European Spine Journal
https://doi.org/10.1007/s00586-026-10046-8

RESEARCH

# An expert-led benchmark using common patient questions: evaluating large language models for adolescent idiopathic scoliosis education

**Cemre Aydin[1] · Asli Beril Karakas[2] · Anil Murat Ozturk[3] · Figen Govsa[4] · Mehmet Asim Ozer[4]**

Received: 5 February 2026 / Revised: 7 May 2026 / Accepted: 23 May 2026


## Abstract

**Aim/background** Large language models (LLMs) are increasingly used by patients to obtain medical information. Adolescent idiopathic scoliosis (AIS), a chronic condition requiring long-term monitoring and treatment decisions, generates substantial demand for reliable and understandable patient education. Although LLMs may function as accessible explanatory tools, their suitability for patient-oriented use remains uncertain. This study aimed to perform an expert-led, patient-centered evaluation of two widely accessible LLMs, Claude Sonnet 4.5 and GPT 5.2, focusing on their ability to deliver accurate, clear, and conceptually adequate responses to common AIS-related patient questions.

**Methods** A cross-sectional comparative design was used with 100 high-frequency patient questions covering ten clinical domains. Responses generated by both models using standardized zero-shot prompts were independently assessed by expert clinicians: factual accuracy by three raters (two orthopedic spine surgeons and one senior pediatric physiotherapist), and clarity and conceptual coverage by two raters (one surgeon and the physiotherapist). A structured evaluation framework examined three dichotomous dimensions relevant to patient education: factual accuracy, clarity and understandability, and conceptual coverage. Model performances were compared using McNemar's test, and inter-model agreement was assessed with Krippendorff's alpha.

**Results** Both models demonstrated equally high factual accuracy (91%). However, clarity was limited, with only one-third of responses rated as sufficiently understandable. A significant difference was observed in conceptual coverage, with Claude Sonnet 4.5 outperforming GPT 5.2 (46% vs. 29%, $p = 0.012$), particularly in domains requiring integrative explanations.

**Conclusion** Despite strong factual accuracy, current LLMs show deficiencies in clarity and conceptual depth, limiting their reliability as standalone patient education tools for AIS. These findings highlight the necessity of clinician mediation and the importance of patient-centered evaluation criteria before clinical adoption.

**Clinical trial registration** As this study is not a clinical trial, clinical trial registration is not applicable.

**Keywords** Large language models · Patient education · Adolescent idiopathic scoliosis · Health literacy · Artificial intelligence · Clinical communication

✉ Asli Beril Karakas
asliberilkarakas@gmail.com

✉ Anil Murat Ozturk
amuratozturk@yahoo.com

Cemre Aydin
caydin9119@gmail.com

Figen Govsa
fgovsa@yahoo.com

Mehmet Asim Ozer
mehmet.asim.ozer@gmail.com

[1] Department of Orthopedics and Traumatology, Istanbul Bağcılar Education Research Hospital, 34200 Istanbul, Turkey

[2] Department of Anatomy, Faculty of Medicine, Kastamonu University, 37150 Kastamonu, Turkey

[3] Department of Orthopedics and Traumatology, Faculty of Medicine, Ege University, 35040 Izmir, Turkey

[4] Digital Imaging and 3D Modelling Laboratory, Department of Anatomy, Faculty of Medicine, Ege University, 35040 Izmir, Turkey

Published online: 10 June 2026

## Introduction

Adolescent idiopathic scoliosis (AIS), defined as a structural spinal curvature greater than 10° with no identifiable cause, is the most common pediatric spinal deformity [1]. Its clinical management involves a prolonged and multifaceted pathway, encompassing ongoing surveillance, potential orthotic bracing, and possible surgical correction. This trajectory generates considerable uncertainty for patients and families regarding disease progression, aesthetic outcomes, functional capacity, and long-term prognosis [2, 3]. Consequently, the provision of accurate, comprehensible, and timely information is a fundamental pillar of ethical clinical practice, critically influencing informed consent, treatment adherence, and psychosocial adaptation [4].

Patients and caregivers are increasingly using large language models (LLMs) for health information, making rigorous evaluation of these tools an urgent priority [5, 6]. LLMs allow users to pose natural language questions and receive immediate, conversational responses, positioning them as on-demand sources of medical explanation [7]. However, LLMs are susceptible to generating inaccurate information ("hallucination") and demonstrate inconsistent performance on complex clinical tasks [8].

In AIS education, prior studies have reported variable accuracy. Yaş et al. [9] evaluated ChatGPT and Google Gemini on six clinical scenarios, finding moderate accuracy rates (57.1% and 52.7%, respectively) and highlighting surgeon concerns about reliability. Lang et al. [10] assessed three LLMs on ten AIS FAQs, reporting that only 39% of ChatGPT-4 responses were rated "excellent". Suresh et al. [11] compared multiple LLMs on common parent questions and noted generally favourable surgeon ratings. Focusing specifically on conservative management, Negrini et al. [12] evaluated ChatGPT-4.0 on 14 frequently asked questions, finding that while clarity was rated highly, content validity varied across topics, with experts expressing caution about factual inaccuracies in certain areas. Other investigations have examined LLM performance in scoliosis classification and surgical decision support, further underscoring the variability of these tools [13, 14]. Beyond AIS, recent work in minimally invasive spine surgery has shown that LLMs can reliably distinguish surgical from non-surgical triage ($\kappa$ = 0.42–0.59) but struggle with specific procedure selection ($\kappa$ = 0.15–0.25), a pattern that mirrors the dissociation between factual knowledge and nuanced application observed in patient education contexts [15].

These investigations, however, predominantly rely on global quality metrics and overlook the distinct informational needs of laypersons, who prioritise practical implications and quality-of-life considerations over abstract biomedical details [16, 17]. A parallel issue is the lack of accessibility in patient-facing materials. Even content authored by professional societies often fails to meet recommended readability standards, a deficit that appears replicated rather than resolved by current LLMs [18, 19]. The utility of medical information for a lay audience cannot be reduced to factual veracity alone. Clarity, the use of accessible language free of unexplained jargon, is essential for genuine comprehension. Conceptual coverage, the inclusion of core contextual elements, alternatives, and limitations, is vital to prevent technically accurate but misleadingly incomplete answers [20].

To address these gaps, this study conducted a structured, multi-dimensional benchmark of two leading publicly accessible LLMs, Claude Sonnet 4.5 and GPT 5.2, using 100 common patient questions. The objectives were: (1) to assess factual accuracy against evidence-based standards, (2) to evaluate clarity and patient-appropriateness, and (3) to analyse conceptual coverage. This expert-led evaluation offers preliminary insights for clinical practice and establishes a methodological foundation for future patient-centred AI validation.

## Methods

### Study design

A cross-sectional, comparative benchmarking design was employed to conduct a structured evaluation of two LLMs. The study aimed to assess the quality of responses generated by these models when presented with patient-centric questions on AIS at a single point in time. Quality was evaluated across three pre-defined, clinically relevant dimensions: factual accuracy, clarity of communication, and conceptual coverage.

### Development and validation of the question set

To maximize ecological validity, the stimulus questions were derived directly from the informational needs of patients and caregivers. A systematic, multi-stage approach was used to curate and validate the final question bank:

1. ***Source identification***: Potential questions were harvested from three primary source types:

a. Patient education portals of major professional societies.
b. Clinical resource websites of leading academic medical centers.
c. Moderated online forums and peer-support communities dedicated to scoliosis.

2. ***Curation and categorization:*** An initial pool of over 300 unique questions was compiled. After removing duplicates, the remaining questions were thematically analyzed and inductively categorized by two researchers (CA, ABK) achieving consensus through discussion into ten clinically coherent domains (I. Diagnosis and Basics, II. Non-Surgical Management, III. Surgical Treatment, IV. Living (Non-Surgical), V. Living (Post-Surgical), VI. Specific Populations, VII. Pain Management, VIII. Prognosis and Outlook, IX. Bracing Specifics, X. Surgical/Recovery Details).
3. ***Prioritization and selection***: Questions within each domain were ranked by their frequency of appearance across the different sources. The final instrument consisted of the 100 highest-frequency questions, ensuring the set reflected the most common and pressing patient concerns.
4. ***Standardization***: Each selected question was linguistically standardized (Each question was edited for grammatical clarity while preserving its original meaning and patient-oriented phrasing.).

Our question curation process was designed to maximize ecological validity by sourcing real patient and caregiver inquiries from diverse, trusted platforms. This approach contrasts with methods that rely solely on search engine results or AI-generated questions, ensuring our benchmark reflects authentic informational needs.

## LLM selection and query execution

Two state-of-the-art, publicly accessible LLMs were selected for head-to-head comparison: Claude Sonnet 4.5 (Anthropic, Application Programming Interface (API) version 2024-05-01) and GPT 5.2 (OpenAI's ChatGPT, API version 2024-05-13). These models were chosen based on their widespread adoption, recognized technical sophistication, and representation of leading architectures from distinct AI research organizations.

To ensure standardization and reproducibility, all queries were executed programmatically via the official APIs for each model. The following protocol was strictly adhered to:

a. ***Prompting strategy***: A zero-shot prompting approach was used. Zero-shot prompting means that each question was submitted as an isolated, context-free prompt without any example responses or prior conversational turns. Each of the 100 standardized questions was submitted as an isolated, context-free prompt.
b. ***Parameter configuration***: API calls utilized standardized parameters (temperature = 1.0 for GPT 5.2, default settings for Claude Sonnet 4.5) to simulate a typical user's interaction with a public chatbot interface.
c. ***Data logging***: For full transparency, every interaction was logged, capturing the exact input prompt, the complete, unedited model response, a timestamp, and the specific model version identifier.

## Development of the patient-centric evaluation framework

Given the absence of a validated tool for assessing patient-oriented AI responses, a novel evaluation framework was developed de novo through an iterative, consensus-driven process involving two board-certified orthopedic spine surgeons. The framework defined three independent, dichotomous rating scales, each targeting a fundamental pillar of effective patient communication.

Unlike composite or Likert-based global quality scores commonly used in prior evaluations, this framework deliberately employs dichotomous rating scales. This design choice was intended to isolate clinically relevant failure modes that are directly actionable from a patient safety perspective. In the context of patient education, partial correctness or borderline clarity may still lead to misunderstanding, anxiety, or inappropriate decision-making. Therefore, responses were classified as either meeting or failing minimum adequacy thresholds for factual accuracy, clarity, and conceptual coverage.

1. ***Factual accuracy***: This scale assessed the veracity of medical information. A response was rated "Completely Accurate" only if all contained statements were fully congruent with the latest evidence-based clinical guidelines and contemporary peer-reviewed literature. The presence of any factual error, critical omission, or reliance on outdated knowledge resulted in a rating of "Not Completely Accurate." Discordant ratings among the three raters were resolved by a consensus meeting to arrive at a final unanimous rating.
2. ***Clarity and understandability***: This scale evaluated the communicative efficacy for a lay audience. A response was rated "Sufficiently Clear and Understandable" if it employed language accessible to a non-specialist, actively avoided or clearly defined technical jargon, was logically structured, and provided necessary context for complex concepts. The rater was instructed to use the ≤8th-grade reading level as a benchmark, informed by established health literacy guidelines [19]. Responses judged to be overly technical, disorganized, or ambiguous received a rating of "Not Sufficiently Clear." A post-hoc Flesch-Kincaid Grade Level (FKGL) analysis

was performed as objective validation of expert ratings; this analysis did not inform rater judgments.

3. ***Conceptual coverage and depth***: This scale appraised the comprehensiveness and contextual richness of the answer. A response was rated "Adequate Coverage" if it addressed the essential conceptual elements and contextual factors required to provide a substantively complete answer to the patient's explicit and implicit query. This included, where pertinent, discussion of indications, limitations, alternative options, or prognostic nuances. Superficial responses that omitted these central pillars were rated "Inadequate Coverage."

Clarity and conceptual coverage were treated as orthogonal constructs. Separating these dimensions allows for more precise identification of communication deficits relevant to patient understanding.

## Expert rating procedure and blinding protocol

All 200 responses (100 per model) were evaluated by a panel of three independent, blinded raters: two orthopedic spine surgeons (15 + years of subspecialty practice), and a senior pediatric physiotherapist specializing in spinal deformities (8 + years of clinical and patient education experience). However, the number of raters per dimension differed based on the nature of the assessment:

**Factual accuracy** *All three raters (two surgeons and one physiotherapist) independently assessed each response. This allowed consensus-based resolution of any factual disputes.*

**Clarity and conceptual coverage** Two raters (one spine surgeon and the physiotherapist) independently assessed each response. These subjective dimensions were evaluated by the rater pair with the most direct patient education experience.

To eliminate assessment bias, a rigorous blinding procedure was implemented:

a. All model outputs were compiled into a single document.
b. All source identifiers (model name, metadata) were removed.
c. The order of the 200 responses was fully randomized.

The rater then assessed each blinded response independently and sequentially against the three defined scales, with no knowledge of its origin. For clarity and conceptual coverage, in cases where the two raters disagreed (split-decisions: $n = 12$ for Clarity, $n = 15$ for Conceptual Coverage), the rating of the senior spine surgeon (who served as the third rater for factual accuracy) was applied as a tie-breaker per a pre-specified protocol.

## Statistical analysis

Inter-rater reliability (IRR) was assessed separately for each dimension according to the number of raters involved. For Factual Accuracy (three raters), Fleiss' Kappa (κ) was calculated to measure initial agreement prior to the consensus meeting. For Clarity and Conceptual Coverage (two raters), Cohen's Kappa (κ) was used. All κ values were interpreted using established benchmarks: <0.00 (Poor), 0.00–0.20 (Slight), 0.21–0.40 (Fair), 0.41–0.60 (Moderate), 0.61–0.80 (Substantial), and 0.81–1.00 (Almost Perfect) [20].

After reliability was established, final ratings were determined as follows: For Factual Accuracy, any initially discordant ratings among the three raters were resolved by a consensus meeting to obtain a unanimous final rating. For Clarity and Conceptual Coverage, final ratings were based on the two raters' independent assessments; in cases of disagreement ($n = 12$ for Clarity, $n = 15$ for Conceptual Coverage), the senior spine surgeon's rating served as the tie-breaker per a pre-specified protocol.

Model performance for each dimension was calculated as the proportion of responses receiving a positive rating (Accurate, Clear, or Adequate Coverage). McNemar's test was used to determine if observed differences in these paired proportional outcomes between Claude Sonnet 4.5 and GPT 5.2 were statistically significant.

To quantify inter-model agreement, defined as the consistency with which both models received identical final ratings, Krippendorff's Alpha reliability coefficient was calculated for each quality dimension. This measure was selected for its documented robustness in handling datasets with potentially high agreement and unbalanced marginal distributions [20].

To objectively validate the clarity assessments, we calculated the FKGL for all 200 responses using Python's textstat library (version 0.7.2). This analysis was performed post-hoc and did not influence the raters' judgments. The mean FKGL was compared between responses rated "Clear" and "Not Clear" for each model using the Mann-Whitney U test, given the non-normal distribution of readability scores.

All statistical analyses were performed using MedCalc Statistical Software version 20.218, with a two-sided alpha level of 0.05 denoting statistical significance.

# Results

The expert evaluation of 200 AI-generated responses (100 per model) across the three predefined dimensions, factual accuracy, clarity, and conceptual coverage, revealed

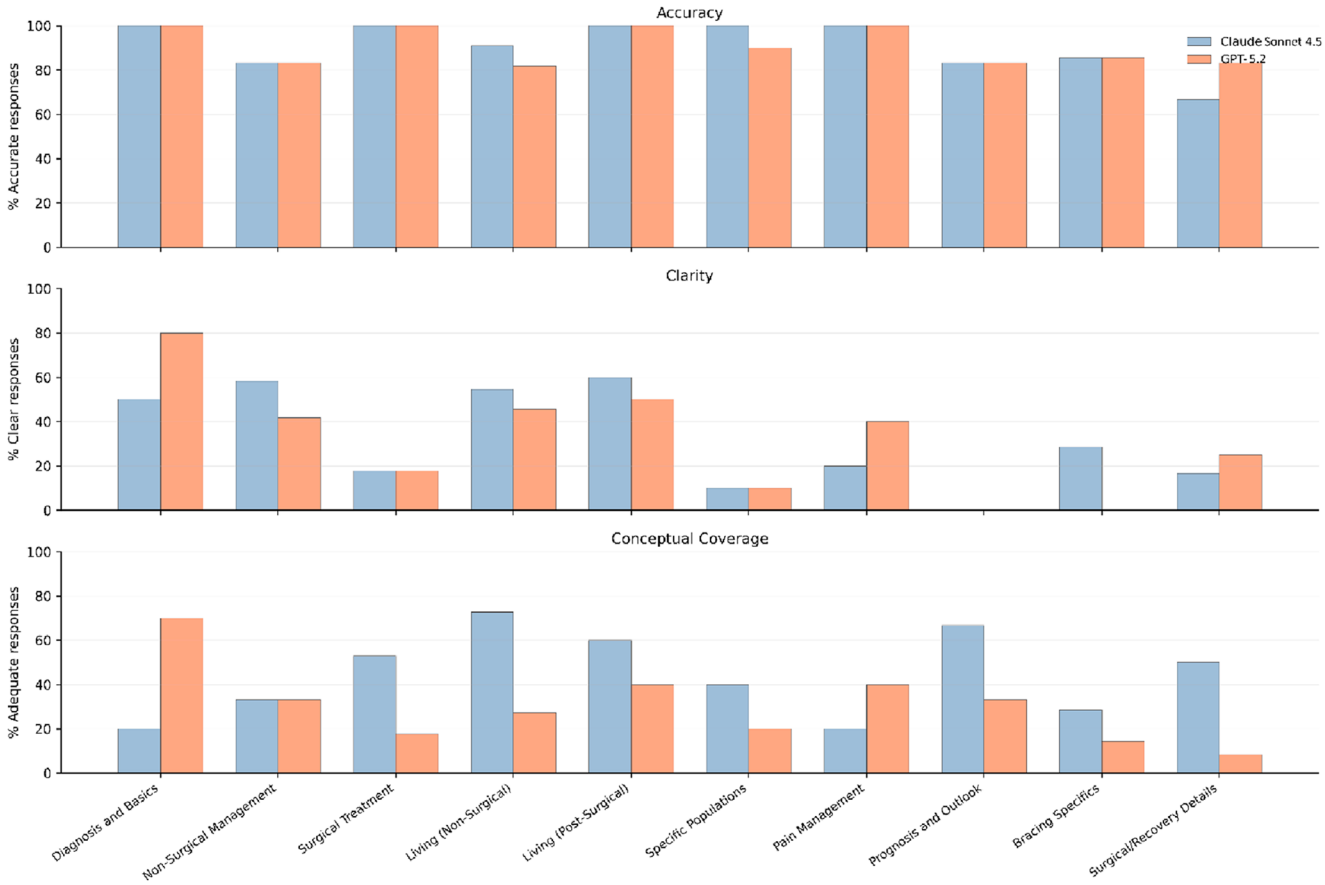


**Fig. 1** Category-based performance comparison

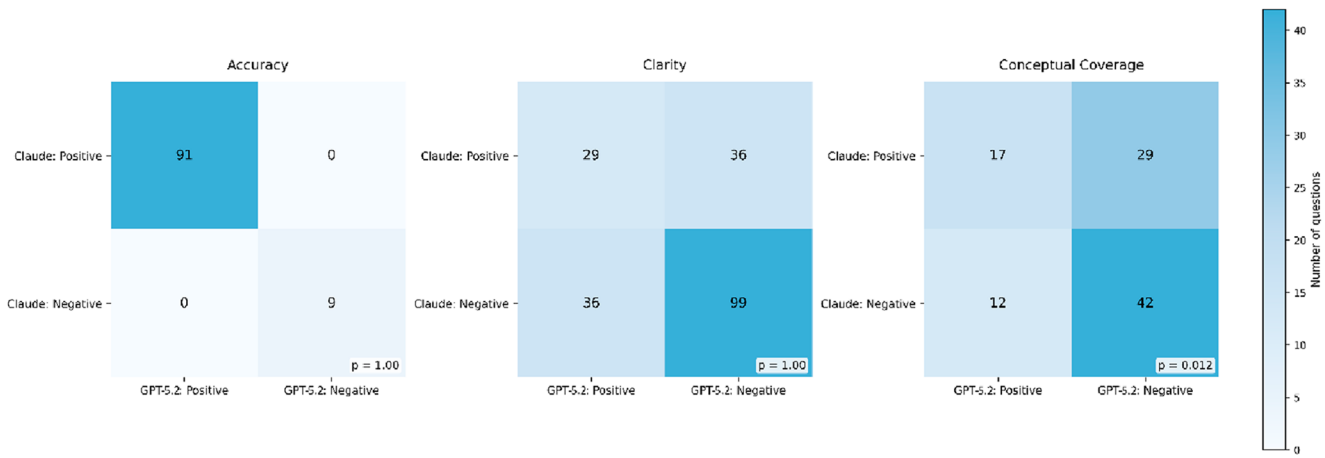


**Fig. 2** Paired agreement between Claude Sonnet 4.5 and GPT 5.2(McNemar analysis)

a complex performance profile. The results demonstrate a critical dissociation between the models' strong factual knowledge and their ability to effectively communicate that knowledge to a lay audience. Category-based performance differences between Claude Sonnet 4.5 and GPT 5.2 across factual accuracy, clarity, and conceptual coverage are presented in Fig. 1. Paired agreement patterns between the two models across all dimensions are shown in Fig. 2.

**Table 1** Accuracy performance stratified by clinical question category

| Question category | Claude sonnet 4.5 (accurate) | Claude sonnet 4.5 (not accurate) | GPT 5.2 (accurate) | GPT 5.2 (not accurate) | Discordant ratings |
|---|---|---|---|---|---|
| I. Diagnosis and basics | 10 | 0 | 10 | 0 | 0 |
| II. Non-surgical management | 10 | 2 | 10 | 2 | 0 |
| III. Surgical treatment | 17 | 0 | 17 | 0 | 0 |
| IV. Living (non-surgical) | 10 | 1 | 9 | 2 | 1 |
| V. Living (post-surgical) | 10 | 0 | 10 | 0 | 0 |
| VI. Specific populations | 10 | 0 | 9 | 1 | 1 |
| VII. Pain management | 5 | 0 | 5 | 0 | 0 |
| VIII. Prognosis and outlook | 5 | 1 | 5 | 1 | 0 |
| IX. Bracing specifics | 6 | 1 | 6 | 1 | 2 |
| X. Surgical/recovery details | 8 | 4 | 10 | 2 | 2 |
| **TOTAL** | **91** | **9** | **91** | **9** | **6** |

Discordant Ratings: Number of question pairs where the two models received different accuracy ratings

**Table 2** Contingency analysis of factual paired accuracy assessments

| | GPT 5.2: accurate | GPT 5.2: not accurate | Row total |
|---|---|---|---|
| **Claude sonnet 4.5: accurate** | 88 | 3 | 91 |
| **Claude sonnet 4.5: not accurate** | 3 | 6 | 9 |
| **Column total** | 91 | 9 | 100 |

McNemar's test, p = 1.00; Total agreement = 94% (both models accurate for 88 questions and inaccurate for 6)

## Inter-rater reliability and determination of final ratings

For Factual Accuracy (three raters), initial inter-rater agreement was substantial (Fleiss' $\kappa = 0.72$, 95% CI: 0.63–0.81). Discordant cases ($n = 6$) were resolved by a consensus meeting to obtain a unanimous final rating for each response. For Clarity and Understandability (two raters), Cohen's κ was 0.71 (95% CI: 0.62–0.80), indicating substantial agreement. The two raters disagreed on 12 of 200 responses (6%); these split-decisions were resolved by the senior spine surgeon's rating per pre-specified protocol. For Conceptual Coverage and Depth (two raters), Cohen's κ was 0.65 (95% CI: 0.56–0.74), also indicating substantial agreement. Disagreements occurred on 15 of 200 responses (7.5%), resolved by the same tie-breaking procedure. All reported performance proportions are based on these final, consolidated ratings.

## Factual accuracy: high concordance with substantial inter-model agreement

Assessment of factual reliability yielded highly convergent results. Both Claude Sonnet 4.5 and GPT 5.2 were rated "Completely Accurate" for 91 of 100 responses (91%). This indicates a high degree of alignment with contemporary evidence-based guidelines and peer-reviewed literature. Performance was consistently robust across most of the ten clinical domains, as detailed in Table 1. Notably, both models achieved perfect scores in categories such as Diagnosis and Basics and Surgical Treatment.

Statistical comparison using McNemar's test confirmed no significant difference in the overall accuracy proportions between the two models ($p = 1.00$). Analysis of inter-model agreement revealed substantial concordance (Krippendorff's $\alpha = 0.636$, 95% CI: 0.294–0.879). The contingency table for paired responses (Table 2) shows an overall agreement rate of 94% (both models accurate for 88 questions and inaccurate for 6), with discordance occurring in only 6 instances without systematic bias toward either model.

## Clarity: a prevalent deficiency with moderate inter-model agreement

In stark contrast to the high accuracy scores, the clarity of communication was markedly deficient for both LLMs. Only 33% of Claude Sonnet 4.5 responses and 32% of GPT 5.2 responses were rated 'Sufficiently Clear and Understandable." This indicates that approximately two-thirds of all generated content was deemed inaccessible due to unexplained jargon, poor structure, or ambiguous phrasing. Performance was particularly weak in categories involving technical or prognostic details (Table 3).

McNemar's test showed no significant difference in clarity performance between models ($p = 1.00$). Inter-model agreement on clarity ratings was moderate (Krippendorff's $\alpha = 0.569$, 95% CI: 0.384–0.734), with an 81% overall agreement rate (Table 4). Category-level analysis revealed the only notable variation: GPT 5.2 provided clearer explanations for basic diagnostic concepts, whereas both models consistently failed to clarify complex topics such as bracing specifics and surgical recovery.

**Table 3** Clarity performance stratified by clinical question category

| Question category | Claude sonnet 4.5 (clear) | Claude sonnet 4.5 (not clear) | GPT 5.2 (clear) | GPT 5.2 (not clear) |
|---|---|---|---|---|
| I. Diagnosis and basics | 5 | 5 | 8 | 2 |
| II. Non-surgical management | 7 | 5 | 5 | 7 |
| III. Surgical treatment | 3 | 14 | 3 | 14 |
| IV. Living (non-surgical) | 6 | 5 | 5 | 6 |
| V. Living (post-surgical) | 6 | 4 | 5 | 5 |
| VI. Specific populations | 1 | 9 | 1 | 9 |
| VII. Pain management | 1 | 4 | 2 | 3 |
| VIII. Prognosis and outlook | 0 | 6 | 0 | 6 |
| IX. Bracing specifics | 2 | 5 | 0 | 7 |
| X. Surgical/recovery details | 2 | 10 | 3 | 9 |
| **TOTAL** | **33** | **67** | **32** | **68** |

**Table 4** Contingency analysis of paired clarity assessments

| | GPT 5.2: clear | GPT 5.2: not clear | Row total |
|---|---|---|---|
| **Claude sonnet 4.5: clear** | 23 | 10 | 33 |
| **Claude sonnet 4.5: not clear** | 9 | 58 | 67 |
| **Column total** | 32 | 68 | 100 |

McNemar's test, p = 1.00; Total agreement = 81%

**Table 5** Flesch-kincaid grade level (FKGL) by clarity rating and model

| Model | Clarity rating | *n* | Mean FKGL (±SD) | FKGL range | Comparison (*p*-value) |
|---|---|---|---|---|---|
| **Claude sonnet 4.5** | Clear | 33 | 6.9 (±1.5) | 4.2–9.8 | <0.001* |
| | Not clear | 67 | 11.4 (±2.1) | 8.1–15.7 | |
| **GPT 5.2** | Clear | 32 | 7.5 (±2.0) | 4.5–10.9 | <0.001* |
| | Not clear | 68 | 11.8 (±2.5) | 8.3–16.2 | |
| **Overall** | Clear | 65 | 7.2 (±1.8) | 4.2–10.9 | <0.001* |
| | Not clear | 135 | 11.6 (±2.3) | 8.1–16.2 | |

*Mann-Whitney U test comparing FKGL between Clear vs. Not Clear responses within each model and overall

### Readability analysis validation

The post-hoc readability analysis provided objective support for the expert's clarity assessments. As shown in Table 5, responses rated as "Sufficiently Clear" had significantly lower FKGL scores compared to those rated "Not Sufficiently Clear" for both models (both $p<0.001$). The overall mean FKGL for "Clear" responses was 7.2 (±1.8), aligning with the ≤8th-grade benchmark, while "Not Clear" responses averaged 11.6 (±2.3), corresponding to high school to college-level reading difficulty. Differences in FKGL scores between responses rated as clear and not clear are illustrated in Fig. 3.

### Conceptual coverage: a significant performance gap with low inter-model consensus

The most pronounced and statistically significant divergence between models was observed in conceptual coverage, which evaluates response depth and comprehensiveness. Claude Sonnet 4.5 provided responses rated as having "Adequate Coverage" for 46% of questions, significantly outperforming GPT 5.2, which achieved adequate coverage in only 29% of cases ($p=0.012$). This represents a 17-percentage-point absolute difference in performance (Table 6). Category-specific patterns of discordant ratings between the two models are presented in Fig. 4.

Claude Sonnet 4.5 demonstrated a particular advantage in domains requiring synthesis of complex, multi-faceted information, most notably within Surgical Treatment and Surgical/Recovery Details. Inter-model agreement on this dimension was slight (Krippendorff's $\alpha=0.130$, 95% CI: −0.080–0.335), indicating low consistency in how the two models determined the necessary breadth and depth for a given question (Table 7). The dissociation between factual accuracy and conceptual coverage at the category level is illustrated in Fig. 5.

### Summary of core findings

Figure 6 provides a consolidated overview of model performance across the three evaluated dimensions, illustrating the dissociation between high factual accuracy and deficiencies in clarity and conceptual coverage.

## Discussion

This multidimensional benchmark reveals a critical dissociation in LLM performance for AIS patient education. Both Claude Sonnet 4.5 and GPT 5.2 demonstrated high factual accuracy (91%), but clarity was poor (~ 33%), and conceptual coverage varied significantly (46% vs. 29%, $p = 0.012$). These findings indicate that current general-purpose LLMs, while factually reliable, are not yet suitable as standalone patient education tools. This conclusion aligns with recent evaluations in both AIS conservative management and minimally invasive spine surgery triage, which similarly found that LLMs require expert oversight despite adequate performance on coarse-grained tasks [12, 15].

### The accuracy-clarity paradox

Our results align with prior AIS studies reporting moderate accuracy and surgeon concerns about LLM reliability [9, 10]. Negrini et al. [12], focusing specifically on

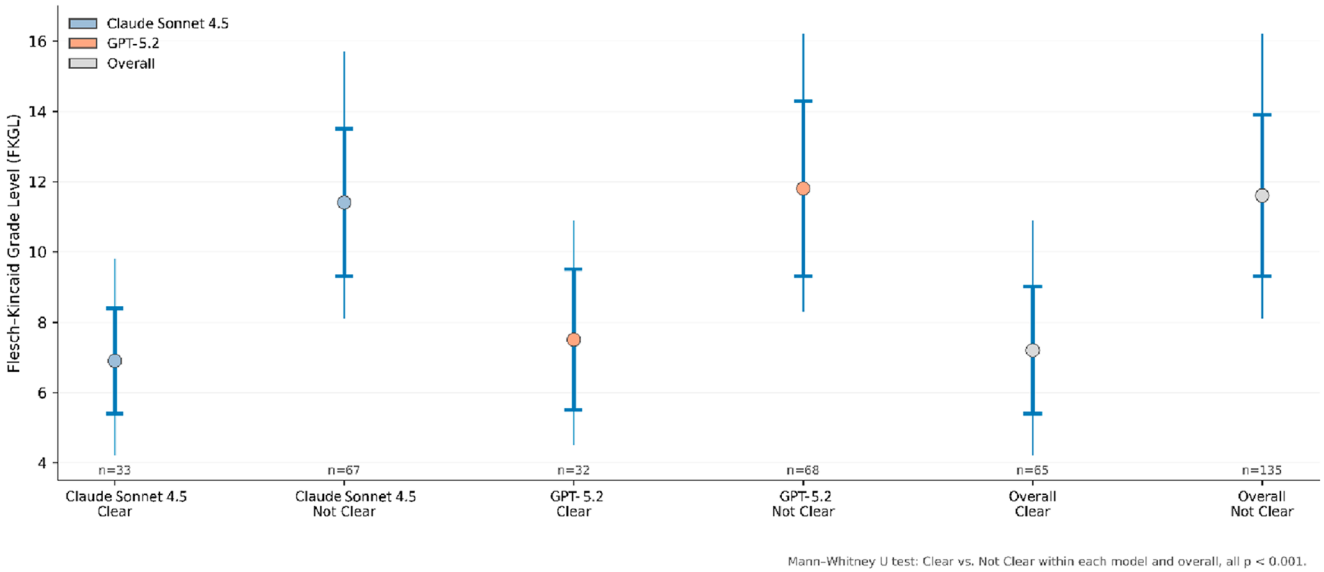


**Fig. 3** Readability validation (FKGL) by clarity rating and model

**Table 6** Conceptual coverage performance stratified by clinical question category

| Question category | Claude sonnet 4.5 (adequate) | Claude sonnet 4.5 (inadequate) | GPT 5.2 (adequate) | GPT 5.2 (inadequate) | Discordant ratings |
|---|---|---|---|---|---|
| I. Diagnosis and basics | 2 | 8 | 7 | 3 | 7 |
| II. Non-surgical management | 4 | 8 | 4 | 8 | 2 |
| III. Surgical treatment | 9 | 8 | 3 | 14 | 8 |
| IV. Living (non-surgical) | 8 | 3 | 3 | 8 | 5 |
| V. Living (post-surgical) | 6 | 4 | 4 | 6 | 4 |
| VI. Specific populations | 4 | 6 | 2 | 8 | 4 |
| VII. Pain management | 1 | 4 | 2 | 3 | 1 |
| VIII. Prognosis and outlook | 4 | 2 | 2 | 4 | 2 |
| IX. Bracing specifics | 2 | 5 | 1 | 6 | 3 |
| X. Surgical/recovery details | 6 | 6 | 1 | 11 | 5 |
| **TOTAL** | **46** | **54** | **29** | **71** | **41** |

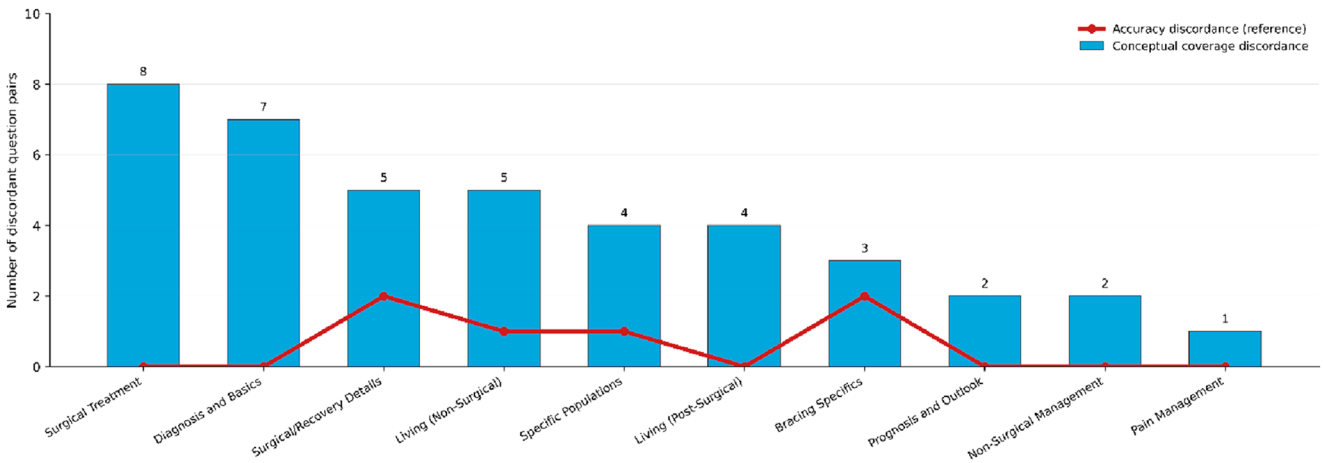


**Fig. 4** Where do the models diverge? Conceptual coverage discordance by category (accuracy shown as reference)

**Table 7** Contingency analysis of paired conceptual coverage assessments

| | GPT 5.2: adequate | GPT 5.2: inadequate | Row total |
|---|---|---|---|
| **Claude sonnet 4.5: adequate** | 17 | 29 | 46 |
| **Claude sonnet 4.5: inadequate** | 12 | 42 | 54 |
| **Column total** | 29 | 71 | 100 |

McNemar's test, p = 0.012; Total agreement = 59%

conservative scoliosis treatment, reported that while ChatGPT-4.0 responses were generally clear (median clarity score = 6/6), content validity varied considerably across topics, with experts identifying factual inaccuracies in responses to questions such as "What is the best sport for scoliosis?". However, unlike global quality ratings used previously, our tripartite framework isolates a foundational problem: correct information is often presented in language inaccessible to laypersons. This "accuracy-clarity paradox" is not unique to AIS; similar deficits have been noted in spine surgery patient education [7]. The post-hoc readability analysis confirmed that responses rated "Clear" averaged a Flesch-Kincaid Grade Level of 7.2, versus 11.6 for "Not Clear" responses ($p < 0.001$). The benchmark of ≤8th grade level is widely recommended for health materials, yet two-thirds of LLM responses exceeded this threshold [21]. This readability deficit is not unique to our study; Lieu et al. [22] reported that ChatGPT responses to scoliosis FAQs required reading levels ranging from 11th grade to college graduate, further demonstrating that current LLMs fail to meet health literacy standards for patient education.

Importantly, FKGL measures only syntactic complexity. Some responses with simple sentence structure but unexplained medical jargon (e.g., "Cobb angle," "Risser sign," "pulmonary function") remained inaccessible. True clarity requires not only short sentences but also active definition and contextualisation of technical terms. Current LLMs rarely provide such explanations, instead assuming a level

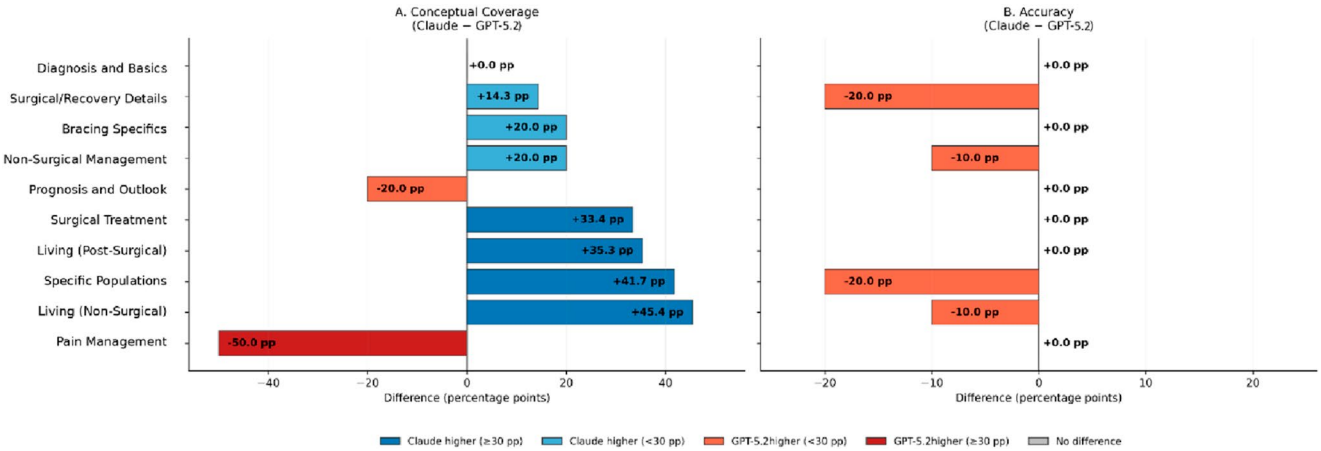


**Fig. 5** Category-based performance differences between models

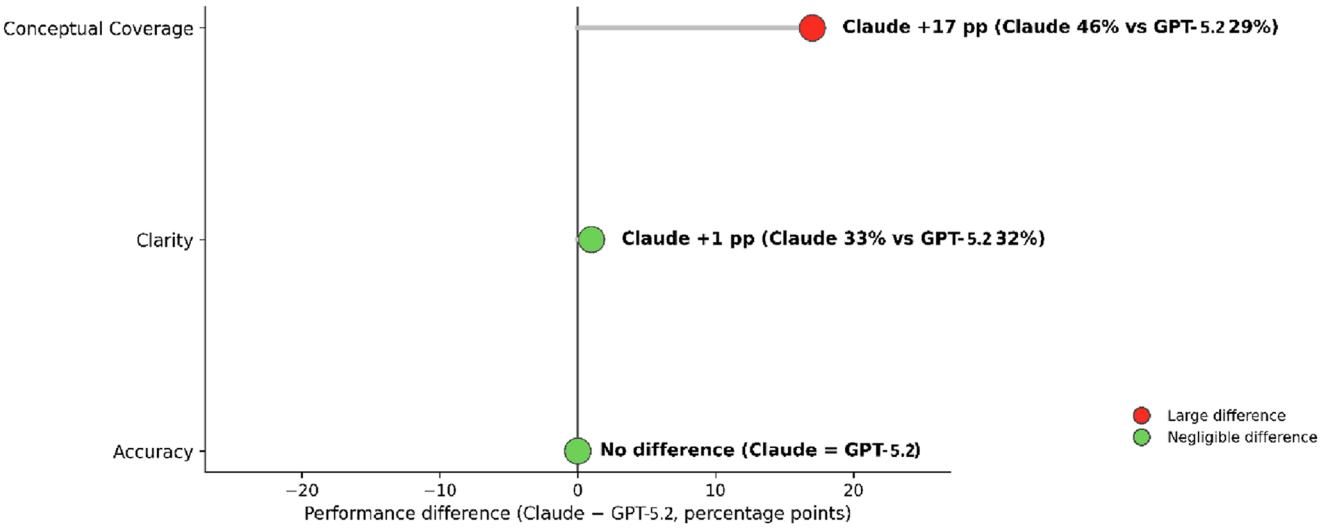


**Fig. 6** Model performance differences at a glance

of biomedical literacy that most patients do not possess. Thus, these tools may inadvertently perpetuate health literacy barriers rather than democratising information [18, 19].

The clinical implication is direct: a parent reading a technically correct but jargon-filled explanation of bracing compliance may not understand why nighttime wear is critical or what signs indicate poor fit. Comprehension is a prerequisite for informed consent and adherence [4]. Therefore, even perfect factual accuracy is insufficient without clarity. Future LLM applications for patient education must prioritise plain language generation, perhaps through targeted prompt engineering or fine-tuning on readability-optimised corpora.

## Conceptual coverage as a discriminating metric

The most pronounced difference between models was in conceptual coverage. Claude Sonnet 4.5 provided "Adequate Coverage" for 46% of questions versus 29% for GPT 5.2 ($p = 0.012$). This advantage was most evident in complex domains such as surgical treatment and recovery details, suggesting a greater capacity for synthesising multifaceted information. For example, when asked about recovery after spinal fusion, Claude Sonnet 4.5 typically included discussion of hospital stay duration, pain management strategies, return to school and sports, and potential long-term activity restrictions. GPT 5.2 often provided shorter, less detailed answers that omitted key contextual elements.

However, even the better-performing model frequently failed when questions required nuanced, individualised judgment (e.g., "Will my child's curve progress after growth stops?"). Such questions demand discussion of known prognostic factors (curve magnitude, skeletal maturity, Risser stage) and acknowledgment of individual variability, content that was often absent. The "slight" inter-model agreement ($\alpha = 0.130$) indicates that the two LLMs operate on fundamentally different internal benchmarks for answer completeness. This dissociation between preserved factual accuracy and deficient nuanced application mirrors findings from surgical decision-support contexts, where Kartal et al. [15] reported that LLMs achieved moderate agreement on binary surgical triage ($\kappa = 0.42$-$0.59$) but poor agreement on specific procedure selection ($\kappa = 0.15$-$0.25$).

The limited conceptual coverage we observed is consistent with other evaluations of LLMs for scoliosis education. Lieu et al. [22] found that ChatGPT responses to frequently asked questions were only "satisfactory requiring minimal to moderate clarification" (Mika score 2.4), indicating that responses, while not incorrect, lacked the depth and completeness expected of a reliable patient education resource. Similarly, Giray et al. [23] reported that 34.4% of responses were "correct but not comprehensive" and that only 26.3% of treatment and follow-up questions were answered correctly. These converging findings suggest that current LLMs consistently struggle to provide complete, contextually rich answers, particularly for complex management questions.

This finding advocates for routine inclusion of conceptual coverage in future LLM evaluations. Factual accuracy alone is insufficient for patient-centred communication; a response that correctly states a Cobb angle threshold but omits discussion of curve progression risk, observation intervals, or when to seek re-evaluation is technically correct but potentially misleading [20].

## Clinical implications

***Model selection matters.*** For verifying a discrete fact (e.g., diagnostic criteria), either model suffices. Comparative studies in surgical decision-making have similarly noted variability across LLMs, reinforcing that model choice can influence the quality of information delivered [24]. For gaining a contextualised understanding of complex management pathways (e.g., bracing vs. surgery), Claude Sonnet 4.5 may offer a more substantive foundation. However, this requires validation with actual patients, as expert judgment of "adequate coverage" may not perfectly align with what patients find useful.

***Mandatory clinician mediation remains non-negotiable.*** The pervasive clarity deficit and incomplete conceptual coverage mean LLM outputs cannot serve as standalone patient education tools. Healthcare professionals must interpret, simplify, and contextualise AI-generated information, adding patient-specific details and addressing emotional concerns. This conclusion aligns with prior spine surgery evaluations and with patient preferences for personal delivery of complex information [7, 25].

Multiple independent studies have reached similar conclusions. Negrini et al. [12] emphasised that "the involvement of expert clinicians remains essential" for conservative scoliosis care, while Kartal et al. [15] stated that "procedure selection should remain an expert-led task". Lieu et al. [22] concluded that ChatGPT responses, despite satisfactory accuracy, "should not be used as the sole source of medical information" for scoliosis patients. Extending these findings, Giray et al. [23] reported that only 26.3% of treatment and follow-up questions were answered correctly, explicitly advising that "patients and parents are advised to consult medical professionals rather than rely solely on AI-generated information for AIS treatment and management". Across four independent evaluations, the message is consistent: AI-generated information requires expert oversight.

The duty of informed consent cannot be delegated to an AI. Moreover, while patient education can improve condition recognition, its sensitivity remains inferior to

professional screening, further underscoring the irreplaceable role of expert clinical judgment [26].

***Informing educational resource design.*** The categories where both models underperformed (bracing specifics, long-term prognosis, surgical recovery) identify gaps in existing patient materials and training data. Clinicians and professional societies should strengthen resources in these areas [18, 27].

### Limitations

Several limitations should be acknowledged. First, data contamination cannot be ruled out, a limitation inherent to most LLM benchmarking studies as noted by Kartal et al. [15]. Our patient-derived question set, sourced from public forums and professional portals, may have overlapped with LLM training corpora, potentially inflating accuracy estimates. Second, this study used expert judgment rather than direct patient assessment of comprehension; while expert evaluation provides rigorous clinical scrutiny, it does not capture how actual patients understand and use AI-generated information. Third, our findings represent a temporal snapshot; LLM performance may change with model updates, and generalisability is limited to the two models tested (Claude Sonnet 4.5 and GPT 5.2). Fourth, the Flesch-Kincaid Grade Level captures syntactic complexity but not full conceptual accessibility; a response with simple sentence structure may still be inaccessible due to unexplained medical jargon. Finally, although our rater panel included experienced clinicians, it excluded patient and caregiver perspectives, which should be incorporated in future research.

### Future directions

a. Longitudinal tracking of LLM performance across versions.
b. Scale development for patient-centred assessment of clarity and coverage.
c. Specialised medical LLMs: The demonstrated gap supports developing models fine-tuned for health communication, including retrieval-augmented generation systems grounded in evidence-based knowledge bases (e.g., Med-PaLM 2 [28]).
d. End-user impact studies measuring comprehension, anxiety, and decision-making.
e. Multimodal integration (visual aids, diagrams) to enhance clarity for anatomical and procedural explanations.

## Conclusion

This multidimensional benchmark reveals that while contemporary LLMs contain robust factual knowledge about AIS, they fall short as standalone patient education resources due to systemic clarity deficits and considerable variability in conceptual coverage. Expert judgment must be validated through direct patient evaluation before conclusions about real-world utility can be drawn. LLMs hold promise as adjuncts to clinical knowledge resources, but their integration into patient care requires judicious caution. Effective and ethical use necessitates mandatory mediation by healthcare professionals, who must vet, interpret, and contextualise AI-generated information to ensure it is accurate, clear, and personally relevant. The path forward lies not in uncritical adoption, but in informed application guided by rigorous understanding of these tools' capabilities and limitations.

**Author contributions** 1. CA, ABK, AMO, FG, MAO: Conceptualization, writing original draft, visualization. 2. CA, ABK, AMO: Formal analysis, data curation, methodology. 3. FG, AMO, MAO: Supervision, project administration, validation. 4. All Authors: Writing-review-editing, and approval of the final manuscript for submission.

**Funding** Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). This research received no specific grant or financial support from funding agencies in the public, commercial, or not-for-profit sectors.

**Data availability** The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

### Declarations

**Ethics approval and consent to participate** This study involved the expert evaluation of AI-generated responses to standardized, non-identifiable patient questions and did not include human participants, patient data, or clinical interventions. In accordance with institutional policy, ethical approval was obtained from the relevant Institutional Ethics Committee (approval number: 26-6T/20). All study procedures were designed in compliance with the ethical principles of the Declaration of Helsinki and its subsequent amendments.

**Informed consent** Not applicable. The study did not involve direct participation of human subjects, patient interaction, or the use of identifiable personal data.

**Competing interests** The authors declare no competing interests.

## References


1. Cheng JC, Castelein RM, Chu WC, Danielsson AJ, Dobbs MB, Grivas TB et al (2015) Adolescent idiopathic scoliosis. Nat Rev Dis Prim 1:15030. https://doi.org/10.1038/nrdp.2015.30
2. Addai D, Zarkos J, Bowey AJ (2020) Current concepts in the diagnosis and management of adolescent idiopathic scoliosis. Childs Nerv Syst 36:1111–1119. https://doi.org/10.1007/s00381-020-04608-4
3. Wang H, Li T, Yuan W, Zhang Z, Wei J, Qiu G et al (2019) Mental health of patients with adolescent idiopathic scoliosis and their parents in China: a cross-sectional survey. BMC Psychiatry 19:147. https://doi.org/10.1186/s12888-019-2128-1
4. Lyu X, Li J, Li S (2024) Approaches to Reach Trustworthy Patient Education: A Narrative Review. Healthcare 12:2322. https://doi.org/10.3390/healthcare12232322
5. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA et al (2023) Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res 25:e47479. https://doi.org/10.2196/47479
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB et al (2023) Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med 183:589. https://doi.org/10.1001/jamainternmed.2023.1838
7. Stroop A, Stroop T, Zawy Alsofy S, Nakamura M, Möllmann F, Greiner C et al (2024) Large language models: Are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? Eur Spine J 33:4135–4143. https://doi.org/10.1007/s00586-023-07975-z
8. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A et al (2023) The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model. MedRxiv Prepr Serv Heal Sci. https://doi.org/10.1101/2023.01.30.23285067
9. Yaş S, Yapar D, Yapar A, Özel T, Tokgöz MA, Baymurat AC et al (2025) Assessing the role of large language models in adolescent idiopathic scoliosis care: a comparison between ChatGPT and Google Gemini. Acta Orthop Traumatol Turc 59:222–9. https://doi.org/10.5152/j.aott.2025.25279
10. Lang S, Vitale J, Galbusera F, Fekete T, Boissiere L, Charles YP et al (2025) Is the information provided by large language models valid in educating patients about adolescent idiopathic scoliosis? An evaluation of content, clarity, and empathy. Spine Deform 13:361–372. https://doi.org/10.1007/s43390-024-00955-3
11. Suresh A, Siahaan J, Marco RA, Klineberg E, Borden T, Vanodia R et al (2025) Comparing the effectiveness of generative AI technology in commonly asked scoliosis questions. J Child Orthop 19:416–21. https://doi.org/10.1177/18632521251359098
12. Negrini F, Malfitano C, Ferriero G, Morone G, Negrini A, Zaina F et al (2026) Evaluating ChatGPT-4.0's accuracy and potential in idiopathic scoliosis conservative treatment: a preliminary study on clarity, validity, and expert perceptions. Eur Spine J 35:422–30. https://doi.org/10.1007/s00586-025-09166-4
13. Fabijan A, Zawadzka-Fabijan A, Fabijan R, Zakrzewski K, Nowosławska E, Polis B (2024) Assessing the accuracy of artificial intelligence models in scoliosis classification and suggested therapeutic approaches. J Clin Med 13:4013. https://doi.org/10.3390/jcm13144013
14. Almekkawi AK, Caruso JP, Anand S, Hawkins AM, Rauf R, Al-Shaikhli M et al (2025) Comparative analysis of large language models and spine surgeons in surgical decision-making and radiological assessment for spine pathologies. World Neurosurg 194:123531. https://doi.org/10.1016/j.wneu.2024.11.114
15. Kartal A, Manalil NF, Cheng CD, Chung LK, Gebhard H, Greenberg M et al (2025) Evaluating Large Language Models for Decision Support in Minimally Invasive Spine Surgery Triage and Procedural Categories. Glob Spine J. https://doi.org/10.1177/21925682251411225
16. Clarke MA, Moore JL, Steege LM, Koopman RJ, Belden JL, Canfield SM et al (2016) Health information needs, sources, and barriers of primary care patients to achieve patient-centered care: a literature review. Health Informatics J 22:992–1016. https://doi.org/10.1177/1460458215602939
17. Wellburn S, van Schaik P, Bettany-Saltikov J (2019) The information needs of adolescent idiopathic scoliosis patients and their parents in the UK: an online survey. Healthcare 7:78. https://doi.org/10.3390/healthcare7020078
18. Eltorai AEM, Sharma P, Wang J, Daniels AH (2015) Most American Academy of Orthopaedic Surgeons' Online Patient Education Material Exceeds Average Patient Reading Level. Clin Orthop Relat Res 473:1181–1186. https://doi.org/10.1007/s11999-014-4071-2
19. Bryson X, Albarran M, Pham N, Salunga A, Johnson T, Hogue GD et al (2025) Artificial intelligence-based large language models can facilitate patient education. J Pediatr Orthop Soc North Am 12:100196. https://doi.org/10.1016/j.jposna.2025.100196
20. Kersting C, Kneer M, Barzel A (2020) Patient-relevant outcomes: what are we talking about? A scoping review to improve conceptual clarity. BMC Health Serv Res 20:596. https://doi.org/10.1186/s12913-020-05442-9
21. Cotugna N, Vickery CE, Carpenter-Haefele KM (2005) Evaluation of literacy level of patient education pages in health-related journals. J Community Health 30:213–219. https://doi.org/10.1007/s10900-004-1959-x
22. Lieu B, Crawford E, Laubach L, Yeramosu T, Sharps C, Horstmann J et al (2025) Patient education strategies in pediatric orthopaedics: using ChatGPT to answer frequently asked questions on scoliosis. Spine Deform 13:1377–89. https://doi.org/10.1007/s43390-025-01087-y
23. Giray E, Korkmaz MD, Illeez OG, Capan N, Karadag Saygi E, Aydin AR (2025) Let's chat (GPT) about adolescent idiopathic scoliosis: accuracy and reliability of chat responses to frequently asked questions. BMC Musculoskelet Disord 26:1075. https://doi.org/10.1186/s12891-025-09315-2
24. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL et al (2023) Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards preparation question bank. Neurosurgery 93:1090–8. https://doi.org/10.1227/neu.0000000000002551
25. Chan CYW, Chong JSL, Lee SY, Ch'ng PY, Chung WH, Chiu CK et al (2020) Parents'/patients' perception of the informed consent process and surgeons accountability in corrective surgery for adolescent idiopathic scoliosis. Spine (Phila Pa 1976) 45:1661–7. https://doi.org/10.1097/BRS.0000000000003641

26. de Groot C, Heemskerk JL, Willigenburg NW, Altena MC, Kempen DHR (2022) Educating parents improves their ability to recognize adolescent idiopathic scoliosis: a diagnostic accuracy study. Children Basel 9. https://doi.org/10.3390/children9040563
27. Kirchner GJ, Kim RY, Weddle JB, Bible JE (2023) Can Artificial Intelligence Improve the Readability of Patient Education Materials? Clin Orthop Relat Res 481:2260–2267. https://doi.org/10.1097/CORR.0000000000002668
28. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M et al (2025) Toward expert-level medical question answering with large language models. Nat Med 31:943–50. https://doi.org/10.1038/s41591-024-03423-7