Journal of Medical Systems (2025) 49:80
https://doi.org/10.1007/s10916-025-02212-0

RESEARCH

Large Language Models and the Analyses of Adherence to Reporting Guidelines in Systematic Reviews and Overviews of Reviews (PRISMA 2020 and PRIOR)

Diego A. Forero[1] · Sandra E. Abreu[2] · Blanca E. Tovar[3] · Marilyn H. Oermann[4]

Received: 28 March 2025 / Accepted: 29 May 2025

Abstract

In the context of Evidence-Based Practice (EBP), Systematic Reviews (SRs), Meta-Analyses (MAs) and overview of reviews have become cornerstones for the synthesis of research findings. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 and Preferred Reporting Items for Overviews of Reviews (PRIOR) statements have become major reporting guidelines for SRs/MAs and for overviews of reviews, respectively. In recent years, advances in Generative Artificial Intelligence (genAI) have been proposed as a potential major paradigm shift in scientific research. The main aim of this research was to examine the performance of four LLMs for the analysis of adherence to PRISMA 2020 and PRIOR, in a sample of 20 SRs and 20 overviews of reviews. We tested the free versions of four commonly used LLMs: ChatGPT (GPT-4o), DeepSeek (V3), Gemini (2.0 Flash) and Qwen (2.5 Max). Adherence to PRISMA 2020 and PRIOR was compared with scores defined previously by human experts, using several statistical tests. In our results, all the four LLMs showed a low performance for the analysis of adherence to PRISMA 2020, overestimating the percentage of adherence (from 23 to 30%). For PRIOR, the LLMs presented lower differences in the estimation of adherence (from 6 to 14%) and ChatGPT showed a performance similar to human experts. This is the first report of the performance of four commonly used LLMs for the analysis of adherence to PRISMA 2020 and PRIOR. Future studies of adherence to other reporting guidelines will be helpful in health sciences research.

Clinical Trial Number

Not applicable.

Keywords Reporting guidelines · Systematic reviews · Overview of reviews · Umbrella reviews · Evidence-based practice · Generative artificial intelligence · Meta-research

Introduction

Evidence-Based Practice (EBP) has changed many aspects of the practice, research and teaching in medicine and other health sciences [1]. In this context, Systematic Reviews (SRs) and Meta-Analyses (MAs) have become cornerstones for the synthesis of research findings (for different types of primary studies) [2–4], being among the most highly cited articles [5] and are major inputs for clinical guidelines [6].

More recently, overviews of reviews, also called *umbrella reviews* [7], have emerged as novel types of articles, summarizing the results from multiple SRs and MAs and being useful for the research synthesis of entire topics and fields [8]. Reporting guidelines have been developed for multiple

✉ Diego A. Forero
dforero41@areandina.edu.co

[1] School of Health and Sport Sciences, Fundación Universitaria del Área Andina, Bogotá, Colombia

[2] Psychology Program, Fundación Universitaria del Área Andina, Medellín, Colombia

[3] Nursing Program, School of Health and Sport Sciences, Fundación Universitaria del Área Andina, Bogotá, Colombia

[4] Duke University School of Nursing, Durham, NC, USA

Published online: 12 June 2025

types of studies in the health sciences to ensure complete and transparent reporting [9] and in recent years, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [10] and the Preferred Reporting Items for Overviews of Reviews (PRIOR) statement [11] are recommended for reporting SRs/MAs and overviews of reviews, respectively. Although these statements are robust, they require careful application and critical evaluation, to adapt to new challenges and advances in the health sciences.

In recent years, advances in Generative Artificial Intelligence (genAI) have been proposed as a potential major paradigm shift in scientific research [12]. In this context, Large Language Models (LLMs) have been explored for use in several processes employed in epidemiological research [13, 14], highlighting the need for their analysis of adherence to major reporting guidelines in health sciences research.

The automated analysis of adherence of reporting guidelines will be useful for meta-research works [15], as it will decrease the time needed to carry out these labor and time-intensive activities (given the large number of items to be extracted), allowing for much larger sizes of analyzed studies [13]. Those large studies of adherence might be a major input for efforts aimed at creating or modifying reporting guidelines [9], particularly for other type of studies (such as those for -omics sciences). In addition, the automated analysis of adherence will be helpful for authors of SRs/MAs and overviews of reviews to easily analyze, or improve, the adherence in their manuscripts before submission [16].

The main aim of this work was to examine the performance of four LLMs for the analysis of adherence to PRISMA 2020 and PRIOR, in a sample of SRs and overviews of reviews.

Methods

In the current study, we tested the free versions of four commonly used LLMs: ChatGPT (GPT-4o) [17], DeepSeek (V3) [18], Gemini (2.0 Flash) [19] and Qwen (2.5 Max) [20]. Further details of the LLMs are described in Table S1. These four LLMs were selected as they are broadly used, free and allow the uploading of pdf files. As the chatbots for these LLMs were used, there was no option to modify parameters, such as those minimizing randomness.

An initial pilot phase was carried out to create working prompts for the LLMs and for both PRISMA 2020 and PRIOR. It involved several phases of refinement [21, 22], testing the improved versions of the prompts for each of the LLMs, and using three articles for both reporting guidelines. The selected prompts were chosen after complete answers were provided by the LLMs, simulating a real-word deployment [23] by health sciences researchers. The authors of this study include PhD-level experts in several fields of health sciences research.

The choice of the PRISMA 2020 and PRIOR statements is based on their recognition as highly used reporting guidelines of SRs/MAs and overviews of reviews, respectively. The exploration of the analysis of adherence by LLMs is of potential interest for many researchers around the world, as it would decrease the time needed to carry out these labor and time-intensive activities (given the large number of items to be extracted), allowing for much larger sizes of analyzed studies [24].

The PRISMA 2020 statement was published in 2021 [10] and is a widely used guideline for the reporting of SRs and MAs. PRISMA 2020 contains 27 items (with a total of 42 subitems; compliance with a larger number of items means a higher adherence), for the different sections of the SRs, such as Title, Abstract, Introduction, Methods, Results and Discussion). It also includes a commonly used flow diagram, which indicates the number of identified and included primary studies in the SRs [10].

The PRIOR statement was published in 2022 [11] and provides a guideline for reporting overviews of reviews, particularly related to health interventions. PRIOR contains 27 items (with a total of 46 subitems; compliance with a larger number of items means a higher adherence) and, similar to PRISMA 2020, it involves multiple aspects related to the different sections of the overviews of reviews and a flow diagram [11].

In order to have definitions of adherence to PRISMA 2020 and PRIOR, carried out by human experts, published studies about these were searched and identified. The supplementary files of Qin et al. [25] (for PRISMA 2020) and Lu et al. [26] (for PRIOR) were used to select randomly 20 SRs/MAs [27] and 20 overviews of reviews (Lists of the included studies are available in Supplementary file 1; the sample size of 20 has been previously used in other similar articles [22, 27, 28]). Qin et al. [25] and Lu et al. [26] were focused on SRs and overviews of reviews in the field of acupuncture and the adherence to PRISMA 2020 and PRIOR were carried out by the consensus of several experts in health sciences research. These two datasets were selected as they provided the complete information for each one of the included articles. The *RAND* function of the MS Excel 365 software (Microsoft Corporation, Redmond, WA) was used for the generation of random numbers, to select the studies to be included. The pdf files with the full text for each one of the selected SRs and overviews of reviews were uploaded (in March 2025) to each one of the LLMs, using the optimized prompts (Supplementary file 2), and their responses (including the explanations for their assessments) were retrieved and stored.

For the statistical analysis, several complementary approaches were used [21]. Based on previous studies [29, 30], for the determination of overall adherence for both PRISMA 2020 and PRIOR, each item was defined by each LLM as having adherence (reported), no adherence (not reported) or partial adherence (partially reported) and were counted as 1, 0 or 0.5, respectively, as previously done [26]. The overall adherence percentage was defined as the total sum divided by the total number of subitems (42 for PRISMA 2020; 46 for PRIOR), multiplied by 100.

The Shapiro–Wilk test was used to explore the normality of the studied numerical variables [31]. ANOVA tests [32] were carried out, followed by Tukey´s tests for adjustment [33], to determine statistical differences in the adherence scores determined by the LLMs and the human experts. A Pearson´s r coefficient [34] was calculated to determine the correlation between the responses of the LLMs and the human experts. The Altman-Bland plots and tables [35] were calculated to analyze in detail the agreement between each of the LLMs and the human experts (the mean differences and the confidence interval: +/- 1.96 SDs of the differences). Additionally, as previously described [36], the accuracy for each of the LLMs was also estimated. A p value < 0.05 was defined as statistically significant.

In addition, the calculation of adherence percentages to sections of PRISMA 2020 and PRIOR (such as for Methods and Results, among others) were also carried out and some examples of texts were extracted to visualize details of the responses generated by the LLMs. We did not calculate parameters such as specificity and sensitivity as they are used for topics with dichotomic variables [37]. The JASP program (version 0.18.3.0) [38] was employed for the statistical analyses. For the reporting of this study, we took into consideration key aspects of the MI-CLEAR-LLM guidelines [39].

Results

For the analysis of the adherence to PRISMA 2020, our analysis identified that there was a higher percentage of adherence in the responses defined by all the four LLMs, in comparison to human subjects: this is shown in box plots (Fig. 1A) and in detailed plots for each one of the included SRs (Fig. 1B). A statistical analysis showed that there were significant differences between each of the LLMs compared to the human experts and that there were no large correlations to the scores defined by human experts (Table 1). An analysis of agreement, using the Bland-Altman plots and tables, showed that the differences between the LLMs and the human experts were large: on average, from 23.1 to 29.7% (Table 1; Fig. 2). The accuracy for each one of the LLMs, for PRISMA 2020, complemented the previously described analyses (Table 3). Overall, these results for PRISMA 2020 indicate a poor performance for all the four tested LLMs.

An analysis of the adherence to sections of the PRISMA 2020 also found that the differences were large for the multiple sections of items (Table S2) and examples of text showed the differences in the texts of the responses of the LLMs (Table S4).

Regarding the PRIOR statement, our analysis showed that, in comparison to the results from PRISMA 2020, there was better concordance for the adherence defined by the LLMs, in comparison to human experts: this is shown in box plots (Fig. 3A) and in detailed plots for each one of the included overviews of reviews (Fig. 3B). A statistical analysis showed that there were no significant differences between the responses generated by ChatGPT and the scores defined by the human experts and that there was also a significant correlation for ChatGPT and human experts (Table 2). The analysis of agreement, using the Bland-Altman plots and tables, showed that the differences between the ChatGPT and the human experts were smaller: on average, 6.1 (Table 2; Fig. 4). The accuracy for each one of the LLMs, for PRIOR, complemented the previously described analyses (Table 3). Overall, these results for PRIOR indicate a poor performance for three of the four tested LLMs.

Finally, the analysis of the adherence to sections of the PRIOR found that the differences, between ChatGPT and the human experts, were smaller for the Results and Discussion sections (Table S3) and examples of text also showed the differences in the texts of the responses of the LLMs (Table S5).

Discussion

This is the first report of the performance of four commonly used LLMs for the automatic identification of adherence to PRISMA 2020 and PRIOR, in a sample of SRs/MAs and overviews of reviews. Among these two statements, PRISMA 2020 [10] is one of the most used and highly cited reporting guidelines in the health sciences, with more than 86,000 citations to date.

In our results, all the four LLMs showed a low performance for the analysis of adherence to PRISMA 2020, in comparison to human experts. In general, the LLMs overestimated the percentage of adherence (in addition to exhibiting low accuracy), and it was more evident for SRs with a low percentage. In contrast, for PRIOR, the LLMs presented lower differences in the estimation of adherence and

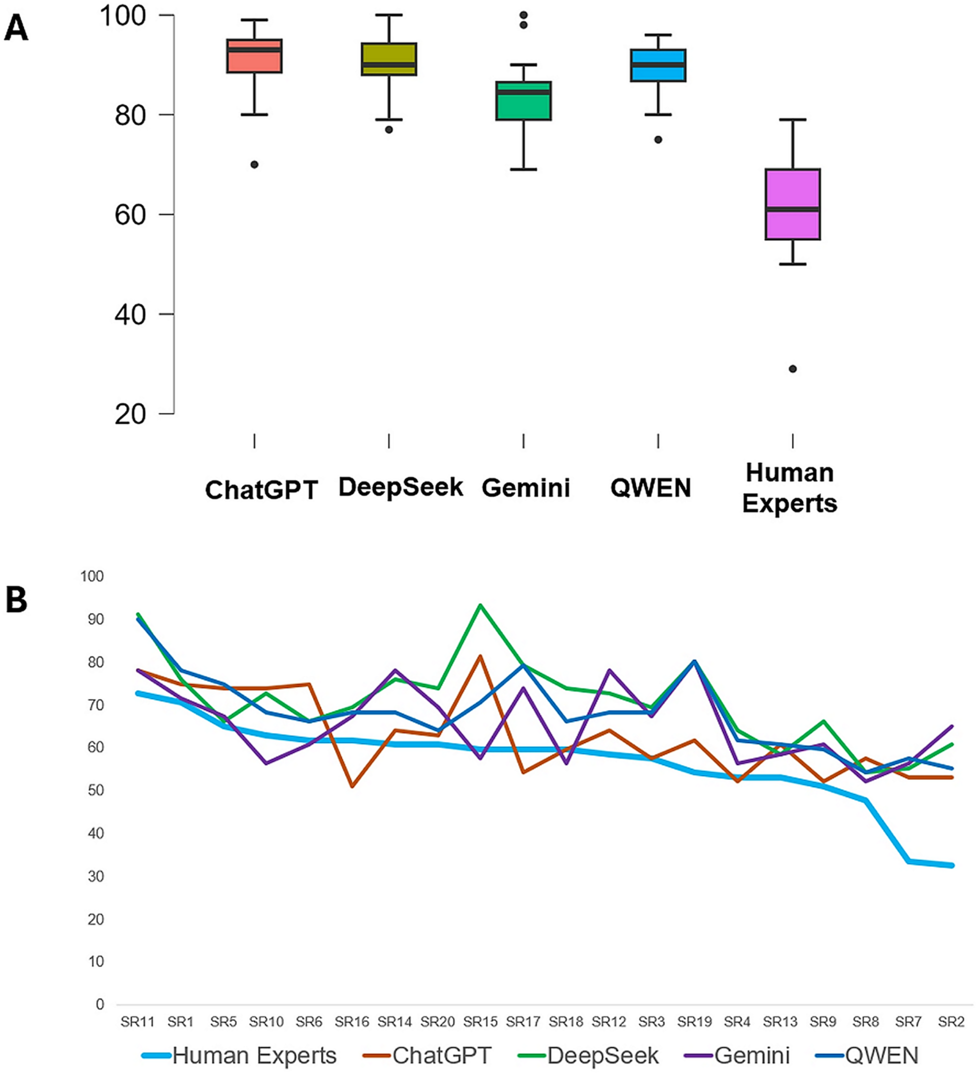

Fig. 1 Analysis of the performance of four LLMs for the estimation of adherence to PRISMA 2020, in a sample of SRs, in comparison to human experts. **A**. A box plot for the overall adherence percentages to PRISMA 2020 estimated by the four LLMs and by human experts. **B**. A detailed plot of the adherence to PRISMA 2020, estimated by each of the four LLMs and the human experts, for each one of the SRs included

Table 1 Analysis of the performance of four LLMs for the analysis of adherence to PRISMA 2020, in a sample of SRs, in comparison to human experts

Measurement	Adherence (a)	*p* value (b)	Correlation (c)	Bland-Altman (d)
Human Experts	61.2 (11.2)	Taken as reference	Taken as reference	Taken as reference
ChatGPT	90.0 (7.0)	< 0.001	0.60 (0.005)	29.7 (12.2–47.2)
DeepSeek	90.6 (6.8)	< 0.001	0.43 (0.06)	29.4 (9.2–49.6)
Gemini	84.3 (8.5)	< 0.001	0.36 (0.12)	23.2 (0.9–45.4)
Qwen	89.2 (5.6)	< 0.001	0.39 (0.09)	28.1 (7.7–48.4)

(a) Presented as mean (SD) of overall adherence percentages. (b) Results from an ANOVA test, with Tukey´s adjustment, in comparison with human experts. (c) Pearson's correlation (*p* value in parenthesis), in comparison with human experts. (d) Results from a Bland-Altman analysis: Mean differences (range of difference in parenthesis), in comparison with human experts

ChatGPT showed a performance similar to human experts. Our current findings are consistent with the results of a recent scoping review highlighting that general-purpose LLMs are not ready for use in research synthesis [24].

A current major challenge, from the perspective of health sciences research, is the presence of errors, commonly defined as hallucinations or confabulations, in the outputs of LLMs [40, 41]. Transparence about the use of genAI models in health sciences research [42] is a major challenge when there is a lack of available information about the articles used for the training of LLMs [43, 44]. In addition, it is possible that some advanced uses of LLMs, such as the evaluation of adherence to PRISMA 2020, require a better performance of the LLMs in functional linguistic competence [45].

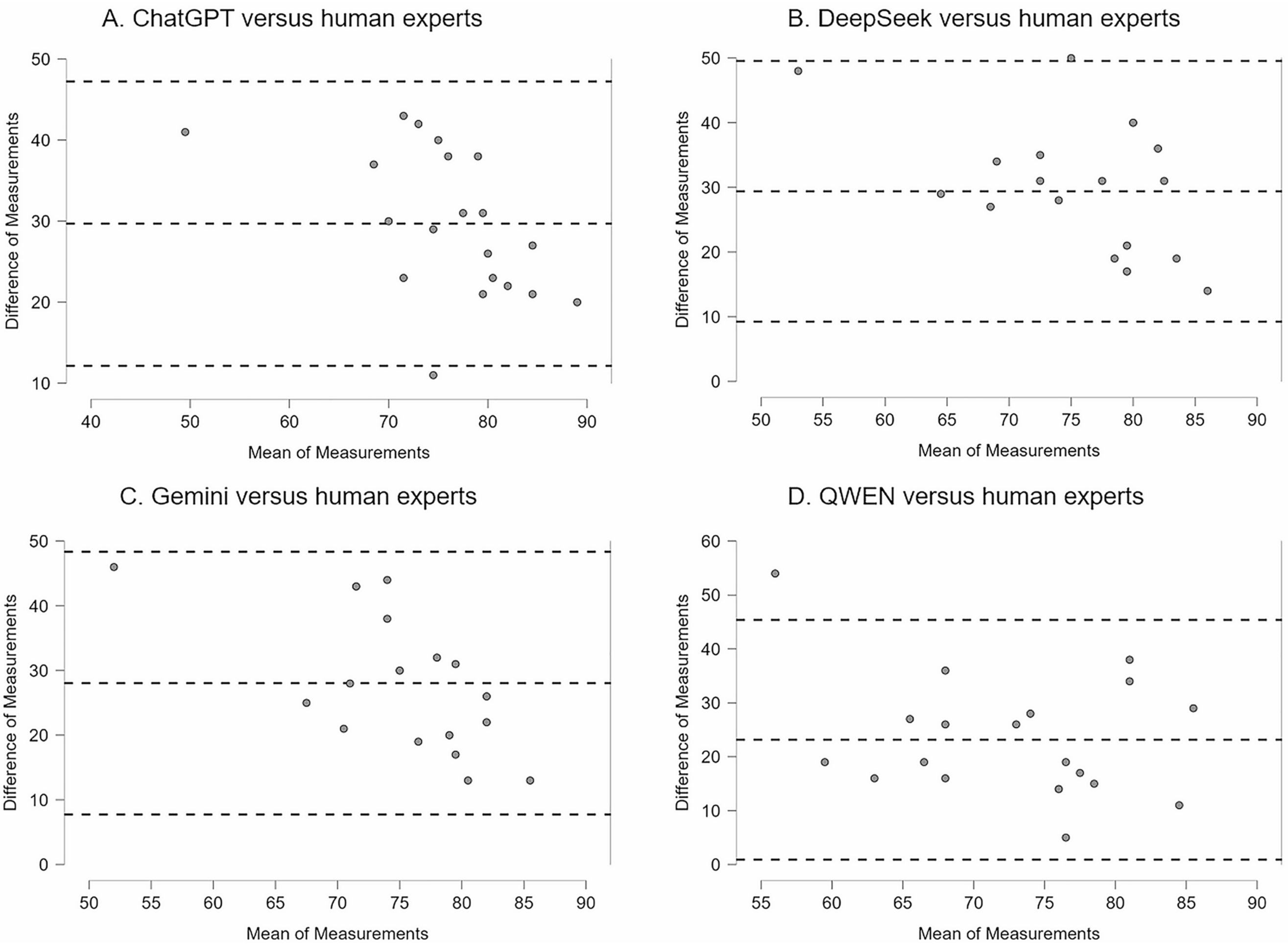

Fig. 2 Bland-Altman plots for the concordance of determination of adherence to the PRISMA 2020 statement (overall adherence percentages), for four LLMs in comparison to human experts

Some recent articles have explored the results from LLMs for related tools. Woelfle et al. explored five versions of three LLMs (including Claude and ChatGPT) for the analysis of PRISMA 2009 (an older version of PRISMA; this study was focused in its current version: PRISMA 2020) in a sample of SRs; they found that the accuracy for the LLMs range from 63 to 70% [13]. Roberts et al. explored ChatGPT 3 for the analysis of the CONSORT-A guidelines and found small differences with scores defined by humans [21]. Other recent studies have focused on LLMs and analysis of risk of bias in SRs/MAs and in primary studies [46–48].

Limitations of the current study include the homogenous nature and relatively small sample size of the included studies [22, 28]. Future studies need the use of more heterogeneous and larger samples of studies and the testing of the potential effect of more complex prompts [49].

Future studies of adherence to other reporting guidelines, including additional LLMs, will be helpful in health sciences research. Additionally, future studies of LLMs designed, or fine-tuned [50], for advanced analyses of epidemiological studies and data [24, 51] are needed. Finally, as the majority of research on LLMs and health sciences research has been carried out in the Global North [52], there is a need for further studies in these topics done in the Global South [43].

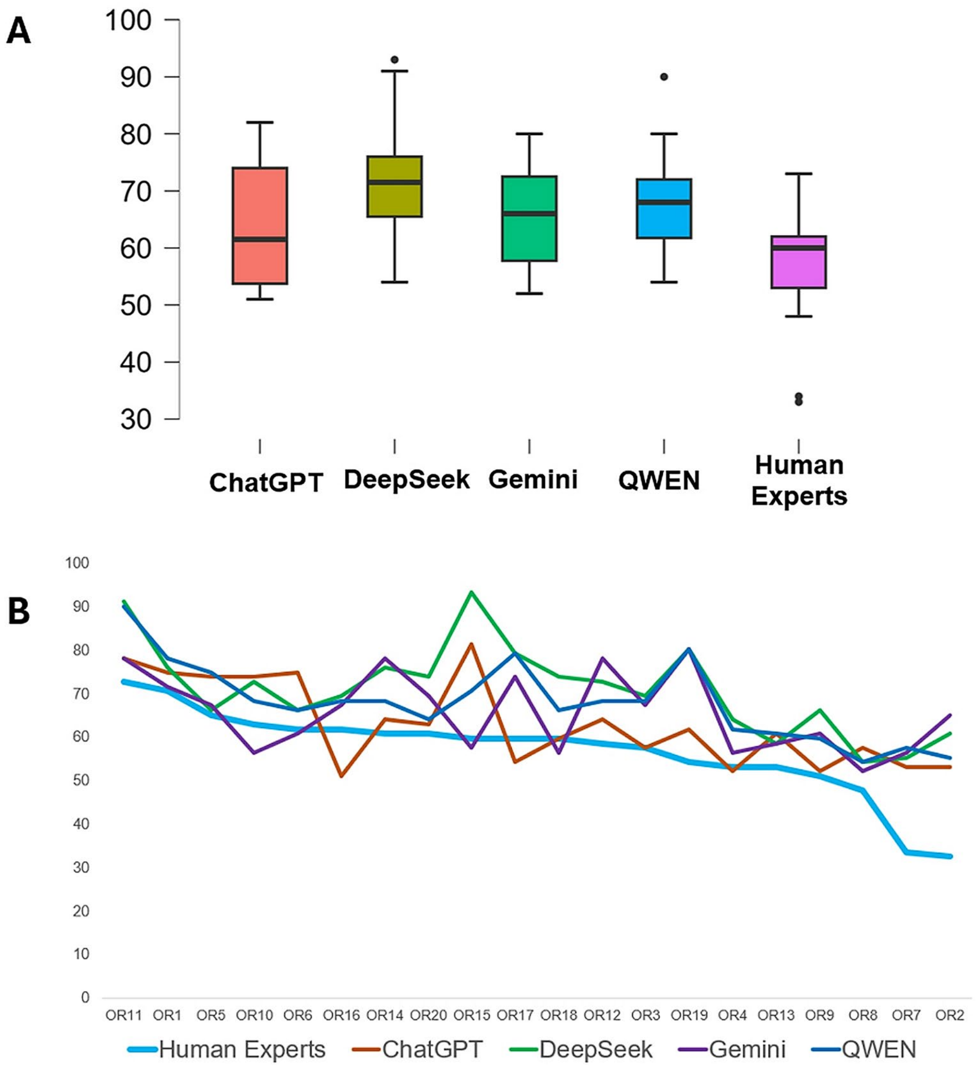

Fig. 3 Analysis of the performance of four LLMs for the estimation of adherence to PRIOR, in a sample of overviews of reviews, in comparison to human experts. **A**. A box plot for the overall adherence percentages to PRIOR estimated by the four LLMs and by human experts. **B**. A detailed plot of the adherence to PRISMA 2020, estimated by each of the four LLMs and the human experts, for each one of the overviews of reviews included

Table 2 Analysis of the performance of four LLMs for the analysis of adherence to PRIOR, in a sample of overviews of reviews, in comparison to human experts

Measurement	Adherence (a)	*p* value (b)	Correlation (c)	Bland-Altman (d)
Human Experts	57.1 (10.1)	Taken as reference	Taken as reference	Taken as reference
ChatGPT	63.2 (9.9)	0.27	0.65 (0.002)	6.1 (-10.3-22.5)
DeepSeek	71.0 (10.3)	<0.001	0.66 (0.001)	14.0 (-2.5-30.4)
Gemini	65.8 (8.7)	0.04	0.41 (0.07)	8.7 (-11.4-28.8)
Qwen	68.0 (9.0)	0.005	0.75 (<0.001)	10.9 (-2.5-24.3)

(a) Presented as mean (SD) of overall adherence percentages. (b) Results from an ANOVA test, with Tukey´s adjustment, in comparison with human experts. (c) Pearson’s correlation (*p* value in parenthesis), in comparison with human experts. (d) Results from a Bland-Altman analysis: Mean differences (range of difference in parenthesis), in comparison with human experts

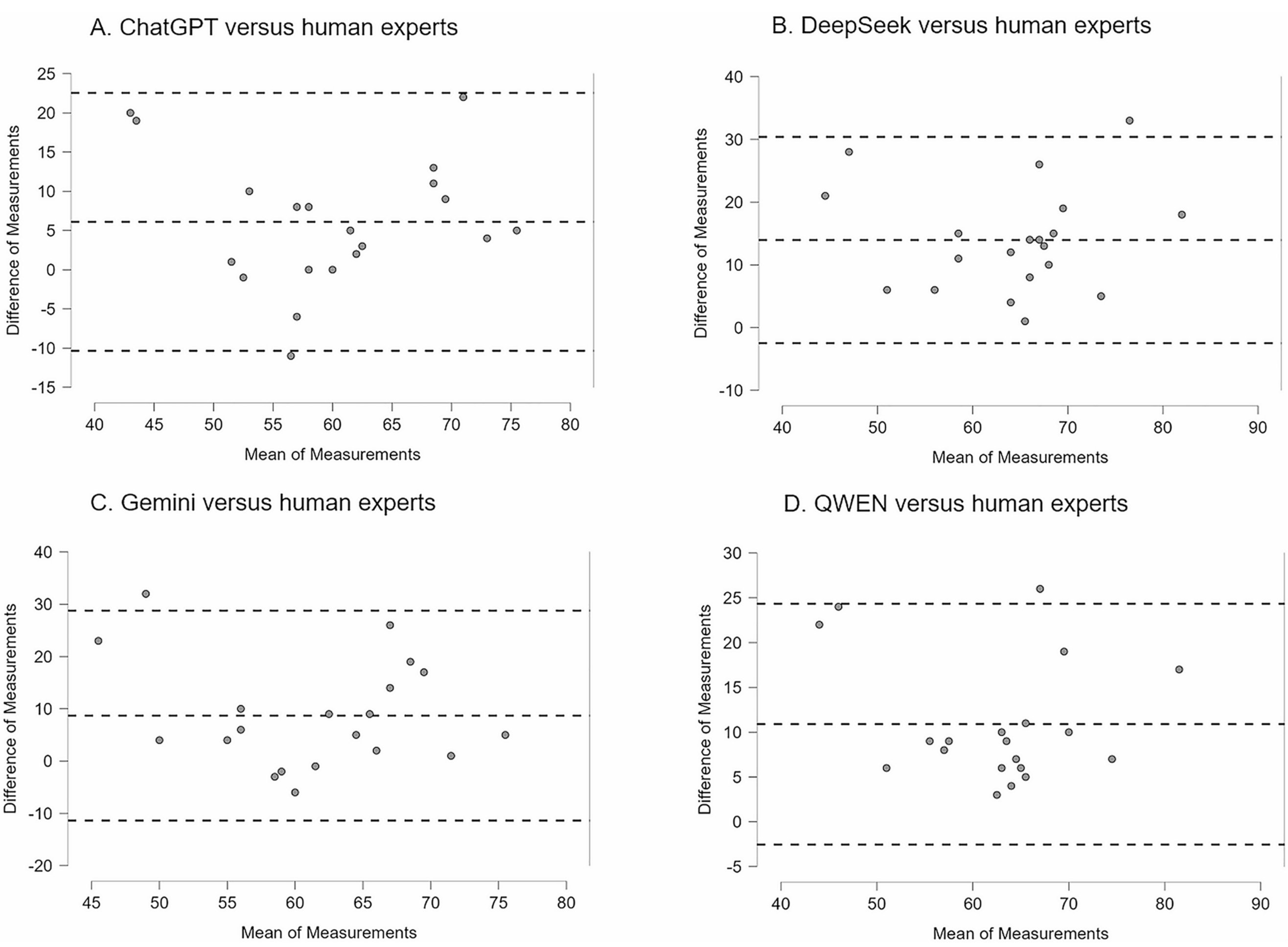

Fig. 4 Bland-Altman plots for the concordance of determination of adherence to the PRIOR statement (overall adherence percentages), for four LLMs in comparison to human experts

Table 3 Mean accuracy of four LLMs for the automatic identification of adherence to the PRISMA 2020 (upper part) and PRIOR (lower part) statements

PRISMA 2020				
LLM	Gemini	DeepSeek	ChatGPT	QWEN
Mean (SD)	70.4 (9.2)	67.3 (9.0)	61.8 (9.3)	60.5 (9.3)
PRIOR				
LLM	DeepSeek	Gemini	QWEN	ChatGPT
Mean (SD)	74.0 (7.2)	72.2 (7.0)	70.8 (7.0)	67.9 (8.2)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10916-025-02212-0.

Acknowledgements DAF has been previously supported by research grants from Minciencias and Areandina.

Author Contributions DAF and SEA extracted the information. DAF and BET analyzed the data. DAF wrote a first version of the manuscript. SEA, BET and MHO contributed to the final manuscript. All authors reviewed the manuscript.

Funding Open Access funding provided by Colombia Consortium
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Data is provided within the manuscript or supplementary information files.

Declarations

Research Involving Human Participants and/or Animals This manuscript describes only computational analyses and did not describe research involving human participants or animal models.

Declaration of generative AI in Scientific Writing No generative AI tools were used for scientific writing.

Competing Interests The authors declare no competing interests.

References

1. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017;390(10092):415–23.
2. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175–82.
3. Forero DA, Lopez-Leon S, Gonzalez-Giraldo Y, Bagos PG. Ten simple rules for carrying out and writing meta-analyses. PLoS Comput Biol. 2019;15(5):e1006922.
4. Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. Eval Health Prof. 2002;25(1):12–37.
5. Patsopoulos NA, Analatos AA, Ioannidis JP. Relative citation impact of various study designs in the health sciences. JAMA. 2005;293(19):2362–6.
6. Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, Prasad K, et al. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. JAMA. 2014;312(2):171–9.
7. Fusar-Poli P, Radua J. Ten simple rules for conducting umbrella reviews. Evid Based Ment Health. 2018;21(3):95–100.
8. Bougioukas KI, Pamporis K, Vounzoulaki E, Karagiannis T, Haidich AB. Types and associated methodologies of overviews of reviews in health care: a methodological study with published examples. J Clin Epidemiol. 2023;153:13–25.
9. Altman DG, Simera I. A history of the evolution of guidelines for reporting medical research: the long road to the EQUATOR Network. J R Soc Med. 2016;109(2):67–77.
10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
11. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of health-care interventions: development of the PRIOR statement. BMJ. 2022;378:e070849.
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40.
13. Woelfle T, Hirt J, Janiaud P, Kappos L, Ioannidis JPA, Hemkens LG. Benchmarking Human-AI collaboration for common evidence appraisal tools. J Clin Epidemiol. 2024;175:111533.
14. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;15(4):616–26.
15. Ioannidis JPA. Meta-research: Why research on research matters. PLoS Biol. 2018;16(3):e2005468.
16. Collins GS. Innovative solutions are needed to overcome implementation barriers to using reporting guidelines. BMJ. 2025;389:r718.
17. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv (preprint). 2023.
18. Liu A, Feng B, Xue B, Wang B, Wu B, Lu C, et al. Deepseek-v3 technical report. arXiv (preprint). 2024.
19. Gemini Team, Anil R, Borgeaud S, Alayrac J-B, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. arXiv (preprint). 2023.
20. Yang A, Yang B, Zhang B, Hui B, Zheng B, Yu B, et al. Qwen2. 5 technical report. arXiv (preprint). 2024.
21. Roberts RH, Ali SR, Hutchings HA, Dobbs TD, Whitaker IS. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. BMJ Health Care Inform. 2023;30(1).
22. Wrightson JG, Blazey P, Moher D, Khan KM, Ardern CL. GPT for RCTs? Using AI to determine adherence to clinical trial reporting guidelines. BMJ Open. 2025;15(3):e088735.
23. Srinivasan A, Berkowitz J, Kivelson S, Friedrich N, Tatonetti N. Evaluating the Reporting Quality of 21,041 Randomized Controlled Trial Articles. Preprint, medRxiv. 2025:2025.03. 06.25323528.
24. Lieberum JL, Tows M, Metzendorf MI, Heilmeyer F, Siemens W, Haverkamp C, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. J Clin Epidemiol. 2025;181:111746.
25. Qin C, Ma H, Mandizadza OO, Xu X, Ji C. Reporting quality of meta-analyses in acupuncture: Investigating adherence to the PRISMA statement. Medicine (Baltimore). 2024;103(39):e39933.
26. Lu T, Liu B, Lu C, Du Z, Yang K, Ge L. Reporting quality of acupuncture overviews: A methodological investigation based on the PRIOR statement. Complement Ther Med. 2024;82:103034.
27. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.
28. Alharbi F, Asiri S. Automated Assessment of Reporting Completeness in Orthodontic Research Using LLMs: An Observational Study. Applied Sciences. 2024;14(22):10323.
29. Innocenti T, Feller D, Giagio S, Salvioli S, Minnucci S, Brindisino F, et al. Adherence to the PRISMA statement and its association with risk of bias in systematic reviews published in rehabilitation journals: A meta-research study. Braz J Phys Ther. 2022;26(5):100450.
30. Panic N, Leoncini E, de Belvis G, Ricciardi W, Boccia S. Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. PLoS One. 2013;8(12):e83138.
31. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth. 2019;22(1):67–72.
32. Bewick V, Cheek L, Ball J. Statistics review 9: one-way analysis of variance. Crit Care. 2004;8(2):130–6.
33. McHugh ML. Multiple comparison analysis testing in ANOVA. Biochem Med (Zagreb). 2011;21(3):203–9.
34. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. Crit Care. 2003;7(6):451–9.
35. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141–51.
36. Li Z, Luo X, Yang Z, Zhang H, Wang B, Ge L, et al. RAPID: Reliable and efficient Automatic generation of submission rePorting checklists with Large language moDels. Preprint, bioRxiv. 2025:2025.02. 13.638015.
37. Gartlehner G, Kahwati L, Nussbaumer-Streit B, Crotty K, Hilscher R, Kugley S, et al. From promise to practice: challenges and pitfalls in the evaluation of large language models for data extraction in evidence synthesis. BMJ Evid Based Med. 2024.
38. Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen J, et al. JASP: Graphical statistical software for common statistical designs. Journal of Statistical Software. 2019;88:1–17.
39. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM). Korean J Radiol. 2024;25(10):865–8.

40. Smith AL, Greaves F, Panch T. Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. PLOS Digit Health. 2023;2(11):e0000388.
41. Kim Y, Jeong H, Chen S, Li SS, Lu M, Alhamoud K, et al. Medical Hallucination in Foundation Models and Their Impact on Healthcare. medRxiv (Preprint). 2025:2025.02. 28.25323115.
42. Miao BY, Chen IY, Williams CYK, Davidson J, Garcia-Agundez A, Sun S, et al. The MI-CLAIM-GEN checklist for generative artificial intelligence in health. Nat Med. 2025.
43. Lobentanzer S, Feng S, Bruderer N, Maier A, BioChatter C, Wang C, et al. A platform for the biomedical application of large language models. Nat Biotechnol. 2025;43(2):166–9.
44. Nguyen T. ChatGPT in Medical Education: A Precursor for Automation Bias? JMIR Med Educ. 2024;10:e50174.
45. Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, Fedorenko E. Dissociating language and thought in large language models. Trends Cogn Sci. 2024;28(6):517–40.
46. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open. 2024;7(5):e2412687.
47. Kuitunen I, Ponkilainen VT, Liukkonen R, Nyrhi L, Pakarinen O, Vaajala M, et al. Evaluating the Performance of ChatGPT-4o in Risk of Bias Assessments. J Evid Based Med. 2024;17(4):700–2.
48. Barsby J, Hume S, Lemmey HA, Cutteridge J, Lee R, Bera KD. Pilot study on large language models for risk-of-bias assessments in systematic reviews: A(I) new type of bias? BMJ Evid Based Med. 2025;30(1):71–4.
49. Lin Z. How to write effective prompts for large language models. Nat Hum Behav. 2024;8(4):611–5.
50. Anisuzzaman D, Malins JG, Friedman PA, Attia ZI. Fine-Tuning Large Language Models for Specialized Use Cases. Mayo Clin Proc Digit Health. 2025;3(1):100184.
51. Cox LA, Jr. An AI assistant to help review and improve causal reasoning in epidemiological documents. Glob Epidemiol. 2024;7:100130.
52. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: A scoping review. iScience. 2024;27(5):109713.

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