical research that  
22 has yet to be evaluated and so should not be used to guide clinical practice.

## 23 Abstract

24 Background: Adherence to established reporting guidelines can improve clinical trial reporting  
25 standards, but attempts to improve adherence have produced mixed results. This exploratory  
26 study aimed to determine how accurate a Large Language Model generative AI system (AI-LLM)  
27 was for determining reporting guideline compliance in a sample of sports medicine clinical trial  
28 reports.

29 Design and Methods: This study was an exploratory retrospective data analysis. The OpenAI  
30 GPT-4 and Meta LLama2 AI-LLMa were evaluated for their ability to determine reporting guideline  
31 adherence in a sample of 113 published sports medicine and exercise science clinical trial reports.  
32 For each paper, the GPT-4-Turbo and Llama 2 70B models were prompted to answer a series of  
33 nine reporting guideline questions about the text of the article. The GPT-4-Vision model was  
34 prompted to answer two additional reporting guideline questions about the participant flow  
35 diagram in a subset of articles. The dataset was randomly split (80/20) into a TRAIN and TEST  
36 dataset. Hyperparameter and fine-tuning were performed using the TRAIN dataset. The Llama2  
37 model was fine-tuned using the data from the GPT-4-Turbo analysis of the TRAIN dataset.

38 Primary outcome measure: Model performance (F1-score, classification accuracy) was assessed  
39 using the TEST dataset.

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

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

50

51 Keywords

52 Machine Learning, peer review, large language model, clinical trials

## 53 Introduction

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

61  
62 Clinical trial reporting standards are improved when medical journals instruct authors to use  
63 reporting guideline checklists, but not all journals require adherence to reporting guidelines, and  
64 enforcement of adherence is inconsistent [11–14]. Recent attempts to improve reporting  
65 standards involve training peer reviewers, authors, or editors, in the use of reporting guideline  
66 checklists with mixed results [15,16]. It might be possible to improve guideline adherence by  
67 providing feedback to authors, either prior to submission or during peer review [16].

68  
69 Using machine learning and AI, especially Large Language Model generative AI systems (AI-  
70 LLM), to check for adherence to reporting guidelines might save authors, reviewers and  
71 publishers time and make the editorial process more efficient [17,18]. An AI-LLM can discern—  
72 with 80-90% accuracy—whether the content of computer science manuscripts corresponds to  
73 reporting guideline checklists [17]. Given this success, generative AI systems hold promise for  
74 evaluating and improving adherence to reporting guidelines.

75  
76 There is increasing interest in the rigour and transparency, or lack thereof, of sports medicine,  
77 exercise science and orthopedic research [19]. Reporting in sports medicine and exercise science  
78 papers is often inadequate, and there are concerns about the reproducibility and veracity of many  
79 findings [19–23]. Improving reporting practices should be a priority for researchers and publishers  
80 [19,24] in all fields, including sports medicine. This exploratory research aimed to answer the  
81 following research question: How accurately can an AI-LLM measure reporting guideline  
82 compliance in a sample of sports medicine clinical trial reports?

## 83 Method

84 This study was an exploratory retrospective data analysis. The study is reported in accordance  
85 with the Minimum Information about CLinical Artificial Intelligence Modeling (MI-CLAIM) standards  
86 [25], and a completed MI-CLAIM checklist is available on the Open Science Framework (OSF;  
87 <https://tinyurl.com/gpt4rct1>).

## 88 Pilot study

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

## 93 Data

94 We used a sub-sample of the dataset provided by Schulz et al. (2022) [19]. In their systematic  
95 review, Schulz and colleagues analyzed the reporting practices, including items from the  
96 CONSORT checklist, of 160 peer-reviewed scientific papers published in Sports Medicine  
97 journals in 2020. Journals were identified using the Scimago Journal Rank indicator (see Schulz  
98 et al. 2022 [19] for details). We extracted all papers from the Schulz et al. dataset that were  
99 available in full-text machine-readable format. Details for the data extraction are shown in the R  
100 notebooks located on the OSF: <https://tinyurl.com/gpt4rctdata>.

101  
102 The data for open-access papers (n=24) were extracted from the PubMed Central database in  
103 machine-readable form. The data for the remaining papers were extracted from articles with  
104 electronic ('Epub') or PDF files accessible by the study lead author (JW). Papers were removed  
105 from analysis if a) the text extraction contained errors or b) the electronic file was inaccessible.  
106 The by-journal distribution of papers included in the analysis is shown in Figure 1. Data were split  
107 into TRAIN (80% of text-question pairs) and TEST (20%) datasets. The split was stratified across  
108 the paper sections (Introduction/Method/Results) so that a single paper could be used in both the  
109 TEST and TRAIN datasets (e.g., the Method section in the TEST dataset and the Results section  
110 in the TRAIN dataset). The characteristics of the datasets are shown in the Supplementary

111 Materials Table S1. We did not create a validation data set because of the relatively low number  
112 of training examples.

113

114 Data extraction

115 The limit on the size of data (the ‘context window’) submitted to the GPT 3.5 (from the pilot study)  
116 and Meta’s Llama 2 (Menlo Park, USA) AI models meant that entire papers could not be analyzed.  
117 Instead, we followed the example of Liu and Shah [17] and divided each paper into three sections:  
118 the Introduction, Method and Results. For each paper, nine pairs of a section of text from the  
119 paper (e.g. methods) and a question about the text (text-question pairs) were created to match  
120 the reporting guideline items that could be assessed using the AI model (see Table 1 below). For  
121 one item (see Question 9, Table 2) we combined the method and results section of the paper.  
122 This was necessary because the label provided in the Schulz et al (2022) data specifically referred  
123 to the primary outcome, which may only have been identified in the method section of each paper.  
124 This question could not be analyzed using the Llama 2 model because of the smaller context  
125 window allowed for this model. The model response was limited to 512 tokens. Only the text-  
126 question pair was removed at this stage, and thus, in the final analysis, some papers did not have  
127 all the text-question pairs (Supplementary Material Table S1).

## 128 Reporting Guideline Items and Data Labelling

129 The extracted papers were assessed for adherence to eleven reporting guideline items. These  
130 items included the nine ‘core’ Method and Results items from the CONSORT 2010 guidelines  
131 [16], with an additional item from the Introduction section (Table 1, Question 1) and Method  
132 section (Table 1, Question 3) sections. The text of each paper was assessed for adherence to  
133 nine reporting guideline items (Table 1, Items 1-9). The participant flow figures from a subset of  
134 the dataset were assessed for adherence to two further reporting items (Table 1, Items 10-11,  
135 see ‘Image Analysis’ below for details). Each reporting guideline item was assessed using a  
136 generative question and answering format, wherein the model was prompted to answer a  
137 question, formulated using natural language, about the text/image extracted from each paper.  
138 The model was required to summarize the text that was relevant to the question, and answer YES  
139 or NO to the question. Each question corresponded to a variable from the labelled dataset  
140 provided by Schulz et al (2022). The label (“ground truth”) for each question (“YES” or “NO”) was  
141 extracted from the systematic analysis by Schulz et al. (2022). Details for how each label was  
142 transformed into a “YES” or “NO” answer can be found in the raw data spreadsheets  
143 (<https://tinyurl.com/gpt4rctdata>).

## 144 Model Choice and Optimization

145 We tested three models; OpenAI's GPT-4-Turbo and GPT-4-Vision, and Meta's Llama 2 70B. In  
146 our pilot study, we found that an earlier version of OpenAI's AI-LLM (GPT-3.5) could achieve  
147 adequate performance in the same task, but required significant tuning to do so. The newer GPT-  
148 4 models perform significantly better on similar tasks [17,26]. We also tested the open-source  
149 Llama 2 AI-LLM. We included an open-source model because we envisage use cases where  
150 authors may want to run models locally or wish to support open-source systems in accordance  
151 with open science principles [27].

152

153 For the GPT-4 analysis, we used the hyperparameter settings from our pilot study (Temperature  
154 = 0.2 and Top-P = 0.2, see the pilot study <https://tinyurl.com/gpt4rct1> for details of the  
155 hyperparameter tuning steps). Unfortunately, at the time of this analysis, we did not have access  
156 to fine-tune the GPT-4 model. Model tuning in the present study was achieved through iterative  
157 "prompt engineering". We initially used the system and user prompts from our pilot study which  
158 were developed using the guidelines provided by OpenAI [28] and included asking the model to  
159 adopt a persona (the system prompt), using delimiters to distinguish parts of the input and  
160 specifying the steps required to complete the task (the user prompt). We then analyzed the  
161 training dataset using these prompts. We subsequently took the first 10 examples from the  
162 training dataset that the model had incorrectly answered, and used the OpenAI ChatGPT  
163 application to help us improve the system and user prompts by asking it to improve the language  
164 of the prompts to increase the likelihood that the model would respond correctly. The training  
165 dataset was rerun with these adjustments, and then a second round of examples was used to  
166 improve the prompts further. We used the text provided by ChatGPT for each prompt verbatim.  
167 The final system prompt attached to each question pair was as follows:

168 *"You are a health researcher reviewing a scientific article for a peer-reviewed sports medicine*  
169 *journal. You will be supplied with text from the article, a description of a task, and a question*  
170 *(delimited by XML tags). Use only the information provided directly in the text to answer the*  
171 *question. Your answer must be accurate and precise. You must answer each question in steps.*  
172 *Delimit each step. Step 1: Summarize the information in the text that is relevant to the task*  
173 *description. Step 2: Answer the question with either 'YES' or 'NO' only."*

174 ChatGPT suggested providing a task description and a question in the user prompt. For example,  
175 the user prompt given to the model with the introduction text was formatted thus:

176 *“Task: Analyze the article focusing specifically on the details of the hypotheses used in the study.*  
177 *Note that acceptable hypotheses are clear, testable propositions or predictions that are*  
178 *specifically formulated for statistical analysis. Your task is to identify whether the hypotheses are*  
179 *clearly stated and if they meet these criteria. If the text only mentions hypotheses in general*  
180 *without clearly stating them or if they are not formulated for statistical analysis, your answer should*  
181 *be 'NO'. Question: Does the article include the study hypotheses?”*

182 For clarity, only the questions included in the user prompt are shown in Table 1. The final full user  
183 prompt for each reporting guideline item is shown in the Supplementary Material Table S2.

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

## 192 Image analysis

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

203

204 The PubMed Identifiers (PMID) for the papers included in the image analysis are available in the  
205 online materials. Because the number of suitable papers was so low, the included papers were  
206 drawn from both the test and training datasets. To ensure the model did not simply label any flow  
207 diagram as a participant flow diagram, a sample of 10 flow diagrams which did not show



208 participant flow through a clinical trial was included in the dataset for one question (Table 2,  
209 Question 10). The details of these images are in the online materials.

210

211 Table 1. The number of text-question pairs for each reporting guideline in the dataset, and the  
212 proportion of papers that adhered to 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%*          |

213 \* Subset of the full dataset, see 'Image analysis' for details

## 214 Analysis

215 Data extraction and analysis were performed using the R (version 4.3.2) and Python (version  
216 3.8.17) programming languages. The primary outcome of this study was the F1-score (%) from  
217 the GPT-4 text analysis. The F1-score is the harmonic mean of the model precision (the ratio of



218 true positives to the total number of identified positives) and recall (the ratio of true positives to  
219 the actual number of positive cases in the data). The F1-score controls for the expected class  
220 imbalances (YES or NO answers) in the dataset. The model classification accuracy (the ratio of  
221 true positives to the total number of cases) and associated 95% confidence interval (95%CI, [29])  
222 were also calculated.

## 223 Results

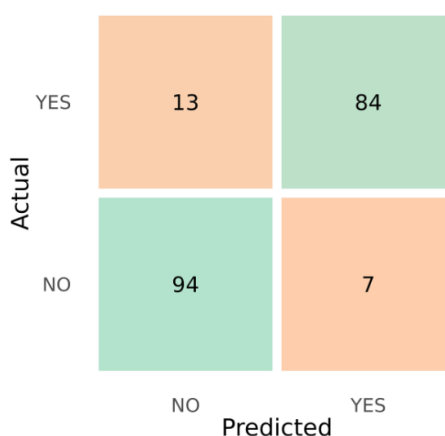
224 The breakdown of included papers (n = 113) by publication name is shown in the Supplementary  
225 Material Figure S1.

## 226 Text analysis

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

### 228 GPT-4 Performance

229 The confusion matrix for the model performance in the analysis of the text from the TEST dataset  
230 is shown in Figure 1. The model performance on each question is shown in Table 2 (Questions  
231 1-9) and in the Supplementary Material Table S3. Across all questions about the article text, the  
232 GPT-4-Turbo AI-LLM demonstrated acceptable performance (F1-score = 0.89, accuracy[95% CI]  
233 = 90%[85-94%]). Accuracy for the reporting guideline items ranged from 83-100%.



234  
235  
236 Figure 1. Confusion matrix from the analysis of the TEST dataset. Actual (y-axis) represents the  
237 answers given by the GPT-4-turbo model, and Predicted (x-axis) represents the labels for each  
238 question extracted from Schulz et al (2022).

## 239 Llama 2 performance

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

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

| #                            | Question   | Accuracy [95%CI] |
|------------------------------|--|------------------|
| <b><u>Text Analysis</u></b>  |  |                  |
| 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%]     |
| <b><u>Image Analysis</u></b> |  |                  |
| 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

## 245 Image analysis

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

## 250 Discussion

251 We wanted to determine how accurate AI-LLMs were for measuring reporting guideline  
252 compliance in a sample of clinical trial reports. The OpenAI GPT-4-Turbo AI-LLM achieved ~90%  
253 accuracy across nine reporting guideline questions. Using an AI-LLM may help journal editors,  
254 publishers, peer reviewers and authors check reporting guideline adherence quickly and  
255 accurately, which could reduce both editorial workloads and waste in research[30].

256  
257 Poor reporting of clinical trials negatively impacts downstream evidence synthesis, such as  
258 systematic reviews and clinical practice guidelines and can impact care and public trust in science  
259 [3,4,31,32]. Interventions to improve clinical trial reporting guideline adherence have been  
260 examined, but the results have been inconsistent [15]. Typically, these interventions target  
261 authors' or peer reviewers' behaviour [15,16] but may fail because they increase the already high  
262 workload for these groups [15]. Interventions targeting publishers are less common [16]. There  
263 are many automated tools that journal publishers and editors use to screen manuscripts for errors  
264 and misconduct [33]. Using a trained AI-LLM, publishers could screen submitted publications for  
265 adherence to relevant reporting guidelines and flag to authors, peer reviewers and editors where  
266 details may be missing. This approach could allow publishers to improve reporting standards  
267 without substantially increasing the workload on editors and peer reviewers.

268  
269 The accuracy of the GPT-4 AI-LLM was slightly higher than the previous version of the model  
270 (GPT-3.5 ~86%) from our pilot study (<https://tinyurl.com/gpt4rct1>) and that reported by Liu and  
271 Shah [17]. As expected, the Open-AI model was more accurate than the open-source Llama 2  
272 model. However, fine-tuning the Llama 2 model significantly improved its accuracy. It is likely that  
273 any eventual tool that uses an AI-LLM to screen unpublished papers will be one in which the  
274 confidentiality of the text, and the geolocation where the application is hosted, can be controlled.  
275 For example, entities that publish clinical trials testing drug efficacy may not wish, or be allowed,

276 to submit potentially sensitive information to corporations whose servers are across international  
277 borders. Open-source models will allow these stakeholders to host and train models in secure  
278 environments.

279  
280 The development of models that allow users to pose natural language-style instructions and  
281 questions about images should expand the possibilities for AI-LLMs to evaluate the contents of  
282 scientific papers. We tested the model's ability to evaluate participant flow diagrams, but these  
283 technologies could also be used to examine the figures provided as evidence of risk of bias in  
284 systematic reviews or the contents of forest plots in meta-analyses. In a small subset of data, we  
285 found that the GPT-4-vision model could identify all of the participant flow diagrams correctly. It  
286 was substantially less accurate when extracting information from the flow diagrams. This suggests  
287 that either the model will require tuning for it to be used to evaluate the contents of scientific  
288 figures and diagrams, or that at present it is not sophisticated enough to perform this task.

289  
290 The less-than-perfect accuracy of the AI-LLMs and the tendency of the current generation of AI-  
291 LLMs to "hallucinate" content [34] means that human confirmation of compliance is required.  
292 However, academics and scientists have long used imperfect automated tools to assess reporting  
293 standards (e.g. [35]). Our results suggest a similar role for AI-LLMs in efforts to improve clinical  
294 trial reporting. To help more clearly define the efficacy and limitations of AI-LLMs, future research  
295 comparing custom, well-trained AI models with author-completed checklists is warranted.

296  
297 The primary limitation of our study was the (small and homogenous) dataset. The homogenous  
298 nature of the clinical trials included in this dataset, and the poor adherence to reporting standards  
299 in sports medicine clinical trial reports [19,21,36] limits the generalizability of our findings to other  
300 fields. Indeed, it is likely that any future AI-LLM developed to assess reporting guideline  
301 adherence will need training on (much larger) datasets that include clinical trials from a broad  
302 range of disciplines. Given that at least 75 clinical trials are published daily, this task appears  
303 achievable.

304  
305 It was beyond the scope of this exploratory study to create a large, well-labelled dataset for model  
306 tuning. The model was therefore trained using data generated by the model (the correct answers  
307 and extracted text from the training data), possibly increasing tendencies to hallucinate and other  
308 errors. Nevertheless, acceptable model performance was achieved, and results should improve  
309 with larger samples. We did not analyze full papers, something that is possible with GPT-4-Turbo

310 because of its larger context window but not for the version of Llama 2 used here (though large  
311 context window Llama 2 models have been developed (e.g. <https://tinyurl.com/gpt4rct3>), though  
312 at the time of analysis we were unable to fine-tune these models on the TogetherAI platform.

313  
314 It is likely that analysis of whole papers will be necessary if AI-LLM are to be used to evaluate  
315 reporting guidelines compliance. For example, it will be necessary to know what the study  
316 hypothesis was when evaluating the statistical methods, results and discussion. Beyond  
317 establishing quantitative information regarding the accuracy of using AI models to assist in an  
318 efficient assessment of the completeness of reporting clinical trials, it is also likely important to  
319 ask a broad spectrum of editors about accuracy thresholds they regard as minimal. For example,  
320 would there be consensus that 70% of a clinical trial report needs to adhere to a reporting  
321 guideline checklist, or should the consensus be 85%, before a journal agrees to consider the  
322 paper for peer review? When journals establish this information, prospective authors may wish to  
323 use the same tools to ensure their clinical report meets these thresholds.

## 324 Conclusion

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

## 328 Acknowledgements

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

## 332 Competing Interests

333 The authors declare they have no competing interests

## 334 Funding

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

## 337 Data availability

338 Code notebooks and data are available on the Open Science Framework  
339 (<https://tinyurl.com/gpt4rctdata>). Copyright issues prevent the sharing of some of the text  
340 extracted from the papers used in this analysis; however, details of the steps needed to reproduce  
341 the extracted text from open and closed papers can be found within these notebooks.



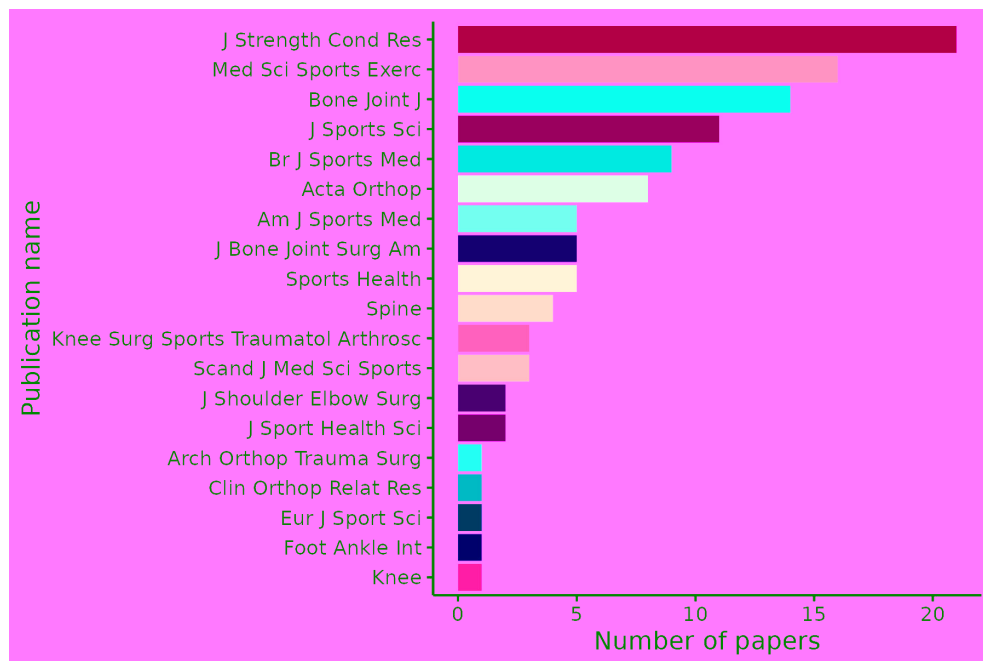
## 342 References

- 343 1 Dechartres A, Trinquart L, Atal I, *et al.* Evolution of poor reporting and inadequate methods  
344 over time in 20 920 randomised controlled trials included in Cochrane reviews: research on  
345 research study. *BMJ*. 2017;357:j2490.
- 346 2 Simera I, Moher D, Hirst A, *et al.* Transparent and accurate reporting increases reliability,  
347 utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC*  
348 *Med*. 2010;8:24.
- 349 3 Duff JM, Leather H, Walden EO, *et al.* Adequacy of Published Oncology Randomized  
350 Controlled Trials to Provide Therapeutic Details Needed for Clinical Application. *JNCI J Natl*  
351 *Cancer Inst*. 2010;102:702–5.
- 352 4 Chalmers I. Underreporting research is scientific misconduct. *JAMA*. 1990;263:1405–8.
- 353 5 Chalmers I, Bracken MB, Djulbegovic B, *et al.* How to increase value and reduce waste  
354 when research priorities are set. *The Lancet*. 2014;383:156–65.
- 355 6 Schulz KF, Altman DG, Moher D, *et al.* CONSORT 2010 Statement: updated guidelines for  
356 reporting parallel group randomised trials. *BMC Med*. 2010;8:18.
- 357 7 Pandis N, Shamseer L, Kokich VG, *et al.* Active implementation strategy of CONSORT  
358 adherence by a dental specialty journal improved randomized clinical trial reporting. *J Clin*  
359 *Epidemiol*. 2014;67:1044–8.
- 360 8 Hopewell S, Ravaud P, Baron G, *et al.* Effect of editors' implementation of CONSORT  
361 guidelines on the reporting of abstracts in high impact medical journals: interrupted time  
362 series analysis. *BMJ*. 2012;344:e4178.
- 363 9 Samaan Z, Mbuagbaw L, Kosa D, *et al.* A systematic scoping review of adherence to  
364 reporting guidelines in health care literature. *J Multidiscip Healthc*. 2013;6:169–88.
- 365 10 Ioannidis JPA, Berkwits M, Flanagan A, *et al.* Peer Review and Scientific Publication at a  
366 Crossroads: Call for Research for the 10th International Congress on Peer Review and  
367 Scientific Publication. *JAMA*. 2023;330:1232–5.
- 368 11 Sarkis-Onofre R, Poletto-Neto V, Cenci MS, *et al.* CONSORT endorsement improves the  
369 quality of reports of randomized clinical trials in dentistry. *J Clin Epidemiol*. 2020;122:20–6.
- 370 12 Turner L, Shamseer L, Altman DG, *et al.* Consolidated standards of reporting trials  
371 (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs)  
372 published in medical journals. *Cochrane Database Syst Rev*. Published Online First: 2012.  
373 doi: 10.1002/14651858.MR000030.pub2
- 374 13 McErlean M, Samways J, Godolphin PJ, *et al.* The reporting standards of randomised  
375 controlled trials in leading medical journals between 2019 and 2020: a systematic review. *Ir*  
376 *J Med Sci 1971 -*. 2023;192:73–80.
- 377 14 Wang P, Wolfram D, Gilbert E. Endorsements of five reporting guidelines for biomedical  
378 research by journals of prominent publishers. *PLOS ONE*. 2024;19:e0299806.
- 379 15 Speich B, Mann E, Schönenberger CM, *et al.* Reminding Peer Reviewers of Reporting  
380 Guideline Items to Improve Completeness in Published Articles: Primary Results of 2  
381 Randomized Trials. *JAMA Netw Open*. 2023;6:e2317651.
- 382 16 Blanco D, Altman D, Moher D, *et al.* Scoping review on interventions to improve adherence  
383 to reporting guidelines in health research. *BMJ Open*. 2019;9:e026589.
- 384 17 Liu R, Shah NB. ReviewerGPT? An Exploratory Study on Using Large Language Models for  
385 Paper Reviewing. 2023. <https://doi.org/10.48550/arXiv.2306.00622>
- 386 18 Wang F, Schilsky RL, Page D, *et al.* Development and Validation of a Natural Language  
387 Processing Tool to Generate the CONSORT Reporting Checklist for Randomized Clinical  
388 Trials. *JAMA Netw Open*. 2020;3:e2014661.
- 389 19 Schulz R, Langen G, Prill R, *et al.* Reporting and transparent research practices in sports  
390 medicine and orthopaedic clinical trials: a meta-research study. *BMJ Open*. 2022;12. doi:

- 391 10.1136/bmjopen-2021-059347  
392 20 Caldwell AR, Vigotsky AD, Tenan MS, *et al.* Moving Sport and Exercise Science Forward: A  
393 Call for the Adoption of More Transparent Research Practices. *Sports Med Auckl NZ.*  
394 2020;50:449–59.  
395 21 Twomey R, Yingling V, Warne J, *et al.* The Nature of Our Literature: A Registered Report on  
396 the Positive Result Rate and Reporting Practices in Kinesiology. *Commun Kinesiol.* 2021;1.  
397 doi: 10.51224/cik.v1i3.43  
398 22 Murphy J, Mesquida C, Warne J. A Survey on the Attitudes Towards and Perception of  
399 Reproducibility and Replicability in Sports and Exercise Science. *Commun Kinesiol.* 2023;1.  
400 doi: 10.51224/cik.2023.53  
401 23 Mesquida C, Murphy J, Lakens D, *et al.* Replication concerns in sports and exercise  
402 science: a narrative review of selected methodological issues in the field. *R Soc Open Sci.*  
403 2022;9:220946.  
404 24 Hansford HJ, Cashin AG, Wewege MA, *et al.* Open and transparent sports science  
405 research: the role of journals to move the field forward. *Knee Surg Sports Traumatol*  
406 *Arthrosc.* 2022;30:3599–601.  
407 25 Norgeot B, Quer G, Beaulieu-Jones BK, *et al.* Minimum information about clinical artificial  
408 intelligence modeling: the MI-CLAIM checklist. *Nat Med.* 2020;26:1320–4.  
409 26 OpenAI. GPT-4 Technical Report. 2023. <https://doi.org/10.48550/arXiv.2303.08774>  
410 27 Open source for open science. CERN. 2023. [https://home.cern/science/computing/open-](https://home.cern/science/computing/open-source-open-science)  
411 [source-open-science](https://home.cern/science/computing/open-source-open-science) (accessed 30 May 2023)  
412 28 OpenAI Platform. <https://platform.openai.com> (accessed 1 March 2024)  
413 29 Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of  
414 the Binomial. *Biometrika.* 1934;26:404–13.  
415 30 Moher D, Glasziou P, Chalmers I, *et al.* Increasing value and reducing waste in biomedical  
416 research: who's listening? *The Lancet.* 2016;387:1573–86.  
417 31 Altman DG. The scandal of poor medical research. *BMJ.* 1994;308:283–4.  
418 32 Heneghan C, Mahtani KR, Goldacre B, *et al.* Evidence based medicine manifesto for better  
419 healthcare. *BMJ.* 2017;357:j2973.  
420 33 Bordewijk EM, Li W, van Eekelen R, *et al.* Methods to assess research misconduct in  
421 health-related research: A scoping review. *J Clin Epidemiol.* 2021;136:189–202.  
422 34 Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. *Crit Care.*  
423 2023;27:180.  
424 35 Straumsheim C. What Is Detected? High. Ed.  
425 <https://www.insidehighered.com/news/2015/07/14/turnitin-faces-new-questions-about->  
426 [efficacy-plagiarism-detection-software](https://www.insidehighered.com/news/2015/07/14/turnitin-faces-new-questions-about-) (accessed 25 October 2023)  
427 36 Bullock GS, Ward P, Impellizzeri FM, *et al.* Up front and open, shrouded in secrecy, or  
428 somewhere in between? A Meta Research Systematic Review of Open Science Practices in  
429 Sport Medicine Research. *J Orthop Sports Phys Ther.* 2023;1–32.  
430  
431

432 Supplementary Material

433



434

435 Figure S1. Breakdown of included papers by journal name.

Table S1: Description of the TRAIN and TEST datasets

|   | <b>TEST,</b><br>N = 198 <sup>1</sup> | <b>TRAIN,</b><br>N = 784 <sup>1</sup> |
|---|--------------------------------------|---------------------------------------|
| Papers (n, unique)  | 96                                   | 113                                   |
| Paper Section (n, %)  |                                      |                                       |
| Introduction  | 23<br>(12%)                          | 90<br>(11%)                           |
| Method  | 152<br>(77%)                         | 604<br>(77%)                          |
| Results   | 23<br>(12%)                          | 90<br>(11%)                           |
| Questions (n, %)  |                                      |                                       |
| Does the article include the study hypotheses?  | 23<br>(12%)                          | 90<br>(11%)                           |
| Does the article define the primary outcome or primary endpoint?  | 28<br>(21%)                          | 80<br>(10%)                           |
| Does the Method include how the sample size was determined?   | 21<br>(11%)                          | 87<br>(11%)                           |
| Does the article define the study's eligibility criteria?   | 22<br>(11%)                          | 86<br>(11%)                           |
| Does the article provide a detailed description of the specific techniques used for the practical implementation of the randomization sequence?   | 16<br>(8%)                           | 92<br>(12%)                           |
| Does the article include the randomization methods used in the study?   | 17<br>(9%)                           | 91<br>(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<br>(11%)                          | 86<br>(11%)                           |
| Does the article include details about who, if anyone, was blinded to the assignment to interventions?  | 26<br>(13%)                          | 82<br>(10%)                           |
| For the PRIMARY outcome, do the results reported in the article include the standardized effect size and confidence interval?   | 23<br>(12%)                          | 90<br>(11%)                           |

<sup>1</sup> = n question-text pairs

Table S2 Full user prompt for each text reporting guideline item

| Guideline Item                                   | User prompt   |
|--|---|
| The study hypotheses                             | <i>"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?"</i>   |
| The primary outcome                              | <i>"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?"</i>   |
| The sample size                                  | <i>"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?"</i>  |
| The eligibility criteria                         | <i>"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?"</i>  |
| The implementation of the randomization sequence | <i>"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?"</i>   |
| The randomization methods                        | <i>"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?"</i> |
| The random allocation sequence & assignment      | <i>"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?"</i>   |

|  |  |
|--|--|
| Blinding   | <i>"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?"</i> |
| Standardized effect sizes and confidence intervals | <i>"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?"</i>   |
| The number of participants randomly assigned       | <i>"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?"</i>  |
| The number of participants lost to follow-up       | <i>"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?"</i>  |

439

440 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

441



442

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%]     |

443