Original Paper

Assessing the Quality of AI Responses to Patient Concerns About Axial Spondyloarthritis: Delphi-Based Evaluation

Jiaxin Bai[1,2*], MM; Xiaojian Ji[2*], MD; Jiali Yu[1,2*], MM; Yiwen Wang[2], MD; Yufei Guo[1,2], MM; Chao Xue[1,2], MM; Wenrui Zhang[1,2], MM; Jian Zhu[2,3], Prof Dr

[1]Medical School of Chinese People's Liberation Army, Beijing, China

[2]Department of Rheumatology and Immunology, The First Medical Center, Chinese People's Liberation Army General Hospital, Beijing, China

[3]State Key Laboratory of Kidney Diseases, Chinese People's Liberation Army General Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Jian Zhu, Prof Dr
Department of Rheumatology and Immunology
The First Medical Center
Chinese People's Liberation Army General Hospital
28 Fuxing Road
Beijing, 100036
China
Phone: 86 010 55499314
Email: jian_jzhu@126.com

Abstract

Background: Axial spondyloarthritis (axSpA) is a chronic autoinflammatory disease with heterogeneous clinical features, presenting considerable complexity for sustained patient self-management. Although the use of large language models (LLMs) in health care is rapidly expanding, there has been no rigorous assessment of their capacity to provide axSpA-specific health guidance.

Objective: This study aimed to develop a patient-centered needs assessment tool and conduct a systematic evaluation of the quality of LLM-generated health advice for patients with axSpA.

Methods: A 2-round Delphi consensus process guided the design of the questionnaire, which was subsequently administered to 84 patients with axSpA and 26 rheumatologists. Patient-identified key concerns were formulated and input into 5 LLM platforms (GPT-4.0, DeepSeek R1, Hunyuan T1, Kimi k1.5, and Wenxin X1), with all prompts and model outputs in Chinese. Responses were evaluated using 2 techniques: an accuracy assessment based on guideline concordance, with independent double blinding by 2 raters (interrater reliability analyzed via Cohen κ), and the AlphaReadabilityChinese analytic tool to assess readability.

Results: Analysis of the validated questionnaire revealed age-related differences. Patients younger than 40 years prioritized symptom management and medication side effects more than those older than 40 years. Distinct priorities between clinicians and patients were identified for diagnostic mimics and drug mechanisms. LLM accuracy was highest in the diagnosis and examination category (mean score 20.4, SD 0.9) but lower in treatment and medication domains (mean score 19.3, SD 1.7). GPT-4.0 and Kimi k1.5 demonstrated superior overall readability; safety remained generally high (disclaimer rates: GPT-4.0 and DeepSeek-R1 100%; Kimi k1.5 88%).

Conclusions: Needs assessment across age groups and observed divergences between clinicians and patients underline the necessity for customized patient education. LLMs performed robustly on most evaluation metrics, and GPT-4.0 achieved 94% overall agreement with clinical guidelines. These tools hold promise as scalable adjuncts for ongoing axSpA support, provided complex clinical decision-making remains under human oversight. Nevertheless, the prevalence of artificial intelligence hallucinations remains a critical barrier. Only through comprehensive mitigation of such risks can LLM-based medical support be safely accelerated.

(JMIR AI 2026;5:e79153) doi: 10.2196/79153

KEYWORDS

axial spondyloarthritis; axSpA; artificial intelligence; AI; large language model; health management; chronic disease

Introduction

Axial spondyloarthritis (axSpA) is a chronic inflammatory disorder that predominantly affects the sacroiliac and axial spinal joints. Early symptoms often include chronic atypical low back pain and morning stiffness, with associated manifestations such as tendinitis and arthritis and extra-articular features such as uveitis, inflammatory bowel disease, and psoriasis frequently observed [1]. Despite substantial research progress on axSpA, most studies have been disease centered, with limited focus on patient-oriented assessment. The insidious onset and nonspecific symptoms frequently contribute to delays in recognition and care. Accurate diagnosis requires the integration of clinical signs; laboratory results; and imaging, such as pelvic X-ray or sacroiliac joint magnetic resonance imaging [2]. Many patients lack a clear understanding of the necessity or implications of these examinations. Therapeutic approaches for axSpA encompass both pharmacological and nonpharmacological strategies [3,4], posing additional challenges regarding patient decision-making and informed participation in care. These factors collectively impact axSpA self-management and highlight the urgent need for enhanced patient education. Furthermore, the rapid advancement of large language models (LLMs) has unlocked considerable health care potential [5,6]. As more patients seek advice from artificial intelligence (AI)–based systems, it remains essential to rigorously evaluate the accuracy and quality of medical guidance they provide within axSpA-related contexts.

This study aimed to systematically identify genuine concerns of patients with axSpA via a questionnaire survey and a parallel analysis of the perspectives from clinicians. Patient-derived questions were presented to LLMs, with resulting health advice assessed across 3 dimensions: readability, accuracy, and health disclaimer. These findings offer data-driven insight for clinicians, enabling them to tailor education to the needs and cognitive patterns of diverse patient populations. The results further inform evaluation of LLMs in health counseling, support more nuanced clinical decision-making in diagnosis and treatment, and guide the development of sustainable patient-centered management strategies.

Methods

Construction of the Questionnaire

The questionnaire development comprised 3 stages [7,8]. Initially, a comprehensive list of knowledge items was extracted from published questionnaires and the 2022 Assessment of Spondyloarthritis International Society–European Alliance of Associations for Rheumatology recommendations for axSpA management. A Delphi process included rheumatologists, rheumatology graduate students, and patients. They first enriched the list by adding items considered potentially useful, and then the list was reduced to obtain the most important items. Participants in the Delphi rounds were enrolled from the department of rheumatology and immunology of the Chinese PLA General Hospital First Medical Center. The rheumatologists and the rheumatology graduate students invited patients to participate.

In the second stage, the initial version of the questionnaire was created based on the first Delphi round results, formulated by XJ, JB, and JY. Each question was mapped to the extracted item list to ensure comprehensive coverage of clinical features, diagnosis, examination methods, medication options, and prognosis related to axSpA. The instrument was designed for all patients with axSpA features regardless of concomitant peripheral SpA, psoriasis, or inflammatory bowel disease manifestations.

In the third stage, the final Delphi round facilitated consensus among all rheumatology experts and rheumatology graduate students to refine the instrument, with questions selected as essential if chosen by more than two-thirds and useful if chosen by more than half but less than two-thirds of participants. Items deemed redundant and overly complex or those lacking clinical relevance were eliminated, resulting in the finalized version. The questionnaire structure and corresponding item numbers are provided in Multimedia Appendix 1.

Data Collection and Analysis

For data collection, the finalized questionnaire was digitized and formatted into an online survey. An additional section at its conclusion collected basic demographic and health-related information to support baseline analysis. Participation was anonymous, with clear disclosure that responses would be used solely for research purposes. Recruitment used a Wenjuanxing (an online survey platform) link, and this link was distributed through hospital outpatient clinics [9]. The collected data were categorized and contrasted according to the baseline characteristics of the respondents, including patient age, sex, and occupational category.

To compare differences in attitudes between health care professionals and patients, a separate online survey was administered to medical staff within the rheumatology and immunology department.

Choice of LLM Chatbots

In selecting LLMs, we included DeepSeek R1 (DeepSeek), Hunyuan T1 (Tencent), Kimi k1.5 (Moonshot AI), Wenxin X1 (Baidu), and GPT-4.0 (OpenAI) [10-13], each possessing strengths in different domains. The comprehensive comparison of these models was intended to more accurately reflect real-world choices and user experiences among patients with axSpA.

Outcomes and Data Synthesis

The LLM-generated answers were systematically collected by a researcher and organized into bullet points. Each question was submitted independently to the models in a 1-time format to prevent AI memory effects and ensure unbiased responses. Both the patient queries and all LLM outputs were generated in Chinese. Full datasets are provided in Multimedia Appendix 2. Response assessment targeted 3 metrics: accuracy, readability, and health advice disclaimers. Accuracy was defined as the degree of correctness in each LLM's response to individual items [6-14] benchmarked against the 2022 Assessment of Spondyloarthritis International Society–European Alliance of Associations for Rheumatology guidelines and the Lancet series

recommendations [4,15-19]. Two independent raters assessed each suggestion based on a published scoring criterion (Multimedia Appendix 3), with arbitration by a third researcher in case of discrepancies. For example, for scoring, if rater A assigned indicator scores of 4, 3, 3, and 1 and rater B assigned scores of 4, 4, 3, and 1, the raters would discuss any discrepancies (here for the second indicator, 3 vs 4). Irreconcilable differences were resolved by an expert's decision. The independent raters acknowledged potential subjective bias favoring AI, possibly leading to higher average ratings than seen in previous literature. Interrater reliability was quantified via the Cohen κ statistic.

Readability was defined as the ease or difficulty of reading each text and quantitatively measured using the AlphaReadabilityChinese tool (Shanghai International Studies University) [20]. This analytic framework assesses 9 dimensions of language complexity. Higher scores in some dimensions signal increased reading difficulty, whereas, for the 5 "precision and clarity" dimensions, higher scores equate to better comprehension (Textbox 1).

The key takeaway was that easier-to-understand texts scored low on dimensions of complexity, such as intricate vocabulary and sentence structure, but high on dimensions of precision and clarity, including the use of specific words and unambiguous phrasing.

"Health disclaimers" were defined as warnings within the response that cautioned about specific risks or promoted appropriate and safe patient behaviors, such as recommending medical attention if symptoms persist. Each LLM response was categorized on the basis of the presence or absence of a health disclaimer [21]. The scope of disclaimers encompassed recommendations to seek professional assistance, urgent care, careful medication use, and general consultative language.

Textbox 1. Dimensions of readability.

Dimensions where higher scores mean the text is harder to read

- Lexical richness indicates the use of diverse and complex vocabulary.
- Syntactic richness refers to longer and structurally intricate sentences.
- Semantic richness reflects a high density of content and information.
- Semantic noise represents the presence of redundant or off-topic information that may obscure the main message.

Dimensions where higher scores mean the text is easier to read

- Noun or verb precision captures the use of specific nouns and action verbs (eg, "MRI scan" instead of "a type of examination" and "reduce pain" instead of "implement analgesic measures").
- Semantic clarity measures how directly and unambiguously information is conveyed.

Statistical Analysis

Statistical analyses were conducted using R (version 3.4.0; R Foundation for Statistical Computing) and RStudio (version 1.0.136; Posit PBC). Assumptions of normality and variance homogeneity informed the use of either ANOVA or Kruskal-Wallis tests for multiple group comparisons of language-difficulty metrics [22,23]; Greenhouse-Geisser or Satterthwaite corrections were applied as needed [24,25]. Categorical data from questionnaire responses were evaluated using chi-square tests or Fisher exact test, where applicable [26,27]. Significance was defined at $P<.05$. Figures were plotted using the *ggplot2* R package.

Ethical Considerations

Before the first Delphi round, this study was approved by the medical ethics committee of Chinese People's Liberation Army General Hospital (S2022-255-03). For patients completing the paper-based questionnaire, a dedicated informed consent form was signed to obtain their consent. For those completing the electronic questionnaire, informed consent was obtained through the "check + click button" method—patients were required to check the box and click the confirmation button to verify that they had read and agreed to all terms. During the data collection process, we ensured patient privacy and maintained strict confidentiality of patient data. No compensation was provided to patients for their participation.

Results

Construction of the Questionnaire

At the first stage, 31 items were extracted from existing survey instruments. Delphi rounds incorporated 1 senior rheumatology expert with more than 30 years of experience, 3 rheumatologists with extensive clinical expertise, 5 rheumatology graduate students, and 8 patients. The first Delphi round expanded the preliminary list to 50 potentially informative items. In the next stage, a graduate student reformulated these into specific questions and compiled them into a draft questionnaire. The final Delphi round selected 42 questions judged "essential" by more than half (9/17, 53%) of the participants. Figure S1 in Multimedia Appendix 4 provides a detailed flowchart of these procedures.

Survey Results

Through the online questionnaire, responses were collected from 84 patients with axSpA. Demographic details and response distributions are presented in Figure 1A and Table 1. The cohort comprised 62 (74%) men and 22 (26%) women, with an average age of 38.01 (SD 10.45) years. Education levels were predominantly bachelor's degree (n=34, 40%), followed by

senior high school (n=24, 29%) and master's or higher degrees (n=13, 15%). Most (n=47, 56%) held sedentary occupations. Parental health status was most often reported as "good" (n=57, 68%), while self-assessed health was frequently rated as "fair" (n=42, 50%). Family history of ankylosing spondylitis was identified in 27 (32%) participants. In total, 57 (68%) participants used the internet for less than 6 hours a day, and 27 (32%) participants exceeded this threshold. Figure 1A shows that question 11 ("My doctor recommended testing for HLA-B27. What does a positive result mean?") was the area of greatest concern. To expand the scope of assessment, 26 responses from health care professionals were gathered (Figure 1B), with question 11 also ranking highly in this group. Health care professionals unanimously identified question 1, question 3, question 14, and question 24 as highly important, with no respondents rating them as "neutral," "unimportant," or "very unimportant."

To explore factors influencing patient prioritization, we compared responses across patient subgroups based on baseline characteristics. The results indicated age was the most significant variable (*P* values ranging from .001 to .05), with 12 questions showing statistically significant age-based differences (question 4, question 13, question 17, question 24, question 27, question 28, question 30, question 31, question 36, question 37, question 38, and question 40; refer to Figures 2A and B. Multimedia Appendix 5 for *P* values). Cross-group analysis of patient versus health care worker priorities revealed statistically significant disparities on 3 questions (question 18, question 26, and question 31; refer to Figures 3A and B. Multimedia Appendix 6 for *P* values).

Figure 1. Questionnaire responses from patients and rheumatologists. (A) Patient questionnaire responses. The lengths of the differently colored bars represent the proportion of respondents who selected each option within the total surveyed population. (B) Rheumatologists' questionnaire responses.

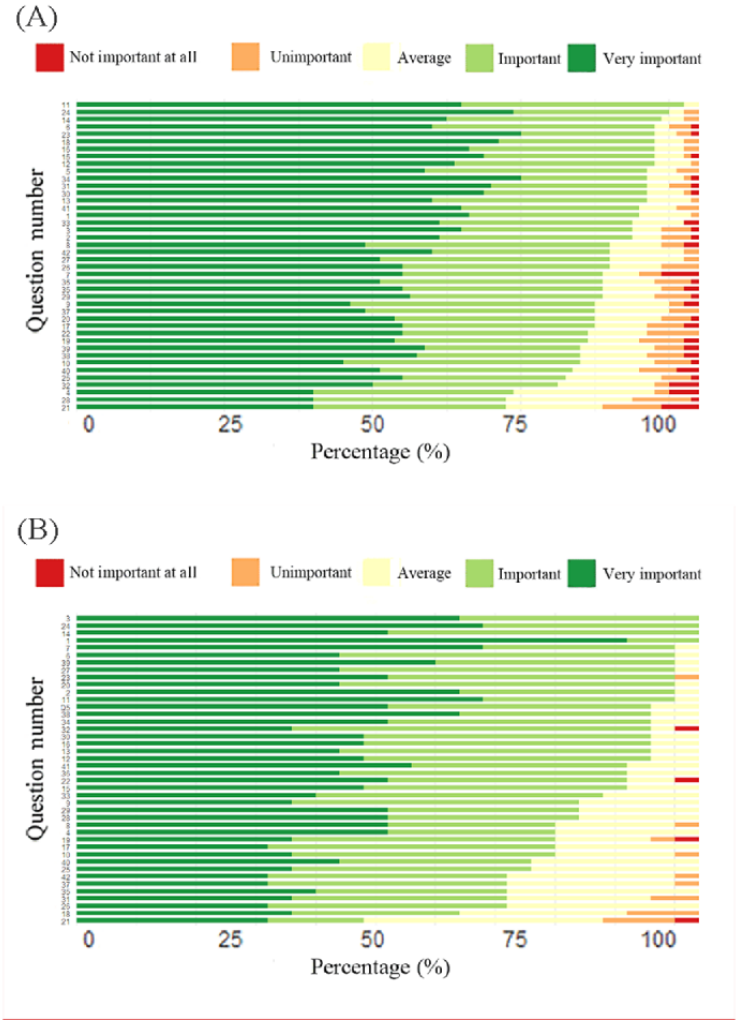


Table 1. Baseline characteristics of the study population (N=84).

Characteristic	Values
Sex, n (%)	
Male	62 (74)
Female	22 (26)
Age (y), mean (SD)	38.01 (10.45)
Education level, n (%)	
Primary school or below	3 (4)
Junior high school	10 (12)
Senior high school	24 (29)
Bachelor’s degree	34 (40)
Master’s degree or above	13 (15)
Sedentary occupation, n (%)	
Yes	47 (56)
No	37 (44)
Parental health status, n (%)	
Good	57 (68)
Fair	23 (27)
Poor	4 (5)
Personal health status, n (%)	
Good	33 (39)
Fair	42 (50)
Poor	9 (11)
Family history of axial spondyloarthritis, n (%)	
Yes	27 (32)
No	57 (68)
Family history of hereditary diseases, n (%)	
Yes	19 (23)
No	65 (77)
Daily internet use duration (h), n (%)	
<6	57 (68)
>6	27 (32)

Figure 2. Age-stratified response discrepancy distribution. (A) Scatter points below the red dashed line indicate $P<.05$, suggesting statistically significant differences in answer choices among different age groups for the specific question. (B) Each color block represents the proportion of respondents who selected that option relative to the total. Group 1 was composed of patients older than 40 years, and group 2 was composed of patients younger than 40 years.

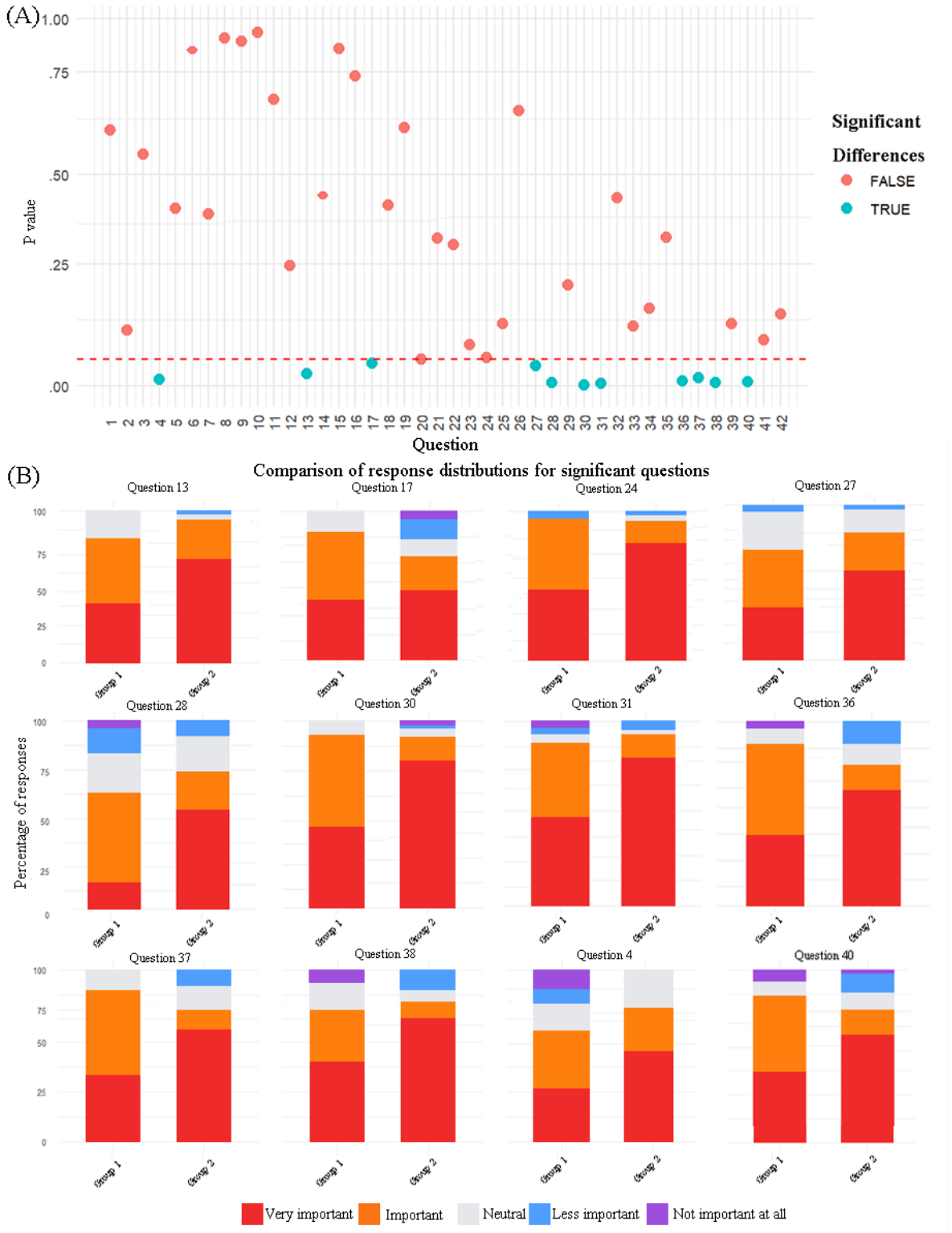
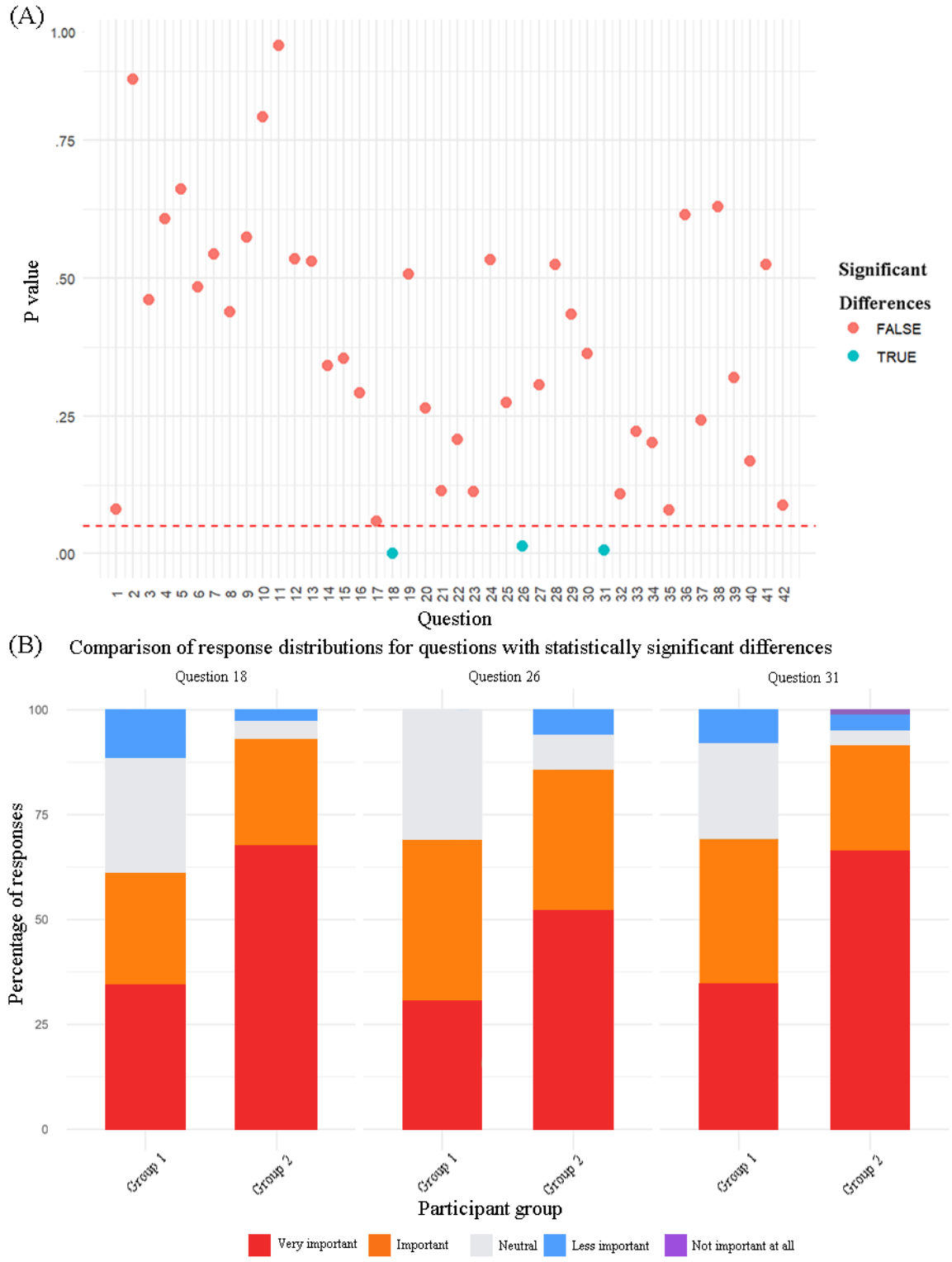


Figure 3. Distribution of response differences between rheumatologists and patients. (A) Scatter points below the red dashed line indicate $P<.05$, suggesting statistically significant differences in answer selection between medical staff and patients for the specific question. (B) Each color block represents the proportion of respondents who selected that option relative to the total. Group 1 was composed of health care professionals, and group 2 was composed of patients.



AI Consultation Opinion Quality Assessment

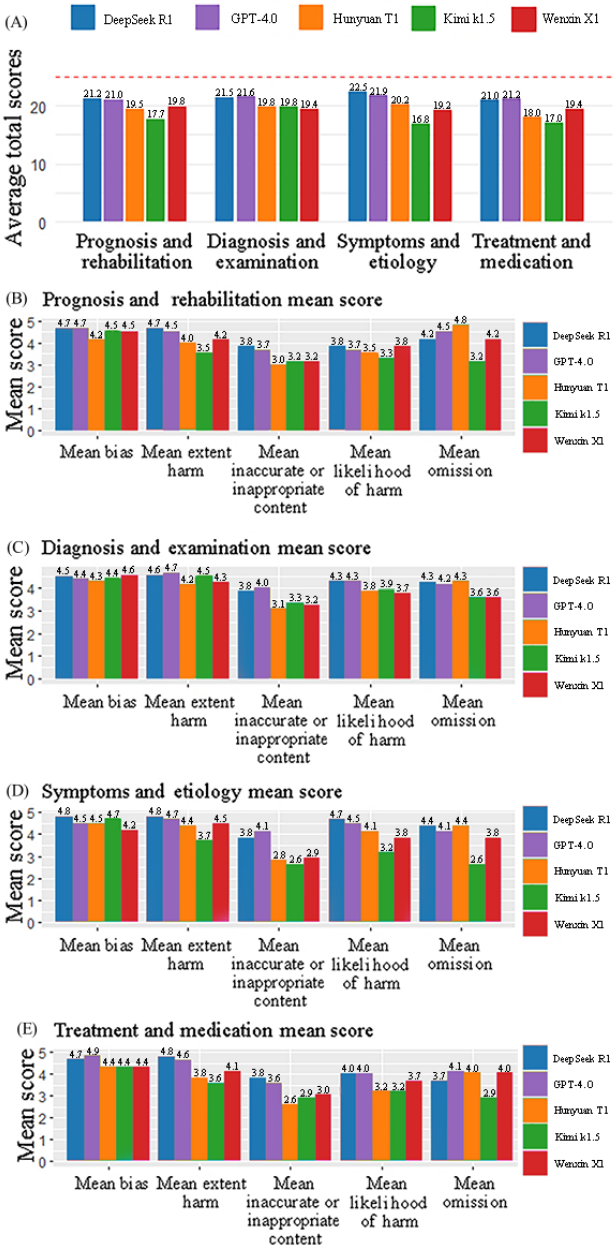
Overview

The 42 patient-derived questions were submitted to all 5 selected LLMs, each generating independent responses to avoid memory bias. Outputs were collected and systematically aggregated into bullet point summaries reflecting health consultation content. Three core attributes—readability, accuracy, and incorporation of health disclaimers—were then assessed for each model's output.

Accuracy

The 5 LLMs generated 1052 recommendations for the 42 items, including repeated suggestions for the same question across models. Interrater reliability was excellent (Cohen κ=0.947; Figure S2 Multimedia Appendix 4). The diagnosis and examination category yielded the highest average accuracy across models (mean score 20.4, SD 0.9), while the treatment and medication domain scored lowest (mean score 19.3, SD 1.7). Model-specific performance data across domains and question items are provided in Figure 4A; additional breakdowns are detailed in Figures 4B-E; Multimedia Appendix 7 presents complete values. Comparative analysis highlighted that the LLMs' lowest scores consistently occurred in the "inaccurate or inappropriate content" category, indicating vulnerability to these errors. In contrast, the highest average scores were in the "bias," suggesting a strong model's ability to avoid bias in health consultation outputs. Overall, model performance was satisfactory, with total accuracy scores ranging from 16.8 to 22.5. The highest scoring questions spanned all domains (question 3: 23.4 points, question 11: 23.2 points, question 38: 18.2 points, and question 40: 22.4 points), while the lowest scores were concentrated in questions involving nuanced or controversial information (question 6: 17.6 points, question 20: 16.4 points, question 34: 16.6 points, and question 38: 18.2 points).

Figure 4. Overall and module-specific score charts. (A) Overall score. (B-E) Scores by module. DS: DeepSeek R1; GPT: GPT-4.0; HY: Hunyuan T1; KM: Kimi k1.5; WX: Wenxin X1.



Readability

The readability of LLM-generated health consultation responses was measured using the AlphaReadabilityChinese tool. Comparative analysis of the 5 LLMs' outputs, as visualized via a heat map in Figure S3 in Multimedia Appendix 4 and detailed in Multimedia Appendix 8, revealed no significant model differences in noun-verb or content-word semantic precision. Kimi k1.5 excelled in lexical richness, verb accuracy, and semantic noise, while GPT-4.0 demonstrated superior syntactic richness, noun accuracy, semantic richness, and semantic clarity. DeepSeek R1, Hunyuan T1, and Wenxin X1 exhibited similar readability performance overall.

Disclaimers About Health Advice

Figure S4 in Multimedia Appendix 4 demonstrates that most LLM outputs contained health advice disclaimers, with GPT 4.0 and DeepSeek R1 including such disclaimers in responses to all 42 questions. Kimi k1.5 provided the fewest responses but still included disclaimers in 37 (88%) of the 42 cases.

Discussion

This study directly addressed real-world concerns of patients with axSpA by fostering collaboration between rheumatologists and patients to develop a comprehensive questionnaire encompassing symptoms, diagnosis, treatment, and prognosis. Subsequent validation with an 84-patient sample demonstrated that the tool reliably reflects patient-identified uncertainties and supports health care professionals in identifying prioritized and neglected issues. This facilitates the creation of targeted educational programs to enhance long-term chronic disease management.

However, marked discrepancies emerged between professionals and patients in the perceived importance of certain topics. For instance, question 18 ("What diseases is this condition likely to be misdiagnosed as?") was rated more highly by patients than by clinicians [28,29]. Question 31 ("Do biologic agents carry addiction potential?") and question 26 ("What are the mechanistic differences between NSAIDs, corticosteroids, and analgesics in pain management?") also showed such divergence [30]. These differences may reflect gaps in professional knowledge transfer, whereby clinicians, familiar with drug mechanisms and risk profiles, may underestimate the informational value these issues hold for patients. This knowledge gap highlights potential inadequacies in current educational practices and underscores the need for efforts to bridge understanding between clinicians and patients in future interventions.

Age is a significant driver of patient perception [31]. Analysis of patients grouped by age (older or younger than 40 years) revealed 12 questions with statistically significant differences, particularly related to symptom management, medication side effects, and prognosis. Younger patients showed increased concern, whereas no significant differences in baseline demographic characteristics were detected (Multimedia Appendix 9). Two main explanations were identified: first, younger patients showed greater interest in novel biological agents and their related mechanisms or risks; second, life stage difference shaped priorities, with patients younger than 40 years demonstrating greater family-planning awareness and early diagnoses mitigating confusion over questions such as question 17. Furthermore, considering axSpA often manifests in early adulthood, older patients, who have lived with the disease for longer, may be more accustomed to standard interventions and less reliant on new information [32]. Collectively, these findings highlight the necessity for age-specific patient education to reflect diverse literacy and life stage requirements, with future health promotion strategies tailored accordingly [33].

A persistent problem observed was AI hallucination, in which LLMs produced confidently stated yet unsourced or inaccurate statistics. For example, in question 41, Hunyuan T1 claimed, "Spinal mobility: 30 minutes of daily yoga can increase the maintenance rate of spinal range of motion by 55% [5-year follow-up data]." While evidence does support mobility benefits of yoga in axSpA through mechanistic pathways, such as muscle strengthening or inflammation reduction, no research corroborates a 55% improvement rate or the alleged 5-year dataset [34]. Although LLMs demonstrated generally strong performance, the safety risk posed by confidently delivered but unfounded claims remains substantial, a threat that cannot be ignored if patients act on these unsubstantiated data. Teaching patients to appraise such claims critically is vital for maximizing LLMs' potential to support chronic disease management while safeguarding patient health [35].

Despite intermodel variability in accuracy for medical advice [36], the LLMs overall performed robustly in this study. Accuracy ratings in this study were higher compared to previous research, which may be attributable to our open-ended, patient-focused question format and relatively accommodating scoring criteria [37,38]. Ongoing advances in AI technology may also explain this improvement. Notably, the "bias" consistently produced high scores, reflecting a strong capacity to provide wide-ranging yet balanced recommendations. However, the inclination for models to sometimes produce superficially authoritative yet insufficiently substantiated advice, especially regarding clinical management, introduces significant risk. For example, in response to glucocorticoid-related queries (question 35), Wenxin X1 recommended glucocorticoids for pain management without thorough context, potentially exposing patients to avoidable complications, including osteoporosis and serious infections [39,40]. These instances typically resulted in lower "inaccurate or inappropriate content" scores.

Our findings showed that high-scoring LLM responses generally addressed well-established topics with strong supporting evidence. As seen in responses to question 40 ("Can Traditional Chinese Medicine [TCM] treatments replace Western pharmacological therapies?"), all models consistently advised against substituting traditional Chinese medicine (TCM) for Western medicine. GPT-4.0's response indicated that TCM currently lacks conclusive evidence comparable to that of Western medicine in key efficacy outcomes such as bone protection and symptom control [41,42]. It further clarified that while TCM can serve as an effective adjunctive therapy, Western medicine should remain the foundational treatment approach. Although TCM or acupuncture may serve as useful adjuncts in the management of ankylosing spondylitis, they

cannot yet replace the central role of Western medications. We recommend that one works with a specialist to build an integrated, individualized treatment plan that is grounded in Western medicine and supplemented by TCM modalities.

Conversely, lower-scoring questions were primarily those related to medication recommendations. Medication management is highly individualized, requiring customized clinical judgment based on expertise and a comprehensive understanding of the patient's profile [36,43,44]. Authoritative but uncontextualized LLM guidance may mislead if presented without real-time clinical oversight, posing a substantial safety risk. Patients must be cautioned that any specific medication recommendations from LLMs must always be reviewed and validated by licensed health care professionals before being acted upon.

Readability was an essential metric; both Kimi k1.5 and GPT-4.0 excelled in generating patient-facing content with concise, clear language and minimal jargon, greatly enhancing accessibility and user comprehension [45,46]. These findings underscore a path for further model refinements to improve the communication of medical information to lay audiences.

Most LLMs systematically incorporated health disclaimers, such as "This information cannot replace professional medical advice." [47,48], which is integral to patient safety. However, inconsistent disclaimer inclusion for less critical questions was observed, calling for the standardization of safety messages across all LLM-generated medical content. Despite generally appropriate use of disclaimers, occasional omissions were noted, representing a residual safety concern, as their absence may increase the risk of patients misinterpreting or misapplying AI-generated advice. To address this, future iterations of medical LLMs should enforce uniform attachment of health advice disclaimers to every health-oriented output, regardless of perceived question severity.

Our study also has some limitations. External generalizability is restricted by the sample size (84 patients and 26 rheumatologists) and single-center, urban tertiary hospital setting, which may limit the applicability of results to broader populations with axSpA with different demographics, health literacy, or health care access. For instance, patients in this top-tier hospital may have distinct expectations, backgrounds, or experiences compared to those in regional or rural centers. In addition, the generalizability of LLM performance and user acceptance may vary by familiarity with digital health tools and local medicolegal contexts. Further multicenter studies spanning diverse socioeconomic and health care environments are necessary to validate these findings and extend the questionnaire's utility. In addition, reliance on 2 raters for accuracy assessments introduces some subjective bias, although this was minimized via strict guideline adherence and a structured arbitration protocol involving a third researcher. Finally, the exclusive use of Chinese-language responses may not fully extrapolate to other linguistic settings.

This research emphasizes the urgency of patient-centered communication tools in axSpA management and illuminates critical shortcomings in current educational practices. The continual evolution of LLMs offers significant promise and unique challenges for supporting chronic disease care with personalized, accessible, and evidence-grounded information. Addressing AI hallucination through improved model development, integrated fact-checking, and explicit cautionary guidance is imperative to ensure responsible and safe adoption of LLMs in patient health care.

Acknowledgments

The authors would like to sincerely thank Jun Zhang for contributing ChatGPT-related insights and answers to this work. The authors would also like to thank all the patients and rheumatologists who participated in this study.

During the preparation of this work, the authors used DeepSeek R1, Hunyuan T1, Kimi k1.5, Wenxin X1, and GPT-4.0. After using these tools or services, the authors reviewed and edited the content as needed. The authors take full responsibility for the content of this study.

Funding

This work was supported by Beijing Natural Science Foundation (grant L242143).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

Conceptualization: JB, XJ, YW, JZ
Data curation: JB, JY, YG
Formal analysis: JB, XJ, YW, CX, WZ
Software: JB, YG
Supervision: XJ, JZ
Writing—original draft: JB
Writing—review and editing: JB, XJ, JY, YW, YG, CX, WZ, JZ

Conflicts of Interest

None declared.

Multimedia Appendix 1

Final version of the questionnaire.
[DOC File , 19 KB-Multimedia Appendix 1]

Multimedia Appendix 2

All original responses from 5 large language models.
[DOC File , 417 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Scoring standard.
[DOC File , 13 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Comparative analysis of the 5 large language models' outputs.
[PDF File (Adobe PDF File), 307 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Specific results of chi-square test 1.
[DOC File , 15 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Specific results of chi-square test 2.
[DOC File , 15 KB-Multimedia Appendix 6]

Multimedia Appendix 7

The scoring results of the various models.
[DOC File , 14 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Specific results of chi-square test 3.
[DOC File , 29 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Baseline characteristics of different age groups.
[DOC File , 16 KB-Multimedia Appendix 9]

References

1. Navarro-Compán V, Sepriano A, Capelusnik D, Baraliakos X. Axial spondyloarthritis. Lancet. Jan 11, 2025;405(10473):159-172. [doi: 10.1016/S0140-6736(24)02263-3] [Medline: 39798984]
2. Maksymowych WP, Carmona R, Weber U, Aydin SZ, Yeung J, Reis J, et al. Features of axial spondyloarthritis in two multicenter cohorts of patients with psoriasis, uveitis, and colitis presenting with undiagnosed back pain. Arthritis Rheumatol. Jan 2025;77(1):47-58. [doi: 10.1002/art.42967] [Medline: 39107875]
3. Ortolan A, Webers C, Sepriano A, Falzon L, Baraliakos X, Landewé RB, et al. Efficacy and safety of non-pharmacological and non-biological interventions: a systematic literature review informing the 2022 update of the ASAS/EULAR recommendations for the management of axial spondyloarthritis. Ann Rheum Dis. Jan 2023;82(1):142-152. [doi: 10.1136/ard-2022-223297] [Medline: 36261247]
4. Ramiro S, Nikiphorou E, Sepriano A, Ortolan A, Webers C, Baraliakos X, et al. ASAS-EULAR recommendations for the management of axial spondyloarthritis: 2022 update. Ann Rheum Dis. Jan 2023;82(1):19-34. [FREE Full text] [doi: 10.1136/ard-2022-223296] [Medline: 36270658]

5. Chen X, Wang L, You M, Liu W, Fu Y, Xu J, et al. Evaluating and enhancing large language models' performance in domain-specific medicine: development and usability study with DocOA. J Med Internet Res. Jul 22, 2024;26:e58158. [FREE Full text] [doi: 10.2196/58158] [Medline: 38833165]
6. Li C, Zhao Y, Bai Y, Zhao B, Tola YO, Chan CW, et al. Unveiling the potential of large language models in transforming chronic disease management: mixed methods systematic review. J Med Internet Res. Apr 16, 2025;27:e70535. [FREE Full text] [doi: 10.2196/70535] [Medline: 40239198]
7. Diekhoff T, Giraudo C, Machado PM, Mallinson M, Eshed I, Haibel H, et al. Clinical information on imaging referrals for suspected or known axial spondyloarthritis: recommendations from the Assessment of Spondyloarthritis International Society (ASAS). Ann Rheum Dis. Nov 14, 2024;83(12):1636-1643. [FREE Full text] [doi: 10.1136/ard-2024-226280] [Medline: 39317418]
8. Beauvais C, Pereira B, Pham T, Sordet C, Claudepierre P, Fayet F, et al. Development and validation of a self-administered questionnaire measuring essential knowledge in patients with axial spondyloarthritis. J Rheumatol. Jan 2023;50(1):56-65. [FREE Full text] [doi: 10.3899/jrheum.211314] [Medline: 35840152]
9. Zhang J, Wang J, Zhang J, Xia X, Zhou Z, Zhou X, et al. Young adult perspectives on artificial intelligence-based medication counseling in China: discrete choice experiment. J Med Internet Res. Apr 09, 2025;27:e67744. [FREE Full text] [doi: 10.2196/67744] [Medline: 40203305]
10. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med. Aug 2025;31(8):2550-2555. [doi: 10.1038/s41591-025-03726-3] [Medline: 40267969]
11. Ibrahim AF, Danpanichkul P, Hayek A, Paul E, Farag A, Mansoor M, et al. Artificial intelligence in gastroenterology education: DeepSeek passes the gastroenterology board examination and outperforms legacy ChatGPT models. Am J Gastroenterol. May 20, 2025. [doi: 10.14309/ajg.0000000000003552] [Medline: 40392256]
12. Kang D, Wu H, Yuan L, Shen W, Feng J, Zhan J, et al. Evaluating the efficacy of large language models in guiding treatment decisions for pediatric refractive error. Ophthalmol Ther. Apr 2025;14(4):705-716. [doi: 10.1007/s40123-025-01105-2] [Medline: 39985747]
13. Su Z, Jin K, Wu H, Luo Z, Grzybowski A, Ye J. Assessment of large language models in cataract care information provision: a quantitative comparison. Ophthalmol Ther. Jan 08, 2025;14(1):103-116. [doi: 10.1007/s40123-024-01066-y] [Medline: 39516445]
14. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Aug 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
15. Meissner Y, Strangfeld A, Molto A, Forger F, Wallenius M, Costedoat-Chalumeau N, et al. EuNeP collaborator group. Pregnancy and neonatal outcomes in women with axial spondyloarthritis: pooled data analysis from the European Network of Pregnancy Registries in Rheumatology (EuNeP). Ann Rheum Dis. Nov 2022;81(11):1524-1533. [doi: 10.1136/ard-2022-222641] [Medline: 35961759]
16. Ribeiro AL, Proft F. Unraveling the challenges of difficult-to-treat spondyloarthritis: SPARTAN 2024 annual meeting proceedings. Curr Rheumatol Rep. Feb 03, 2025;27(1):18. [doi: 10.1007/s11926-025-01183-y] [Medline: 39899221]
17. Poddubnyy D, Sieper J. Treatment of axial spondyloarthritis: what does the future hold? Curr Rheumatol Rep. Jul 20, 2020;22(9):47. [FREE Full text] [doi: 10.1007/s11926-020-00924-5] [Medline: 32691259]
18. McGonagle D, Ramonda R, Scagnellato L, Scriffignano S, Weddell J, Lubrano E. A strategy towards disentangling treatment refractory from misdiagnosed axial spondyloarthritis. Autoimmun Rev. Jan 2024;23(1):103405. [doi: 10.1016/j.autrev.2023.103405] [Medline: 37543288]
19. Bechman K, Yang Z, Adas M, Nagra D, S Uğuzlar A, Russell MD, et al. Incidence of uveitis in patients with axial spondylarthritis treated with biologics or targeted synthetics: a systematic review and network meta-analysis. Arthritis Rheumatol. May 2024;76(5):704-714. [doi: 10.1002/art.42788] [Medline: 38116697]
20. Lei L, Wei Y, Liu K. AlphaReadabilityChinese: a tool for the measurement of readability in Chinese texts and its applications. Foreign Lang Teach. 2024;46(1):83-93. [doi: 10.13458/j.cnki.flatt.004997]
21. Anibal JT, Huth HB, Gunkel J, Gregurick SK, Wood BJ. Simulated misuse of large language models and clinical credit systems. NPJ Digit Med. Nov 11, 2024;7(1):317. [FREE Full text] [doi: 10.1038/s41746-024-01306-2] [Medline: 39528596]
22. Li X, Wang H, Zhao R, Wang T, Zhu Y, Qian Y, et al. Elevated extracellular volume fraction and reduced global longitudinal strains in participants recovered from COVID-19 without clinical cardiac findings. Radiology. May 2021;299(2):E230-E240. [doi: 10.1148/radiol.2021203998] [Medline: 33434112]
23. Zivanovic S, Papic M, Vucicevic T, Miletic Kovacevic M, Jovicic N, Nikolic N, et al. Periapical lesions in two inbred strains of rats differing in immunological reactivity. Int Endod J. Jan 2022;55(1):64-78. [doi: 10.1111/iej.13638] [Medline: 34614243]
24. Wucherpfennig L, Wuennemann F, Eichinger M, Seitz A, Baumann I, Stahl M, et al. Long-term effects of lumacaftor/ivacaftor on paranasal sinus abnormalities in children with cystic fibrosis detected with magnetic resonance imaging. Front Pharmacol. Apr 10, 2023;14:1161891. [FREE Full text] [doi: 10.3389/fphar.2023.1161891] [Medline: 37101549]

25. Plavén-Sigray P, Hedman E, Victorsson P, Matheson GJ, Forsberg A, Djurfeldt DR, et al. Extrastriatal dopamine D2-receptor availability in social anxiety disorder. Eur Neuropsychopharmacol. May 2017;27(5):462-469. [FREE Full text] [doi: 10.1016/j.euroneuro.2017.03.007] [Medline: 28377075]
26. Thorolfsson B, Lundgren M, Snaebjornsson T, Karlsson J, Samuelsson K, Senorski EH. Lower rate of acceptable knee function in adolescents compared with young adults five years after ACL reconstruction: results from the Swedish National Knee Ligament Register. BMC Musculoskelet Disord. Aug 19, 2022;23(1):793. [FREE Full text] [doi: 10.1186/s12891-022-05727-6] [Medline: 35982445]
27. Xu JT, Li K, Lin Y, Cheng T, Gu J, Chen YK, et al. Diverse impacts of different rpoB mutations on the anti-tuberculosis efficacy of capreomycin. EBioMedicine. Jul 2025;117:105776. [FREE Full text] [doi: 10.1016/j.ebiom.2025.105776] [Medline: 40449326]
28. Bittar M, Khan MA, Magrey M. Axial spondyloarthritis and diagnostic challenges: over-diagnosis, misdiagnosis, and under-diagnosis. Curr Rheumatol Rep. Mar 2023;25(3):47-55. [doi: 10.1007/s11926-022-01096-0] [Medline: 36602692]
29. Marques ML, Ramiro S, van Lunteren M, Stal RA, Landewé RB, van de Sande M, et al. Can rheumatologists unequivocally diagnose axial spondyloarthritis in patients with chronic back pain of less than 2 years duration? Primary outcome of the 2-year SPondyloArthritis Caught Early (SPACE) cohort. Ann Rheum Dis. Apr 11, 2024;83(5):589-598. [doi: 10.1136/ard-2023-224959] [Medline: 38233104]
30. Bittar M, Deodhar A. Axial spondyloarthritis: a review. JAMA. Feb 04, 2025;333(5):408-420. [doi: 10.1001/jama.2024.20917] [Medline: 39630439]
31. Capelusnik D, Boonen A, Ramiro S, Nikiphorou E. The role of social determinants of health on disease outcomes in axial spondyloarthritis: a narrative review. Autoimmun Rev. Apr 30, 2025;24(5):103762. [doi: 10.1016/j.autrev.2025.103762] [Medline: 39922474]
32. Ørnbjerg LM, Georgiadis S, Kvien TK, Michelsen B, Rasmussen S, Pavelka K, et al. Impact of patient characteristics on ASDAS disease activity state cut-offs in axial spondyloarthritis: results from nine European rheumatology registries. RMD Open. Nov 02, 2024;10(4):e004644. [FREE Full text] [doi: 10.1136/rmdopen-2024-004644] [Medline: 39489531]
33. Garrido-Cumbrera M, Gálvez-Ruiz D, Delgado-Domínguez CJ, Poddubnyy D, Navarro-Compán V, Christen L, et al. EMAS working group. Impact of axial spondyloarthritis on mental health in Europe: results from the EMAS study. RMD Open. Nov 2021;7(3):e001769. [FREE Full text] [doi: 10.1136/rmdopen-2021-001769] [Medline: 34740979]
34. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of openAI's GPT-4 and Google's Gemini. Comput Biol Med. Feb 2025;185:109545. [FREE Full text] [doi: 10.1016/j.compbiomed.2024.109545] [Medline: 39667055]
35. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. J Med Internet Res. Apr 05, 2024;26:e52935. [FREE Full text] [doi: 10.2196/52935] [Medline: 38578685]
36. Schmidgall S, Harris C, Essien I, Olshvang D, Rahman T, Kim JW, et al. Evaluation and mitigation of cognitive biases in medical language models. NPJ Digit Med. Oct 21, 2024;7(1):295. [FREE Full text] [doi: 10.1038/s41746-024-01283-6] [Medline: 39433945]
37. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med. Mar 29, 2024;7(1):82. [FREE Full text] [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
38. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). Jan 21, 2025;5(1):26. [FREE Full text] [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
39. Antiperovitch P, Liu I, Mokhtar AT, Tang A. Evaluating large language models in cardiovascular antithrombotic care: performance, accuracy, and implications for clinical practice. Can J Cardiol. Aug 2025;41(8):1584-1591. [doi: 10.1016/j.cjca.2025.04.008] [Medline: 40239865]
40. Williams CY, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. Nat Commun. Oct 08, 2024;15(1):8236. [FREE Full text] [doi: 10.1038/s41467-024-52415-1] [Medline: 39379357]
41. Danve A, Deodhar A. Treatment of axial spondyloarthritis: an update. Nat Rev Rheumatol. Apr 2022;18(4):205-216. [doi: 10.1038/s41584-022-00761-z] [Medline: 35273385]
42. Long Z, Deng Y, He Q, Yang K, Zeng L, Hao W, et al. Efficacy and safety of iguratimod in the treatment of ankylosing spondylitis: a systematic review and meta-analysis of randomized controlled trials. Front Immunol. Mar 03, 2023;14:993860. [FREE Full text] [doi: 10.3389/fimmu.2023.993860] [Medline: 36936924]
43. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. Nat Med. Jun 25, 2024;30(6):1574-1582. [FREE Full text] [doi: 10.1038/s41591-024-02933-8] [Medline: 38664535]
44. Vordenberg SE, Nichols J, Marshall VD, Weir KR, Dorsch MP. Investigating older adults' perceptions of AI tools for medication decisions: vignette-based experimental survey. J Med Internet Res. Dec 16, 2024;26:e60794. [FREE Full text] [doi: 10.2196/60794] [Medline: 39680885]

45. Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E. Large language models may help patients understand peer-reviewed scientific articles about ophthalmology: development and usability study. J Med Internet Res. Dec 24, 2024;26:e59843. [FREE Full text] [doi: 10.2196/59843] [Medline: 39719077]
46. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. Radiology. Mar 01, 2024;310(3):e231593. [doi: 10.1148/radiol.231593] [Medline: 38530171]
47. Menz BD, Modi ND, Abuhelwa AY, Ruanglertboon W, Vitry A, Gao Y, et al. Generative AI chatbots for reliable cancer information: evaluating web-search, multilingual, and reference capabilities of emerging large language models. Eur J Cancer. Mar 11, 2025;218:115274. [FREE Full text] [doi: 10.1016/j.ejca.2025.115274] [Medline: 39922126]
48. Seo J, Choi D, Kim T, Cha WC, Kim M, Yoo H, et al. Evaluation framework of large language models in medical documentation: development and usability study. J Med Internet Res. Nov 20, 2024;26:e58329. [FREE Full text] [doi: 10.2196/58329] [Medline: 39566044]

Abbreviations

AI: artificial intelligence
axSpA: axial spondyloarthritis
LLM: large language model
TCM: traditional Chinese medicine

Edited by F Dankar; submitted 16.Jun.2025; peer-reviewed by S Biswas, H Wang, J Grosser; comments to author 03.Sep.2025; revised version received 28.Oct.2025; accepted 31.Oct.2025; published 07.Jan.2026

Please cite as:
Bai J, Ji X, Yu J, Wang Y, Guo Y, Xue C, Zhang W, Zhu J
Assessing the Quality of AI Responses to Patient Concerns About Axial Spondyloarthritis: Delphi-Based Evaluation
JMIR AI 2026;5:e79153
URL: https://ai.jmir.org/2026/1/e79153
doi: 10.2196/79153
PMID: