European Spine Journal
https://doi.org/10.1007/s00586-025-09166-4

RESEARCH

Evaluating ChatGPT-4.0's accuracy and potential in idiopathic scoliosis conservative treatment: a preliminary study on clarity, validity, and expert perceptions

Francesco Negrini[1,2] · Calogero Malfitano[3,4] · Giorgio Ferriero[1,2] · Giovanni Morone[5,6] · Alberto Negrini[7] · Fabio Zaina[7] · Irene Ferrario[7] · Carlotte Kiekens[8] · Stefano Negrini[8,9] · Jacopo Vitale[10,11]

Received: 14 April 2025 / Revised: 28 June 2025 / Accepted: 14 July 2025

Abstract

Purpose This study aimed to evaluate the scientific accuracy, content validity, and clarity of ChatGPT-4.0's responses on conservative management of idiopathic scoliosis. The research explored whether the model could effectively support patient education in an area where non-surgical treatment information is crucial.

Methods Fourteen frequently asked questions (FAQs) regarding conservative scoliosis treatment were identified using a systematic, multi-step approach that combined web-based inquiry and expert input. Each question was submitted individually to ChatGPT-4.0 on December 6, 2024, using a standardized patient prompt ("I'm a scoliosis patient. Limit your answer to 150 words"). The generated responses were evaluated by a panel of 37 experts from a specialized spinal deformity center via an online survey using a 6-point Likert scale. Content validity was assessed using the Content Validity Ratio (CVR) and Content Validity Index (CVI), and inter-rater reliability was calculated with Fleiss' kappa. Experts also provided categorical feedback on reasons for any rating discrepancies.

Results Eleven out of 14 responses met the CVR threshold (≥0.38), yielding an overall CVI of 0.68. Three responses - addressing "What is scoliosis?", "Can exercises or physical therapy cure scoliosis?", "What is the best sport for scoliosis?"- showed lower validity (CVR scores: 0.37, 0.37, and −0.58, respectively), primarily due to factual inaccuracies and insufficient detail. Clarity received the highest ratings (median = 6), while comprehensiveness, professionalism, and response length each had a median score of 5. Inter-rater reliability was slight (Fleiss' kappa = 0.10).

Conclusion ChatGPT-4.0 generally provides clear and accessible information on conservative idiopathic scoliosis management, supporting its potential as a patient education tool. Nonetheless, variability in response accuracy and expert evaluation underscores the need for further refinement and expert supervision before wider clinical application.

Keywords Scoliosis · Rehabilitation · Artificial intelligence · Natural Language processing · Patient education as topic

Introduction

The integration of Artificial Intelligence (AI) in medicine has been discussed for some decades, but only in the last five years there has been a surge in clinical studies and investments in the field [1]. A recent overview of reviews has highlighted how AI has been proposed for all phases of clinical medicine, from diagnosis to follow-up, including clinical decision-making and prognosis estimation [2]. Oncology and radiology are the most represented medical domains [2], rehabilitation, and orthopaedic surgery remaining underrepresented. A breakthrough in AI has been the introduction of Large Language Models (LLMs) such as ChatGPT, Gemini, or Deepseek, tools with high potential that remain largely unexplored [3, 4]. One area where LLMs can provide a significant contribution to clinical medicine is patient education [3]. Specifically, LLMs have shown great potential in answering clinical questions that patients may have, and when properly optimized, they appear to be valuable clinical support. Some studies have analyzed LLMs in

Extended author information available on the last page of the article

Published online: 21 July 2025

various areas of clinical medicine, such as orthopedic and spinal surgery [5–7], hepatology [8], radiology [9], and dental medicine [10], mostly with positive outcomes [8]. For the rehabilitative treatment of idiopathic scoliosis (IS) the application of AI appears to be an extremely promising technology. This treatment requires high patient adherence, which is largely influenced by a thorough understanding of the condition and effective patient education [11]. Most IS patients are adolescents, a demographic more exposed to digital technologies and AI, who can quickly adapt to new technological tools [12]. A recent study conducted by surgeons on different LLMs found that ChatGPT-4.0 appears to be the most reliable among the three tested LLMs (ChatGPT-4.0, ChatGPT-3.5, and Bard) in providing information about adolescent IS, particularly in the surgical field [13]. However, ChatGPT has not yet been tested in providing information on rehabilitative treatment. To our knowledge, this is the first study to assess the scientific accuracy and communicative effectiveness of a LLM in answering frequently asked questions about the conservative treatment of scoliosis. Therefore, we aimed to assess whether ChatGPT-4.0 - the most accurate and professional-sounding tool, in the context of surgical management [13] - provides evidence-based, appropriate, and comprehensive answers to common questions about scoliosis conservative treatment. We hypothesize that the majority of ChatGPT's responses would meet established thresholds for content validity and be rated positively across multiple quality dimensions. It is also expected that lower-scoring responses would be associated with specific issues such as factual inaccuracies or insufficient detail.

Table 1 The 14 faqs on non-surgical (conservative) management of scoliosis

FAQs on scoliosis management	
Q1	What is scoliosis?
Q2	What causes scoliosis?
Q3	Can scoliosis get worse over time?
Q4	Can scoliosis lead to disability?
Q5	Can scoliosis be cured?
Q6	Do osteopathy, chiropractic, and manual therapy cure scoliosis?
Q7	Should I use insoles or a heel lift to correct scoliosis?
Q8	Does wearing dental braces affect scoliosis?
Q9	Could my child's heavy school backpack make their scoliosis worse?
Q10	Can exercises or physical therapy cure scoliosis?
Q11	How effective is bracing in stopping or correcting scoliosis progression?
Q12	How many hours per day do I need to wear the brace, and for how many years?
Q13	When is surgery needed for scoliosis?
Q14	What is the best sport for scoliosis?

Materials and methods

Study design

This study was conducted between November and December 2024 and followed a structured two-phase design. The first phase involved the identification and formulation of 14 frequently asked questions (FAQs) related to the conservative treatment of IS. The second phase involved the evaluation of ChatGPT-4.0's responses to these questions by a panel of experts using an online survey. Ethical approval was not required for this study, as confirmed by internal institutional guidance, since no patient data were collected, and all participants were professionals voluntarily assessing AI-generated content. The reporting guidelines for the early-stage clinical evaluation of decision support systems driven by AI (DECIDE-AI) were followed [14]. The study was conducted entirely in English.

Identification of frequently asked questions (FAQs)

The members of the research team with clinical experience with scoliosis patients (FN, FZ, IF, SN– 21.5 ± 10.4 years of clinical experience) developed a list of 14 FAQs concerning the conservative management of IS. These questions were selected using multi-step approach: (1) identification of relevant FAQs on rehabilitation for IS by conducting a search through the first 20 pages of Google results using the query: ('frequently asked questions' OR 'FAQ') AND ('idiopathic scoliosis' OR 'scoliosis') AND ('conservative treatment' OR 'non-surgical treatment' OR 'brace therapy' OR 'physiotherapy'), (2) a list of potential FAQs generated by ChatGPT-4.0 when prompted to produce common questions about scoliosis treatment, (3) a synthesis of the information gathered on point 1 and 2 by ChatGPT-4.0, and (4) a selection of the most relevant questions found in the report generated at step 3 by the clinical team. The experts selected 14 questions able to cover all the main areas typically addressed by patient FAQs (diagnosis, prognosis, treatment, and lifestyle). After initial generation, the list was reviewed and refined collaboratively by the investigators to ensure clarity, clinical relevance, and diversity of topics across diagnosis, treatment, lifestyle, and prognosis (Table 1).

ChatGPT-4.0

Each of the 14 FAQs was submitted individually to ChatGPT-4.0 via the official online platform on the same day (December 6, 2024). The responses were generated using a standardized prompt to simulate a patient perspective and to ensure uniform response length and tone: *"I'm a scoliosis patient. Limit your answer to 150 words."* The complete set

of responses was compiled into a Word document and formatted for survey distribution. No additional edits or modifications were made to the LLM responses before evaluation.

Expert panel

A total of 37 professionals affiliated with a tertiary clinic specialised in spinal deformity were recruited for the evaluation phase. The panel consisted of a multidisciplinary group with expertise in physical therapy, rehabilitation, orthopaedic surgery, psychology and clinical research related to scoliosis. All participants had experience in scoliosis management and were familiar with evaluating clinical information and patient education materials.

Online survey and evaluation procedure

The expert panel received a link to an online Google Forms survey that included the full set of ChatGPT-generated responses. Each answer was evaluated independently and anonymously using a 6-point Likert scale to assess perceived appropriateness of the response. The scale ranged from 1 ("Strongly disagree") to 6 ("Strongly agree"), and scores of 4 to 6 were classified as "appropriate," while scores from 1 to 3 were classified as "non-appropriate." Experts were also asked to indicate the reason for any disagreement with the response using predefined categories: (1) off-topic/not pertinent; (2) clear mistakes in the answer; (3) Too much information/not all necessary; (4) too little information/not enough for an exhaustive answer; (5) Language issues/not suitable for patients; (6) Other reasons (with space for comments). Additionally, experts were asked to assess each response across four secondary dimensions: (1) comprehensiveness; (2) clarity, (3) professionalism; (4) Length; all rated on the same 6-point Likert scale. In the final section of the survey, experts were asked general questions about their perceptions of ChatGPT-4.0's usefulness in patient education, potential risks, and whether the responses met their expectations regarding conservative scoliosis treatment.

Table 2 Characteristics of the expert research team

Variable	Descriptive statistics	Value
Female sex	*(n, %)*	*n* = 16, 55.2%
Age*(years)*	Mean ± SD	41.2 ± 10.8
	Range (min– max)	24.0–65.0
	95% CI (lower– upper)	37.0–44.9
Years of experience*(years)*	Mean ± SD	11.9 ± 9.6
	Range (min– max)	3.0–40.0
	95% CI (lower– upper)	8.4–15.4
Main profession(s)*(n, %)**	Researcher/scientist	2, 6.9%
	Physiotherapist	19, 65.5%
	Professor (academia)	2, 6.9%
	Massage therapist	2, 6.9%
	Medical doctor/physician	7, 24.1%

Values are presented as means ± standard deviations, range (minimum-maximum), and 95% confidence interval. *: more than one option was possible

Statistical analysis

Ordinal and non-normally distributed data are reported as median and interquartile range (IQR), normally distributed continuous data are reported as mean ± standard deviation (SD), and categorical variables are reported as percentages. For the inter-rater reliability, the Fleiss' kappa was calculated and interpreted as none (0–0.19), minimal (0.20–0.39), weak (0.40–0.59), moderate (0.60–0.79), strong (0.80–0.90), and almost perfect (> 0.91) [15]. The ChatGPT-4.0's content validity was assessed using the Content Validity Ratio (CVR). The CVR is used to quantify the degree of agreement among experts on the appropriateness of the answers by ChatGPT-4.0 and it is calculated as:

$$CVR = \frac{\left(ne - \left(\frac{N}{2}\right)\right)}{\left(\frac{N}{2}\right)}$$

where n_e is the number of experts indicating an item "appropriate," and N is the total number of experts. Therefore, considering the sample size of experts ($n = 29$) and the recognized Lawshe's CVR critical values [16], a CVR minimum value of 0.38 was expected for each answer to be considered valid (corresponding to a 69% agreement). Specifically, CVR was calculated by converting the 6-point Likert scale into a binary decision as follows: scores from 1 to 3 were classified as "non-appropriate" and scores from 4 to 6 as "appropriate". The Content Validity Index (CVI) was calculated as the average of the CVRs for the 10 answers. Further, a Chi-square test was used to assess the differences in the reasons for inappropriate responses among the 14 answers provided by ChatGPT-4.0. Statistical analyses were performed using GraphPad Prism version 9.5.1 for Windows (GraphPad Software, San Diego, CA, USA) and the open-source Pingouin and Statsmodels packages in Python 3.9. The level of significance was set at $p < 0.05$.

Results

Expert group and inter-rater reliability

The experts' answer rate was 78% (29/37) (Table 2). The Fleiss' Kappa value for inter-rater reliability was 0.10, indicating slight agreement among experts.

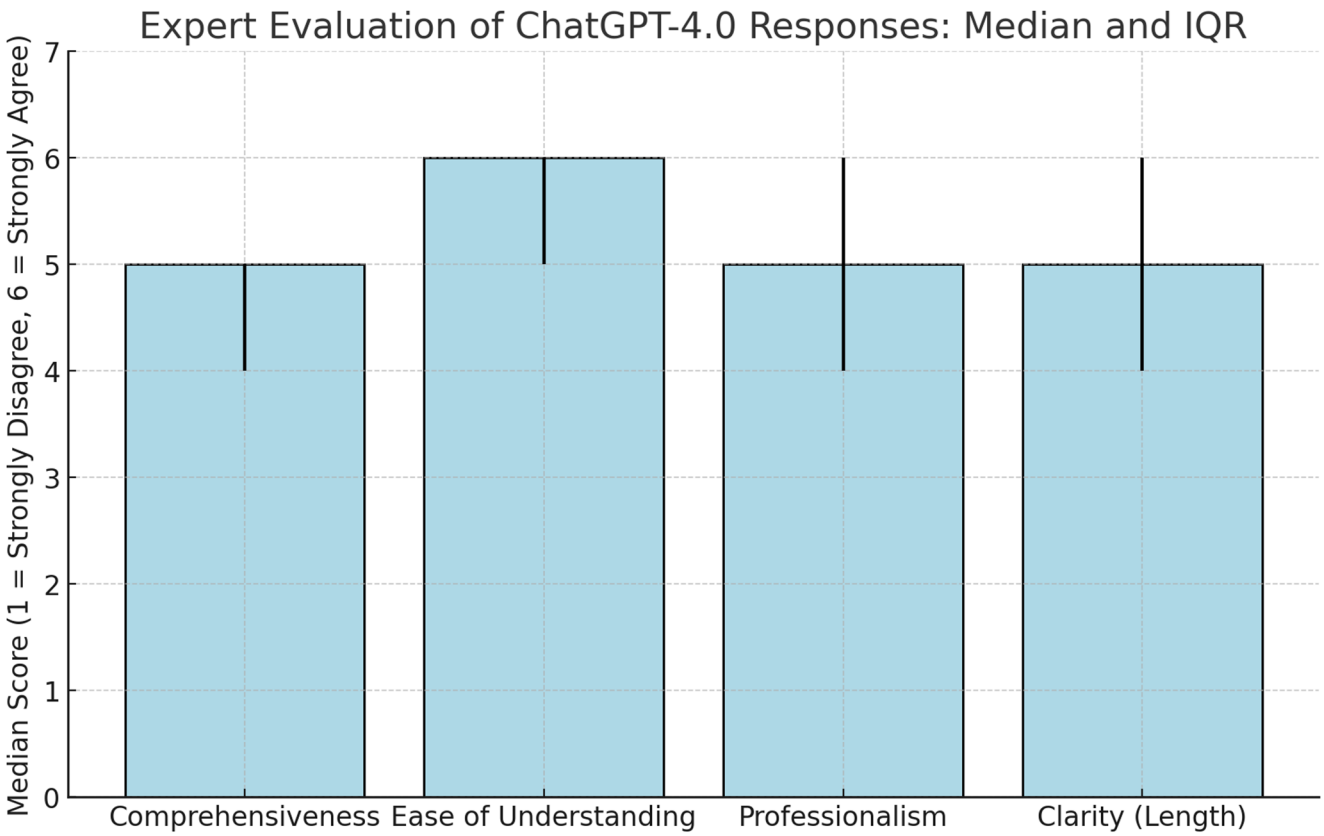

Fig. 1 Expert evaluation of the three questions with the lowest Content Validity Ratio (CVR). Interestingly, these items also showed the lowest inter-rater agreement, reflecting a higher level of disagreement among experts regarding the appropriateness of ChatGPT's responses. Panel A: Q1– What is scoliosis? Panel B: Q10– Can exercises or physical therapy cure scoliosis? Panel C: Q14– What is the best sport for scoliosis?

Table 3 Content validity ratios (CVR) and content validity indexes (CVI) for each question (Q1 - Q14) on scoliosis management for chat GPT-4.0

ChatGPT-4.0	Raters' scores	Scores of 4–6	CVR	CVI
Q1	4.0 (2.0)	20 (68%)	0.37	**0.69**
Q2	6.0 (1.0)	**29 (100%)**	**1.0**	
Q3	6.0 (1.0)	**29 (100%)**	**1.0**	
Q4	6.0 (1.0)	**29 (100%)**	**1.0**	
Q5	5.0 (2.0)	**25 (86.2%)**	**0.72**	
Q6	6.0 (1.0)	**26 (89.7%)**	**0.79**	
Q7	6.0 (2.0)	**24 (82.8%)**	**0.65**	
Q8	6.0 (1.0)	**27 (93.1%)**	**0.86**	
Q9	6.0 (0.0)	**28 (96.6%)**	**0.93**	
Q10	5.0 (2.0)	20 (68%)	0.37	
Q11	6.0 (1.0)	**27 (93.1%)**	**0.86**	
Q12	6.0 (1.0)	**26 (89.7%)**	**0.79**	
Q13	6.0 (1.0)	**28 (96.6%)**	**0.93**	
Q14	2.0 (2.0)	6 (20.7%)	−0.58	

Data on raters' scores are reported as median (IQR). In brackets is the percentage of agreement. In bold: CVR or CVI≥0.38

ChatGPT 4.0 content validity

The CVR met the threshold (≥0.38) in 78.5% (11/14) of responses provided by ChatGPT-4.0. Specifically, answers to "What is scoliosis?", Can exercises or physical therapy cure scoliosis?", and "What is the best sport for scoliosis?" had the lowest CVR scores of 0.37, 0.37, and −0.58, respectively (Fig. 1). For these three responses, the most common reasons for inappropriate answers were "clear mistakes in the answer" (37.1%) and "too few information, not enough for an exhaustive answer" (27.1%). Conversely, answers regarding the causes of scoliosis, worsening of scoliosis over time, and future disability achieved full consensus, with 100% agreement and a CVR of 1.0. Overall, the answers provided by ChatGPT-4.0 on scoliosis management were rated as valid by the experts with a CVI of 0.68. Table 3 shows the CVRs for the 14 questions.

Reasons for the inappropriateness of ChatGPT-4.0 answers

About scoliosis treatment, 33.2% reported a lack of sufficient detail, categorized as "too few information, not enough for an exhaustive answer". "Clear mistakes in the answer" were reported in 22.5% of cases, whereas other critiques were "too much information, not all necessary" (2.7%) and "language issues, not suitable for patients" (2.1%). 38.5% fell into the "other reasons" category, including diverse and context-specific critiques such as: "off topic, the answer is not pertinent to the question", "forgot to mention exercises against progression", and "does not mention specialized healthcare provider". Results showed a statistically significant difference in the distribution of inappropriate reasons across questions ($\chi^2 = 112.49$, $p = 0.0002$), suggesting that certain queries were particularly prone to issues like factual inaccuracy. In detail, "What is the best sport for scoliosis?" attracted the highest volume of critical feedback, with a particularly high number of factual errors cited (19, 65.5%). This indicates a significant concern regarding the accuracy and appropriateness of that specific answer.

Comprehensiveness, clarity, professionalism, and length

Expert evaluations indicated overall positive perceptions of the ChatGPT-generated responses. The "ease of understanding" dimension achieved the highest ratings, with a median score of 6 and an IQR of 1, suggesting uniform clarity and readability. Comprehensiveness, professionalism, and clarity (Length) each received a median score of 5, though variability differed across these aspects. Notably, professionalism and clarity exhibited higher IQR (IQR = 2), indicating more varied opinions regarding tone and detail, whereas comprehensiveness had an IQR of 1, reflecting consistent agreement on content coverage (Fig. 2).

Advantages, risks, and expectations

In terms of expectations, experts assigned a median score of 5, with an IQR of 1. This suggests that most experts felt that ChatGPT performed in line with or exceeded their expectations. A similarly positive pattern emerged for attitudes toward AI/LLM integration in rehabilitation practice. Experts reported a median attitude score of 5, with an IQR of 1, indicating a strong general openness to the adoption of large language models within the field (Fig. 2, left panel). Perceptions of risk versus advantage presented a more nuanced picture. When asked whether the integration of LLMs in rehabilitation represented an advantage or a hazard, 17 experts (58.6%) viewed it as a moderate advantage, while 3 experts (10.3%) considered it a huge advantage. However, 5 experts (17.2%) expressed concerns, rating it as a moderate risk, and 1 expert (3.4%) classified it as a huge risk. Three experts (10.3%) opted for a neutral position, suggesting ambivalence or a balanced view of both opportunities and risks (Fig. 3, right panel).

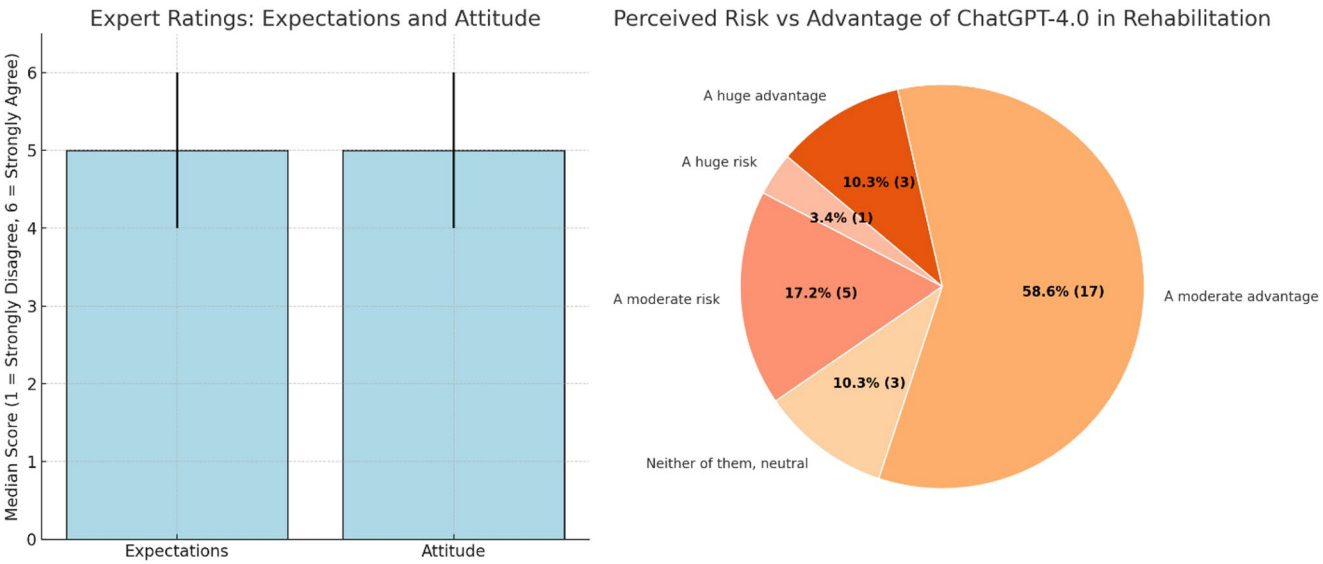

Fig. 2 Median expert ratings and interquartile ranges (IQR) for responses generated by ChatGPT 4.0 across four evaluation dimensions: Comprehensiveness, Ease of Understanding, Professionalism, and Clarity (Length). Each bar represents the median score assigned by a panel of 29 expert raters on a 6-point Likert scale (1 = Strongly Disagree, 6 = Strongly Agree). Error bars indicate IQRs

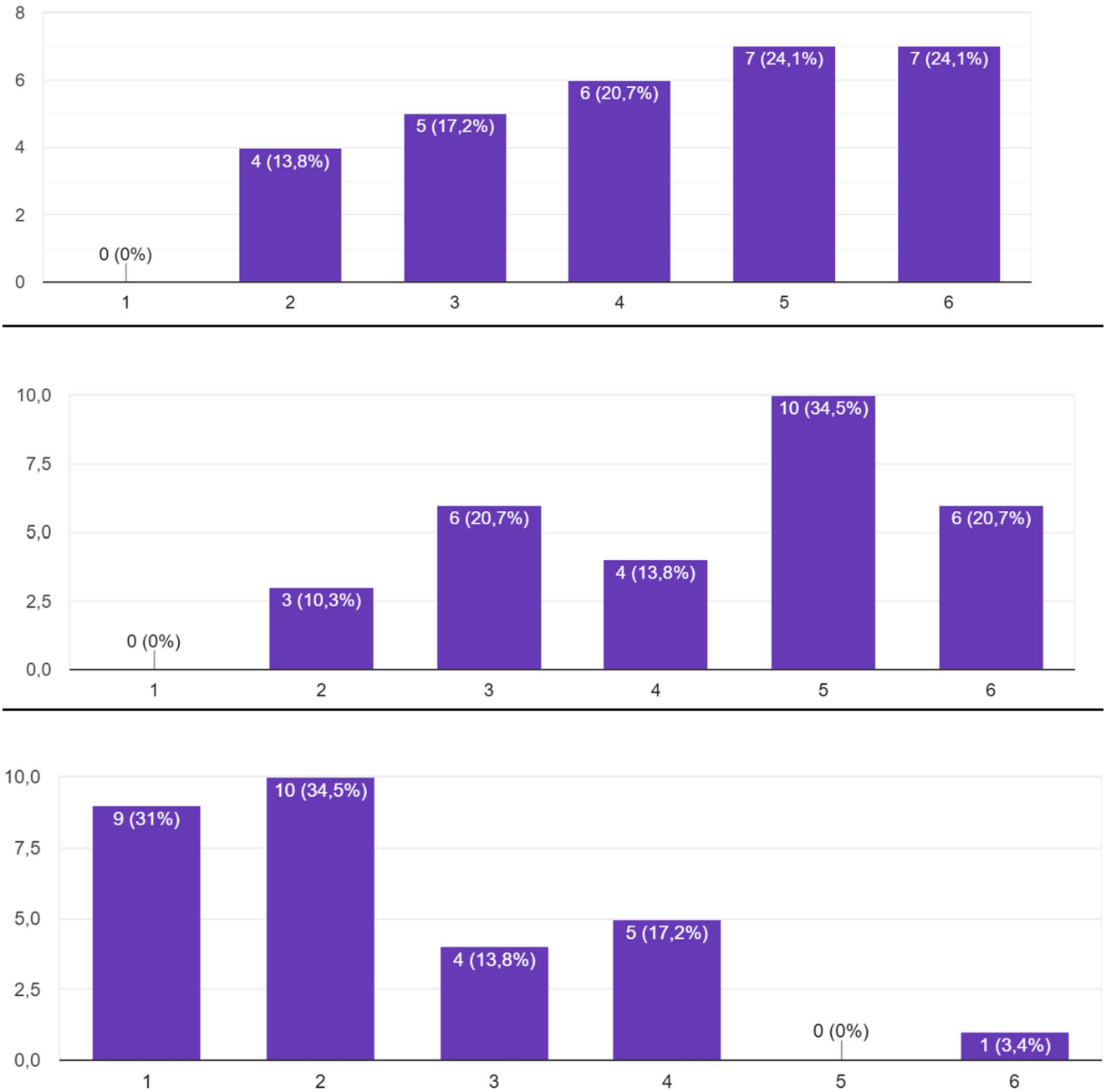

Fig. 3 Expert evaluation of ChatGPT 4.0 in the context of rehabilitation. Panel A shows the median scores (with interquartile ranges, IQR) assigned by the experts for two dimensions: Expectations (i.e., whether ChatGPT-4.0 met expert expectations) and Attitude (i.e., the general openness to integrating AI/LLMs in rehabilitation). Panel B displays a pie chart summarizing expert perceptions of the potential risk or advantage associated with using ChatGPT 4.0 in rehabilitation

Discussion

This study aimed to evaluate the accuracy, clarity, and perceived usefulness of ChatGPT-4.0 in responding to FAQs concerning the rehabilitative treatment of idiopathic scoliosis. The findings demonstrate a generally positive perception among experts regarding ChatGPT-4.0's answers and as a supportive tool in patient education and clinical communication within the context of scoliosis care. Comprehensiveness, professionalism, and appropriateness of length also received favorable ratings, although with more variability, particularly in terms of tone and level of detail. Eleven out of 14 responses met the minimum content validity ratio (CVR≥0.38), suggesting that ChatGPT-4.0's content was judged as appropriate and valid by most expert reviewers. Experts rated highly for clarity and ease of

understanding, reflecting a strong consensus on ChatGPT-4.0's ability to communicate complex medical information in a clear and accessible language. General perceptions of the model's usefulness in patient education were positive, with most experts viewing its integration in rehabilitation as a moderate to high advantage. Some caution was expressed regarding the potential risks and limitations of such tools.

Despite all raters coming from the same institution, with numerous opportunities for discussion and alignment on the clinical topics considered, the Fleiss' Kappa value indicated only slight agreement. Interestingly, the questions with the lowest CVR scores (Q1, Q10, and Q14) were those that showed the lowest inter-rater agreement. This finding highlights the challenge of reaching consensus among experts on certain questions, for which scientific literature has yet to provide clear or definitive answers.

Lang et al. conducted a comparative evaluation of three LLMs in the context of surgical management of adolescent IS, identifying ChatGPT-4.0 as the most reliable tool [13]. Building upon their findings, we explored a complementary area—conservative treatment—and applied a more structured and quantitative evaluation framework based on CVR and CVI metrics. While both studies emphasize the clarity and communicative effectiveness of ChatGPT-4.0, our results reveal how its scientific accuracy varies considerably depending on topic complexity and clinical consensus. Çıracıoğlu and Erdoğan (2025) [17] found good performance for general questions about scoliosis but reported limitations in treatment-specific responses and a high required reading level. Our study focused specifically on conservative management, involved a larger and more specialized experts panel, and applied structured content validity metrics (CVR, CVI). Despite the raters' shared background, we observed relatively low inter-rater agreement especially for clinically ambiguous questions, underlining the complexity of interpreting AI-generated health information in real-world scenarios. This result was likely because we involved a significantly higher number of experts than Çıracıoğlu and Erdoğan, who involved only two raters. We suspect that further increasing the number of evaluators - including professionals from other centers and countries - could lead to even lower agreement levels, reflecting broader variability in clinical interpretation and cultural perspectives on patient education.

A comparison with the recent study by Scaff et al. [6], which assessed LLM-chatbots' responses to common questions about low back pain, highlights interesting differences. While both studies explored the performance of ChatGPT-4.0 in musculoskeletal patient education, Scaff et al. reported a lower overall accuracy (56%) and high rates of misinformation. This discrepancy may be partly explained by the nature of the condition itself: Low back pain is a highly prevalent and controversial topic, often surrounded by misinformation and inconsistent narratives in online sources.

Our study has several limitations. First, while the expert panel was multidisciplinary and experienced, it was composed exclusively of professionals affiliated with a single institution, which may limit generalizability. Future studies could address this limitation. Second, ChatGPT responses were assessed at a single time point, and only one version of each answer was rated, despite the model's known variability in output. Therefore, possible fluctuations may have been missed. Third, the survey did not assess the long-term impact of using AI-generated information on patients' perception, understanding, trust or clinical outcomes, which are crucial metrics for eventual implementation. Lastly, while Likert-based ratings are informative, they do not fully capture the nuanced judgments experts may hold, despite the inclusion of categorical reasoning and comment fields. Future qualitative studies with focus groups could address this limitation. The study also presents notable strengths. The multi-phase design enabled the systematic and transparent generation of FAQs, ensuring both clinical and AI-informed relevance. The use of validated frameworks, such as CVR and CVI, offers a rigorous measure of content validity, which has been seldom applied in the AI-for-health literature to date. Moreover, the addition of risk/benefit perception analysis further enriches the understanding of the potential role and reception of AI tools in clinical rehabilitation settings.

Our results highlight two specific needs: (1) the importance of establishing greater consensus among experts on clinically significant questions that currently lack clear answers in the literature; (2) the necessity of training LLMs using input and supervision from leading international experts in improving the quality of contents.

Looking ahead, future studies should prioritize cross-institutional and multicultural designs to better understand potential differences in expert judgment and educational needs across diverse healthcare contexts. Moreover, involving patients directly in the evaluation process, particularly in assessing the comprehensibility, emotional resonance, and perceived usefulness of AI-generated responses, will be essential to ensure real-world relevance and safety. While quantitative methods such as CVR and CVI offer important objectivity, they should be supplemented by qualitative analyses (e.g., thematic content analysis or discourse analysis) to capture more nuanced aspects of communication and interpretation. Lastly, future model development efforts should consider the active participation of domain-specific experts in fine-tuning and adapting LLMs within specialized clinical contexts, including rehabilitation services and spine care units. This human-in-the-loop approach may represent

a key strategy for improving both the accuracy and trustworthiness of AI outputs in patient-facing applications.

Conclusions

This study suggests that the information provided by LLMs is generally accurate and understandable enough to serve as a preliminary source of information for patients regarding scoliosis. However, for more detailed and personalized analyses — especially for application in everyday clinical practice, where a high level of precision is required and even a single error can lead to serious health consequences — the involvement of expert clinicians remains essential. Future research should explore patients' perspectives, focusing on their understanding, emotional response, and trust when accessing AI-generated information about scoliosis.

Acknowledgements The work of FN, GF, GM, CK, and SN was supported and funded by the Italian Ministry of Health - Ricerca Corrente.

Author contributions Study concept and design: F. Negrini, C. Malfitano and J. Vitale; Acquisition of data: F. Negrini, S. Negrini, I. Ferrario, J. Vitale; Statistical analysis: J. Vitale; Analysis and interpretation of data: all authors; Drafting of the manuscript: all authors; Critical revision of the manuscript for important intellectual content: all authors.

Data availability The anonymised data collected are available as open data via Zenodo using the DOI 10.5281/zenodo.15213885.

Declarations

Competing interests SN and AN own stock of ISICO (Italian Scientific Spine Institute). FN related to SN and AN.

References

1. Alowais SA, Alghamdi SS, Alsuhebany N et al (2023) Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 23. https://doi.org/10.1186/S12909-023-04698-Z
2. Morone G, De Angelis L, Martino Cinnera A et al (2025) Artificial intelligence in clinical medicine: a state-of-the-art overview of systematic reviews with methodological recommendations for improved reporting. Front Digit Health 7. https://doi.org/10.3389/FDGTH.2025.1550731
3. Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 388:1233–1239. https://doi.org/10.1056/NEJMSR2214184
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K et al (2023) Large Language models in medicine. Nat Med 29:1930–1940. https://doi.org/10.1038/S41591-023-02448-8
5. Quinn M, Milner JD, Schmitt P et al (2024) Artificial intelligence large Language models address anterior cruciate ligament reconstruction: superior clarity and completeness by gemini compared with ChatGPT-4 in response to American academy of orthopaedic surgeons clinical practice guidelines. https://doi.org/10.1016/J.ARTHRO.2024.09.020. Arthroscopy
6. Scaff SPS, Reis FJJ, Ferreira GE et al (2025) Assessing the performance of AI chatbots in answering patients' common questions about low back pain. Ann Rheum Dis 84. https://doi.org/10.1136/ARD-2024-226202
7. Lang S, Vitale J, Fekete TF et al (2024) Are large Language models valid tools for patient information on lumbar disc herniation? The spine surgeons' perspective. Brain Spine 4. https://doi.org/10.1016/J.BAS.2024.102804
8. Yeo YH, Samaan JS, Ng WH et al (2023) Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 29:721–732. https://doi.org/10.3350/CMH.2023.0089
9. Lyu Q, Tan J, Zapadka ME et al (2023) Translating radiology reports into plain Language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 6. https://doi.org/10.1186/S42492-023-00136-5
10. Eggmann F, Weiger R, Zitzmann NU, Blatz MB (2023) Implications of large Language models such as ChatGPT for dental medicine. J Esthet Restor Dent 35:1098–1102. https://doi.org/10.1111/JERD.13046
11. Li X, Huo Z, Hu Z et al (2022) Which interventions May improve bracing compliance in adolescent idiopathic scoliosis? A systematic review and meta-analysis. PLoS ONE 17. https://doi.org/10.1371/JOURNAL.PONE.0271612
12. Su J, Ng DTK, Chu SKW (2023) Artificial intelligence (AI) literacy in early childhood education: the challenges and opportunities. Computers Education: Artif Intell 4:100124. https://doi.org/10.1016/J.CAEAI.2023.100124
13. Lang S, Vitale J, Galbusera F et al (2024) Is the information provided by large Language models valid in educating patients about adolescent idiopathic scoliosis? An evaluation of content, clarity, and empathy: the perspective of the European spine study group. https://doi.org/10.1007/S43390-024-00955-3. Spine Deform
14. Vasey B, Nagendran M, Campbell B et al (2022) Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. https://doi.org/10.1136/BMJ-2022-070904. BMJ 377:
15. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 22:276. https://doi.org/10.11613/bm.2012.031
16. Lawshe CH (1975) A quantitative approach to content validity. Pers Psychol 28:563–575. https://doi.org/10.1111/J.1744-6570.1975.TB01393.X
17. Çıracıoğlu AM, Dal Erdoğan S (2025) Evaluation of the reliability, usefulness, quality and readability of chatgpt's responses on scoliosis. Eur J Orthop Surg Traumatol 35. https://doi.org/10.1007/S00590-025-04198-4

Authors and Affiliations

Francesco Negrini[1,2] · Calogero Malfitano[3,4] · Giorgio Ferriero[1,2] · Giovanni Morone[5,6] · Alberto Negrini[7] · Fabio Zaina[7] · Irene Ferrario[7] · Carlotte Kiekens[8] · Stefano Negrini[8,9] · Jacopo Vitale[10,11]

✉ Calogero Malfitano
calogero.malfitano@unimi.it

Francesco Negrini
francesco.negrini@uninsubria.it

Giorgio Ferriero
giorgio.ferriero@uninsubria.it

Giovanni Morone
giovanni.morone@univaq.it

Alberto Negrini
alberto.negrini@isico.it

Fabio Zaina
fabio.zaina@isico.it

Irene Ferrario
irene.ferrario@isico.it

Carlotte Kiekens
carlotte.kiekens@isico.it

Stefano Negrini
stefano.negrini@unimi.it

Jacopo Vitale
Jacopo.Vitale@kws.ch

1 Department of Biotechnology and Life Sciences, University of Insubria, Varese, Italy

2 Institute of Tradate, Istituti Clinici Scientifici Maugeri IRCCS, Tradate, Italy

3 Department of Biomedical Sciences for Health, University of Milan, Milan, Italy

4 Azienda di Servizi alla Persona Istituti Milanesi Martinitt e Stelline e Pio Albergo Trivulzio, Milan, Italy

5 Department of Life, Health and Environmental Sciences, University of L Aquila, L'Aquila, Italy

6 IRCCS Fondazione Santa Lucia, Rome, Italy

7 ISICO (Italian Scientific Spine Institute), Milan, Italy

8 IRCCS Galeazzi– Sant'Ambrogio Hospital, Milan, Italy

9 Department of Biomedical, Surgical and Dental Sciences, University of Milan, Milan, Italy

10 Spine Center, Schulthess-Klinik, Zurich, Switzerland

11 Institute of Sports Sciences, University of Physical Culture in Cracow, Cracow, Poland