# GPT for RCTs?: Using AI to measure adherence to reporting guidelines

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

Affiliations:

1. Faculty of Medicine, University of British Columbia, B.C. Canada
2. Centre for Aging SMART, University of British Columbia, BC, Canada
3. Department of Family Practice, University of British Columbia, BC, Canada
4. School of Kinesiology, University of British Columbia, BC, Canada
5. Department of Physical Therapy, University of British Columbia, BC, Canada
6. Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia

Corresponding author:

Dr. Clare Ardern
2177 Wesbrook Mall
Vancouver, BC V6T 1Z3
Canada
clare.ardern@ubc.ca

# Abstract

Background: Adherence to established reporting guidelines can improve clinical trial reporting standards, but attempts to improve adherence have produced mixed results. This exploratory study aimed to determine how accurately a Large Language Model generative AI system (AI-LLM) could measure reporting guideline compliance in a sample of sports medicine clinical trial reports.

Methods: The OpenAI GPT-3.5 AI-LLM was evaluated for its ability to determine reporting guideline adherence in a sample of 113 published sports medicine and exercise science clinical trial reports. For each paper, the model was prompted to answer a series of nine reporting guideline questions. The dataset was randomly split (80/20) into a TRAIN and TEST dataset. Hyperparameter and model fine-tuning were performed using the TRAIN dataset. Model performance (F1-score, classification accuracy) was assessed using the TEST dataset.

Results: Across all questions, the AI-LLM demonstrated acceptable performance (F1-score = 86%). However, there was significant variation in performance between different reporting guideline questions (accuracy between 70-100%). The model was most accurate when asked to identify a defined primary objective or endpoint and least accurate when asked to identify an effect size and related confidence interval.

Discussion: The AI-LLM showed promise as a tool for assessing reporting guideline compliance. Next steps should include developing a cost-effective, open-source AI-LLM and exploring methods to improve model accuracy.

Keywords

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

# 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, but results have been mixed [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.

There is increasing interest in the rigour and transparency, or lack thereof, of Sports Medicine, Exercise Science and Orthopaedic research [16]. Reporting in sports medicine and exercise science papers is often inadequate, and there are concerns about the reproducibility and veracity of many findings [16–20]. Improving reporting practices in sports medicine should be a priority for researchers and publishers [16,21]. This exploratory research aimed to answer the following research question: How accurately can an AI-LLM measure reporting guideline compliance in a sample of sports medicine 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 [22], and a completed MI-CLAIM checklist is available at: https://osf.io/tyx5s/?view_only=b7c57738230a4eb29d7b1a358f806761

## Data

We used a sub-sample of the dataset provided by Schulz et al. (2022) [16]. 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. Journals were identified using the Scimago Journal Rank indicator (see Schulz et al. 2022 [16] for details). We extracted all papers from the Schulz et al. dataset that were available in full-text machine-readable format. Details for the data extraction are shown in the R notebooks located on the Open Science Framework: https://osf.io/4shmt/?view_only=f0ee0ac3225444b9a198edad5f78a147.

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 Table 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 (a minimum of 50 examples are required for model fine-tuning).

### Data extraction

The limit on the size of data submitted to the AI model meant that entire papers could not be analyzed. Instead, we followed the example of Liu and Shah [15] and split 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). Data were removed if the word count of any text-question pair was too long. The latter was necessary because, at the time of analysis, the OpenAI API [23] limited the size of the data for model fine-tuning (4096 tokens, ~ 3500 words). 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 (Table 1). Full details of these steps are shown in the notebooks.

## Reporting guideline items

Each paper was assessed for adherence to nine reporting guideline items, modified from eleven items in the 2010 CONSORT parallel group randomised trials checklist [6]. An initial list of eleven reporting guideline items was extracted from those used in the analysis of Schulz et al. (2022). These questions were piloted independently on a sample of five trial reports by two of us (JW, PB). From this analysis, nine questions that could be answered using individual sections of a paper were developed. These questions include most of the previously identified “core” CONSORT questions [24]. The questions were amended to meet the model's prompt requirements and clarify text identified as ambiguous in the pilot. Questions that required analysis of multiple paper sections (e.g., required an analysis of text in both the Method and Results sections) were excluded. Details and rationales for these amendments can be found in the supplementary material.

At the time of the analysis (April 2023-September 2023), prompts to the OpenAI GPT-3.5 model used in this study (see Model Selection and Optimization below) could not include images. Therefore, questions that required figures (e.g. a CONSORT flow chart) or tables were not included. Included questions, the corresponding CONSORT checklist item and the adherence in the sample dataset are shown in Table 1. The relevance of these questions for reporting standards is detailed elsewhere [16,25]

## Data labelling

Each text-question pair was labelled with a single-word answer to the question: "YES” or “NO”. The initial label (“ground truth”) for each question was extracted from the systematic analysis by Schulz et al. (2022). For full details of that analysis and the complete analysis dataset, see Schulz et al. (2022). For some papers, labels were adjusted by the lead author (JW) if the information contained in the paper (for a YES answer) was not in the relevant section text (e.g.

the study hypothesis was not in the Introduction section) or we had amended the question so that the label provided by Schulz et al. was no longer correct (e.g. the paper included confidence intervals for the primary outcome but not the secondary outcome, and thus would be labelled “YES” in the present study but may have been labelled “NO” in the study by Schulz et al. Details of these changes are in the online material.

## Model Selection and Optimization

We used the OpenAI GPT AI-LLM [23]. Although the newer GPT4 AI-LLm was available via an API, at the time of analysis (April-September 2023), only the GPT-3.5 turbo model could be fine-tuned via an API. Prompts were developed using the guidelines provided by OpenAI and included asking the model to adopt a persona, using delimiters to distinguish parts of the input and specifying the steps required to complete the task. The final prompt provided to the model was:

*“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 and a question (delimited by XML tags). Use the article text to answer the question. You must answer the question in steps. Delimit each step. Step 1: Summarize the information in the text relevant to the question. Step 2: Answer the question 'YES' or 'NO'".*

The response was limited to 512 tokens. The model was tuned using three text-question pairs per paper (one each from the Introduction, Method and Results sections) in the TRAIN dataset. The two hyperparameters tuned were “Temperature” and “Top-P”, which control the randomness and diversity of text generated by the model [23]. Values of 0.2, 0.5 or 1.0 were tested for both hyperparameters, where lower values make the mode output more deterministic (values of 0 for both were rejected during pilot testing). The values that resulted in the highest model accuracy (the F1-score, see Analysis) were chosen. The model was subsequently fine-tuned using the OpenAI ‘fine-tuning’ system using the OpenAI Python library (3,555,801 tokens, epochs = 3, ‘training loss’ = 0.0104). The data submitted as examples for tuning included the answers (“YES”/”NO”) and the relevant text from the paper extracted by the model for all correctly answered questions from the TRAIN dataset. Model robustness/sensitivity was assessed using text perturbations [26] (see Supplementary material for rationale and methods).

## 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 (%). 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, [27]) were also calculated.

# Results

The breakdown of included papers by publication name is shown in Figure 1. The questions, associated CONSORT items, number of text-question pairs and adherence of included papers are shown in Table 1.

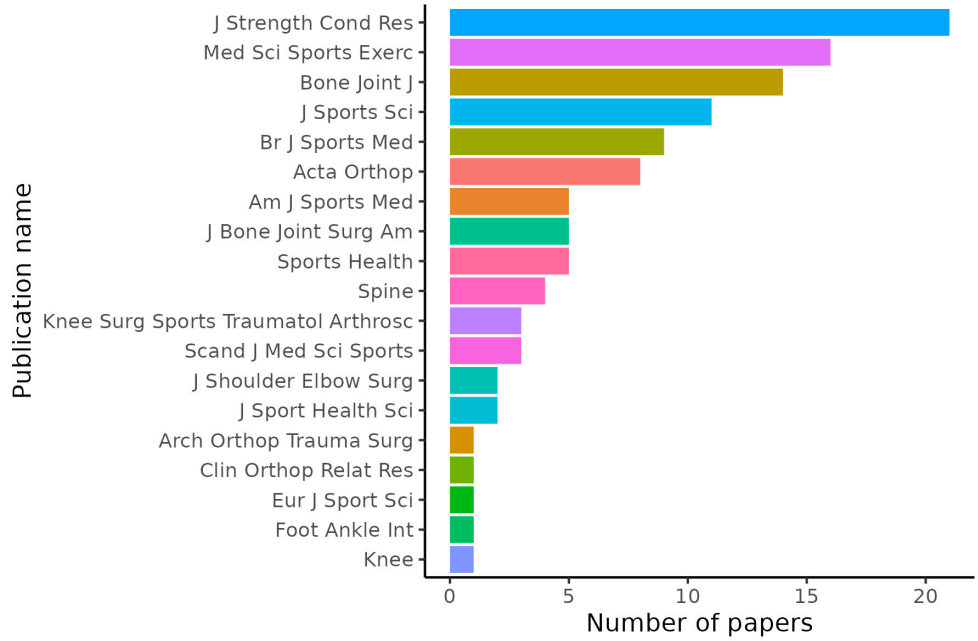


Figure 1. Breakdown of included papers by journal name.

Table 1. Questions, associated CONSORT checklist item, number of text-question pairs in the dataset and % of papers with a YES answer

| Section | Question | CONSORT | Pairs (n) | "YES" |
|---|---|---|---|---|
| Introduction | Are the hypotheses for the study included in the Introduction? | 2b | 113 | 58% |
| Method | Does the Method define the pre-specified primary outcome measure or primary endpoint? | 6a | 108 | 44% |
| Method | Does the Method include how the sample size was determined? | 7a | 108 | 66% |
| Method | Does the Method include the eligibility criteria for the participants? | 4a | 108 | 77% |
| Method | Does the Method include the method used to generate the random allocation sequence? | 8a | 108 | 58% |
| Method | Does the Method include the type of randomisation and details of any restriction (such as blocking and block size)? | 8b | 108 | 47% |
| Method | Does the Method include the mechanism used to implement the random allocation sequence and any steps taken to conceal the sequence? | 9 | 108 | 26% |
| Method | Does the Method include who was blinded after assignment to interventions? | 11a | 108 | 45% |
| Results | Do the Results include the estimated effect size and confidence interval? | 17a | 113 | 30% |

## Model Optimization

Hyperparameter values of Temperature = 0.2 and Top-P = 0.2 resulted in the most accurate model of the TRAIN dataset (F1 score = 82%, accuracy[95%CI] = 81%[79-84%], precision = 0.82, recall = 0.82).

## Model Performance

Model Performance was evaluated in the TEST (20% held back) dataset. Two models were compared for performance: the base model and a model fine-tuned on the correctly categorized

question-text pairs from the TRAIN dataset. The fine-tuned model was more accurate (F1 score = 86%, accuracy [95%CI] = 86% [80-90%], Figure 2) than the base model (F1 score = 79%, accuracy [95%CI] = 80% [74-85%]). The confusion matrix for the fine-tuned model performance in the TEST dataset is shown in Figure 2. The fine-tuned model performance on each question in the TEST dataset is shown in Table 2. The Supplementary Material (Table S2) shows model performance for each question. Results of the sensitivity analysis (see Supplementary Material for details) suggested the model was robust to perturbations to the text (F1 score = 82-85%).

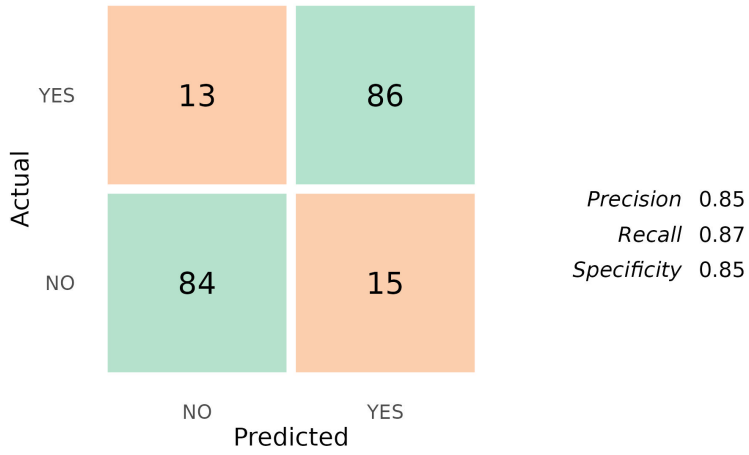


Figure 2. Confusion matrix from the analysis of the TEST dataset. Precision is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives), Recall is the ratio of true positives to the sum of true positives and false negatives, and Specificity is the ratio of true negatives to the sum of true negatives and false positives.

Table 2. Model classification accuracy for each question

| Question | Accuracy (%) |
|---|---|
| Are the hypotheses for the study included in the Introduction? | 91% |
| Does the Method define the pre-specified primary outcome measure or primary endpoint? | 90% |
| Does the Method include how the sample size was determined? | 100% |
| Does the Method include the eligibility criteria for the participants? | 75% |
| Does the Method include the method used to generate the random allocation sequence? | 75% |
| Does the Method include the type of randomization and details of any restriction (such as blocking and block size)? | 88% |
| Does the Method include the mechanism used to implement the random allocation sequence and any steps taken to conceal the sequence? | 91% |
| Does the Method include who was blinded after assignment to interventions? | 92% |
| Do the Results include the estimated effect size and confidence interval? | 70% |

Accuracy = number of correct answers/number of answers *100

# Discussion

We wanted to determine how accurate an AI-LLM is for measuring reporting guideline compliance in a sample of clinical trial reports. The OpenAI GPT3.5 AI-LLM achieved ~86% accuracy across nine reporting guideline questions. Using an AI-LLM may help journal editors and publishers check reporting guideline adherence without increasing workloads for peer reviewers.

Poor reporting of clinical trials negatively impacts care and is unethical [3,4,28,29]. Several other interventions to improve clinical trial reporting guideline adherence have been examined, but the results have been inconsistent [13]. Typically, these 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 are less common [14]. There are many automated tools that journal publishers and editors use to screen manuscripts for errors and misconduct [30]. 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 substantially increasing the workload on editors and peer reviewers. The accuracy of the AI-LLM in the present study is similar to that reported by Liu and Shah [15] and is equal to, or better than, the accuracy (<80%) of author-submitted CONSORT checklists reported by Blanco et al. (2019) [12]. The less-than-perfect accuracy of the model and the tendency of the current generation of AI-LLMs to "hallucinate" content [31] means that human confirmation of compliance is required. However, academics and scientists have long used imperfect automated tools to assess reporting standards (e.g. [32]). The present 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.

Accuracy across the nine reporting guideline items ranged from 70% to 100%. The causes of the variations in accuracy are not clear. The model was most accurate when answering questions about the blinding of experimenters and participants, the presence of a hypothesis and the definition of the primary outcome. These questions may have had the easiest-to-identify tokens (i.e. keywords or phrases) in the text. For example, the word hypothesis in the Introduction or the phrases "sample size calculation" or "power analysis" in the Method section. Conversely, while the model extracted the relevant text to identify the presence of the participant eligibility criteria, it was less stringent than the human analysis of Schulz et al. and incorrectly answered YES for 25% of analyzed papers (all errors were false positives). LLMs have well-documented limitations with numerical processing [33], so it is perhaps unsurprising that the model was least accurate when trying to confirm the presence of effect sizes and confidence intervals. These errors highlight the limitations of current AI-LLMs for editorial tasks; they are a tool that can assist with, but not replace or supersede, human evaluation of the scientific text. Nonetheless, it is possible that some of the errors seen here simply reflect the impact of this very small dataset on model training. Indeed, as a post hoc experiment, we trained the model on each of the least accurate questions individually, and performance was greatly improved for that question, suggesting model performance could be higher with a larger training dataset.

The primary limitation of this study is the small dataset, which would have impacted the ability to train the model, either by over or underfitting the model parameters and limiting the generalizability of these results to other datasets. It was beyond the scope of this exploratory study to create a large, well-labelled dataset for model tuning. The model was therefore trained using data generated by the model (the correct answers and extracted text from the training data), possibly increasing tendencies to hallucinate and other errors. Nevertheless, acceptable model performance was achieved, and results should improve with larger samples. Other technical limitations imposed by the model choice (e.g. the inability to process entire papers for each question) should be resolved as access to superior AI-LLM models increases. We were also limited in the questions we could ask of the extracted data due to the inability to (at the time of analysis) upload figures to extract data from tables reliably, primarily due to the lack of machine-readable open-access text. For example, some of the 'core' CONSORT questions [24] require analysis of the CONSORT flow figures. As AI technology evolves, many of these issues may be resolved.

## Conclusion

An AI-LLM was 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

## 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

Code notebooks and data are available on the Open Science Framework (relevant links within the text). Copyright issues prevent the sharing of some of 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, Moher D, Dickersin K, Boutron 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: 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: 702–705. doi:10.1093/jnci/djq117
4. Chalmers I. Underreporting research is scientific misconduct. JAMA. 1990;263: 1405–1408. doi:10.1001/jama.1990.03440100121018
5. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. The Lancet. 2014;383: 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: 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, Debono VB, Dillenburg R, Zhang S, 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: 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: 60. doi:10.1186/2046-4053-1-60
12. Blanco D, Biggane AM, Cobo E, Altman D, Bertizzolo L, Boutron I, et al. Are CONSORT checklists submitted by authors adequately reflecting what information is actually reported in published papers? Trials. 2018;19: 80. doi:10.1186/s13063-018-2475-0
13. Speich B, Mann E, Schönenberger CM, Mellor K, Griessbach AN, Dhiman P, 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: 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: e026589. doi:10.1136/bmjopen-2018-026589
15. Liu R, Shah NB. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. arXiv; 2023. doi:10.48550/arXiv.2306.00622
16. 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. doi:10.1136/bmjopen-2021-059347
17. Caldwell AR, Vigotsky AD, Tenan MS, Radel R, Mellor DT, Kreutzer A, et al. Moving Sport and Exercise Science Forward: A Call for the Adoption of More Transparent Research Practices. Sports Med Auckl NZ. 2020;50: 449–459. doi:10.1007/s40279-019-01227-1
18. Twomey R, Yingling V, Warne J, Schneider C, McCrum C, Atkins W, et al. The Nature of Our Literature: A Registered Report on the Positive Result Rate and Reporting Practices in Kinesiology. Commun Kinesiol. 2021;1. doi:10.51224/cik.v1i3.43
19. 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. doi:10.51224/cik.2023.53
20. 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: 220946. doi:10.1098/rsos.220946
21. Hansford HJ, Cashin AG, Wewege MA, Ferraro MC, McAuley JH, Jones MD, et al. Open and transparent sports science research: the role of journals to move the field forward. Knee Surg Sports Traumatol Arthrosc. 2022;30: 3599–3601. doi:10.1007/s00167-022-06893-9
22. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26: 1320–1324. doi:10.1038/s41591-020-1041-y
23. OpenAI Platform. [cited 1 Nov 2023]. Available: https://platform.openai.com
24. Blanco D, Schroter S, Aldcroft A, Moher D, Boutron I, Kirkham JJ, et al. Effect of an editorial intervention to improve the completeness of reporting of randomised trials: a randomised controlled trial. BMJ Open. 2020;10: e036799. doi:10.1136/bmjopen-2020-036799
25. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340: c869. doi:10.1136/bmj.c869
26. Moradi M, Samwald M. Evaluating the Robustness of Neural Language Models to Input Perturbations. arXiv; 2021. doi:10.48550/arXiv.2108.12237
27. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika. 1934;26: 404–413. doi:10.2307/2331986
28. Altman DG. The scandal of poor medical research. BMJ. 1994;308: 283–284. doi:10.1136/bmj.308.6924.283
29. 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
30. Bordewijk EM, Li W, van Eekelen R, Wang R, Showell M, Mol BW, 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
31. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. Crit Care. 2023;27: 180. doi:10.1186/s13054-023-04473-y
32. Straumsheim C. What Is Detected? In: Inside Higher Ed [Internet]. [cited 25 Oct 2023]. Available: https://www.insidehighered.com/news/2015/07/14/turnitin-faces-new-questions-about-efficacy-plagiarism-detection-software
33. Razeghi Y, Logan IV RL, Gardner M, Singh S. Impact of Pretraining Term Frequencies on Few-Shot Reasoning. arXiv; 2022. doi:10.48550/arXiv.2202.07206

# Supplementary Material

Table S1: Description of the TRAIN and TEST datasets

| | **TEST**, N = 198[1] | **TRAIN**, N = 784[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, %) | | |
| Are the hypotheses for the study included in the Introduction? | 23 (12%) | 90 (11%) |
| Do the Results include the estimated effect size and confidence interval? | 23 (12%) | 90 (11%) |
| Does the Method define the pre-specified primary outcome measure or primary endpoint? | 21 (11%) | 87 (11%) |
| Does the Method include how the sample size was determined? | 22 (11%) | 86 (11%) |
| Does the Method include the eligibility criteria for participants? | 28 (14%) | 80 (10%) |
| Does the Method include the mechanism used to implement the random allocation sequence and any steps taken to conceal the sequence? | 22 (11%) | 86 (11%) |
| Does the Method include the method used to generate the random allocation sequence? | 16 (8.1%) | 92 (12%) |
| Does the Method include the type of randomisation and details of any restriction (such as blocking and block size)? | 17 (8.6%) | 91 (12%) |
| Does the Method include who was blinded after assignment to interventions? | 26 (13%) | 82 (10%) |

[1] = n question-text pairs

Table S2: model performance for each question in the TEST dataset

| **Question** | **FN**, N = 13 | **FP**, N = 15 | **TN**, N = 84 | **TP**, N = 86 | Correct (%) |
|---|---|---|---|---|---|
| Does the Method include how the sample size was determined? | 0 | 0 | 9 | 13 | 100 |
| Does the Method include who was blinded after assignment to interventions? | 1 | 1 | 16 | 8 | 92 |
| Are the hypotheses for the study included in the Introduction? | 0 | 2 | 10 | 11 | 91 |
| Does the Method include the mechanism used to implement the random allocation sequence and any steps taken to conceal the sequence? | 1 | 1 | 10 | 10 | 91 |
| Does the Method define the pre-specified primary outcome measure or primary endpoint? | 2 | 0 | 12 | 7 | 90 |
| Does the Method include the type of randomisation and details of any restriction (such as blocking and block size)? | 2 | 0 | 9 | 6 | 88 |
| Does the Method include the eligibility criteria for participants? | 0 | 7 | 0 | 21 | 75 |
| Does the Method include the method used to generate the random allocation sequence? | 2 | 2 | 4 | 8 | 75 |
| Do the Results include the estimated effect size and confidence interval? | 5 | 2 | 14 | 2 | 70 |

FN; False negative, FP; False positive, TN; True negative, TP; True positive

## Sensitivity analysis

The MI-CLAIM guidelines stipulate a "model examination" step for all A.I. studies to understand better how the model performance is affected by the structure of the underlying inputs [22]. AI-LLM's performance may impaired by even small changes to the input text [26], perhaps limiting the generalizability to real-world problems. To test the model's robustness, the text for each text-question pair in the TEST dataset was perturbed using either word deletion of word replacement. For word deletion, a random sample of 10% of the words in the text of each text-question pair was deleted. For word replacement, a random sample of 20% of the words from the text with a synonym were replaced with their synonym. Word deletion model performance: F1 score = 82%, accuracy [95%CI] = 83% [77-88%]. Word replacement model performance: F1 score = 85%, accuracy [95%CI] = 85% [79-90%]. Compared to the unadjusted text, there were some changes to the best and worst-performing questions in each analysis. The two best and worst-performing questions for each perturbation are shown in Table S3.

Table S3. Model classification accuracy for each question

| Perturbation | Question | Accuracy (%) |
|---|---|---|
| Word deletion | Does the Method include who was blinded after assignment to interventions? | 96% |
| | Does the Method include how the sample size was determined? | 91% |
| | Do the Results include the estimated effect size and confidence interval? | 74% |
| | Does the Method include the method used to generate the random allocation sequence? | 62% |
| Word replacement | Does the Method include how the sample size was determined? | 100% |
| | Does the Method include who was blinded after assignment to interventions? | 92% |
| | Does the Method include the method used to generate the random allocation sequence? | 80% |
| | Does the Method include the eligibility criteria for participants? | 75% |