Abstract

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

Authors. Francesco Negrini[1,2], Jacopo Vitale[3], Calogero Malfitano[4,5], Alberto Negrini[6], Fabio Zaina[6], Irene Ferrario[6], Stefano Negrini[7,8]

Affiliations.

1. Istituti Clinici Scientifici Maugeri, Institute of Tradate, IRCCS, Tradate, Varese, Italy
2. Department of Biotechnology and Life Sciences, University of Insubria, Varese, Italy.
3. Spine Center, Schulthess Klinik, Zürich, Switzerland
4. Department of Biomedical Sciences for Health, University "La Statale", Milan, Italy
5. Azienda di Servizi alla Persona Istituti Milanesi Martinitt e Stelline e Pio Albergo Trivulzio, Milan, Italy
6. ISICO (Italian Scientific Spine Institute), Milan, Italy.
7. Department of Biomedical, Surgical and Dental Sciences, University "La Statale", Milan, Italy
8. IRCCS Istituto Ortopedico Galeazzi, Milan, Italy.

Background. Large Language Models (LLMs) like ChatGPT have become accessible tools for patients and caregivers seeking medical guidance. However, their accuracy, clarity, and validity in specialized contexts, such as scoliosis conservative treatment, are largely unexamined.

Objective. This study evaluates whether ChatGPT 4.0 provides evidence-based, appropriate, and comprehensive answers to common questions about scoliosis management.

Study design. Cross-sectional observational

Methods. Between November and December 2024, 14 FAQs on scoliosis conservative treatment were identified through expert input and LLM-generated suggestions. These were submitted to ChatGPT 4.0 on the same day (06/12/2024) with the prompt: "I'm a scoliosis patient. Limit your answer to 150 words." Responses were evaluated by 29 multidisciplinary (see Figure 1 for details on the professionals involved) scoliosis experts (mean age 41.2 ± 10.8 years; 55.2% females, 44.8% males; median years of experience 9.5, IQR 6.0-14.0) using a 6-point Likert scale through Google Forms surveys (response rate 29/37, 78%). All data are considered as categorical variables and are reported as percentages (%). Inter-rater reliability was measured using Fleiss' Kappa. Content validity was assessed using the Content Validity Ratio (CVR). CVR was calculated as follows: CVR = $[n_e-(N/2)]/(N/2)$ where n_e is the number of experts rating an answer as "appropriate" and N is the total number of experts. Scores of 4–6 (Likert scale) were considered as "appropriate" and 1–3 as "non-appropriate." A minimum CVR of **0.38** indicated valid answers (69% agreement).

Results. Fleiss' Kappa showed slight agreement (0.10). The CVR threshold (≥0.38) was met in 78.5% (11/14) of responses. Answers to "What is scoliosis?", "Can exercises or physical therapy

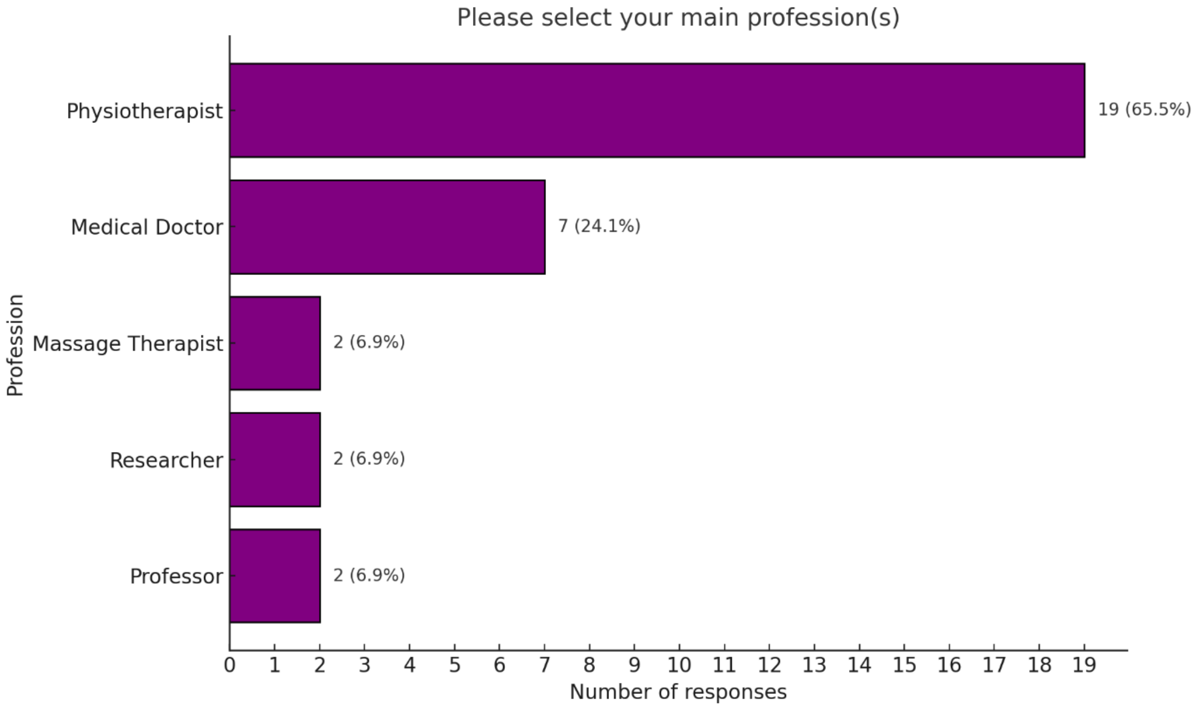
cure scoliosis?", and "What is the best sport for scoliosis?" scored 0.37, 0.37, and -0.58, respectively, due to factual errors (37.1%) and incomplete information (27.1%). Conversely, responses on the causes of scoliosis, its progression, and future disability achieved full consensus (CVR 1.0).

For comprehensiveness, 72.4% of experts agreed, while 6.9% disagreed. Clarity received 86.2% agreement with no disagreements. Professionalism was rated positively by 68.9%, with minor disagreement (6.8%). Minimal risks were perceived, and only 20.7% of experts expressed moderate concerns. Overall, 89.7% felt their expectations were met.

Conclusion. The study highlights that ChatGPT 4.0 delivers clear and professional answers, especially on well-documented medical topics like the causes and progression of scoliosis. However, its performance was uneven in some areas, with factual errors and incomplete answers surfacing. Experts' mixed evaluations point to the challenges of interpreting responses in a complex field like scoliosis treatment. Training LLMs with expert-reviewed guidelines is crucial to enhance reliability and accuracy.

Clinical Significance. Despite some inconsistencies, the generally positive feedback suggests that AI, like ChatGPT, can be a valuable tool for patient education—provided its limitations are carefully managed.

Figure 1.


Please select your main profession(s)
Profession
Physiotherapist
19 (65.5%)
Medical Doctor
7 (24.1%)
Massage Therapist
2 (6.9%)
Researcher
2 (6.9%)
Professor
2 (6.9%)
Number of responses