



# BMJ Open Evaluating the performance of large language models in health education for patients with ankylosing spondylitis/spondyloarthritis: a cross-sectional, single-blind study in China

Yong Ren <sup>1,2</sup>, Yue-ning Kang,<sup>3</sup> Shuang-yan Cao,<sup>3</sup> Fanxuan Meng,<sup>3</sup> Jingyu Zhang,<sup>3</sup> Ruyi Liao,<sup>3</sup> Xiaomin Li,<sup>3</sup> Yuling Chen,<sup>3</sup> Ya Wen,<sup>3</sup> Jiayun Wu,<sup>3</sup> Wenqi Xia <sup>3</sup>, Liling Xu,<sup>3</sup> Shenghui Wen,<sup>3</sup> Huifen Liu,<sup>3</sup> Yuanqing Li,<sup>4,5</sup> Jieruo Gu,<sup>3,6</sup> Qing Lv<sup>3</sup>

**To cite:** Ren Y, Kang Y, Cao S, *et al.* Evaluating the performance of large language models in health education for patients with ankylosing spondylitis/spondyloarthritis: a cross-sectional, single-blind study in China. *BMJ Open* 2025;**15**:e097528. doi:10.1136/bmjopen-2024-097528

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2024-097528>).

YR, Y-nK and S-yC contributed equally.

Received 04 December 2024  
Accepted 28 February 2025



© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

## Correspondence to

Qing Lv; [lvqing@sysush.com](mailto:lvqing@sysush.com),  
Jieruo Gu; [gujieruo@163.com](mailto:gujieruo@163.com)  
and  
Yuanqing Li;  
[auyqli@scut.edu.cn](mailto:auyqli@scut.edu.cn)

## ABSTRACT

**Objectives** To evaluate the potential of large language models (LLMs) in health education for patients with ankylosing spondylitis (AS)/spondyloarthritis (SpA), focusing on the accuracy of information transmission, patient acceptance and performance differences between different models.

**Design** Cross-sectional, single-blind study.

**Setting** Multiple centres in China.

**Participants** 182 volunteers, including 4 rheumatologists and 178 patients with AS/SpA.

### Primary and secondary outcome measures

Scientificity, precision and accessibility of the content of the answers provided by LLMs; patient acceptance of the answers.

**Results** LLMs performed well in terms of scientificity, precision and accessibility, with ChatGPT-4o and Kimi models outperforming traditional guidelines. Most patients with AS/SpA showed a higher level of understanding and acceptance of the responses from LLMs.

**Conclusions** LLMs have significant potential in medical knowledge transmission and patient education, making them promising tools for future medical practice.

## INTRODUCTION

Ankylosing spondylitis/spondyloarthritis (AS/SpA) is a chronic autoimmune inflammatory disease that primarily affects the sacroiliac joints, spine and peripheral joints, leading to inflammation, bone spurs<sup>1</sup> and, eventually, in severe cases, fusion of the spine, significantly impacting a patient's quality of life and ability to work.<sup>1–3</sup> The total prevalence of AS was 0.1–0.5% and continues to increase, with a male-to-female ratio of 2:1 to 4:1.<sup>4–6</sup> Currently, the primary treatment strategy for AS/SpA involves a combination of anti-inflammatory drugs, physical therapy and exercise rehabilitation.<sup>7</sup> The goal of

## STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The study included a diverse group of participants from multiple centres in China.
- ⇒ The use of a cross-sectional, single-blind design helps to minimise bias.
- ⇒ The evaluation of multiple large language models (LLMs) provides a comprehensive comparison.
- ⇒ The study is limited by its relatively small sample size and the potential for questionnaire fatigue among participants.
- ⇒ The study only evaluated four LLMs, and future models may offer different performance characteristics.

these treatments is to alleviate symptoms, control inflammation, inhibit bone formation and thus maintain and improve spinal and peripheral joint function, reducing the risk of disability.<sup>4,8–11</sup> Growing evidence suggests that for a chronic disease like AS/SpA, the prognosis is closely tied to adherence to the basic treatment strategy, that is, patient compliance.<sup>12</sup> Educating patients about the disease, conducting regular check-ups and assessments and encouraging patients to engage in self-management, such as adhering to medication, engaging in appropriate exercise and quitting smoking and drinking, are essential for improving patient compliance.<sup>13–18</sup>

How to educate patients about AS/SpA using professional yet easily understandable language, making it easier for patients to comprehend and accept related knowledge, and address their questions and concerns in real-time has always been a challenging task.<sup>19</sup> Healthcare professionals, while possessing expert knowledge, often lack sufficient time to provide detailed answers to patients'

questions at any given moment. Medical textbooks, on the other hand, often contain complex terminology and content that are difficult for patients to understand. Whether there is a more cost-effective and convenient way to complete or assist in this task is a question worth exploring. Therefore, finding a more economical and convenient way to provide or assist in providing concise, reliable and easy-to-understand medical information to patients with AS/SpA has become a significant challenge in clinical practice.

In recent years, the rapid development of artificial intelligence (AI), particularly the emergence of large language models (LLMs), has provided new solutions for acquiring and disseminating medical information.<sup>20–29</sup> Advanced AI models such as ChatGPT4o possess powerful natural language understanding and text generation capabilities, enabling them to process vast amounts of information and generate high-quality responses in a short period.<sup>30</sup> Trained on massive datasets using deep learning algorithms such as the transformer architecture, these models can produce coherent and contextually relevant responses, simulating the response style of human experts to a certain extent.<sup>31–32</sup> The application scenarios of large models have expanded to various fields, including medical report generation, demonstrating their potential in medical information processing.<sup>33–35</sup>

However, it is important to note that LLMs also have limitations, including potential biases and the occurrence of hallucinations, which can affect the accuracy and reliability of their responses. These issues need to be carefully considered and addressed in the development and application of such models.<sup>36</sup> Additionally, the performance of different models in terms of response quality and patient acceptance needs further comparison and research to determine the most suitable model for practical applications.<sup>37</sup>

Therefore, we apply LLM technology to health education for patients with AS/SpA to effectively convey disease-related knowledge, help patients better understand the disease and treatment plans, enhance their self-management awareness and ability, improve patient adherence to treatment and ultimately improve disease prognosis and quality of life.<sup>36–38</sup>

The aim of this study is to evaluate the potential of LLMs in health education for patients with AS/SpA. Specifically, we focus on the accuracy of information transmission, patient acceptance and performance differences between different models.

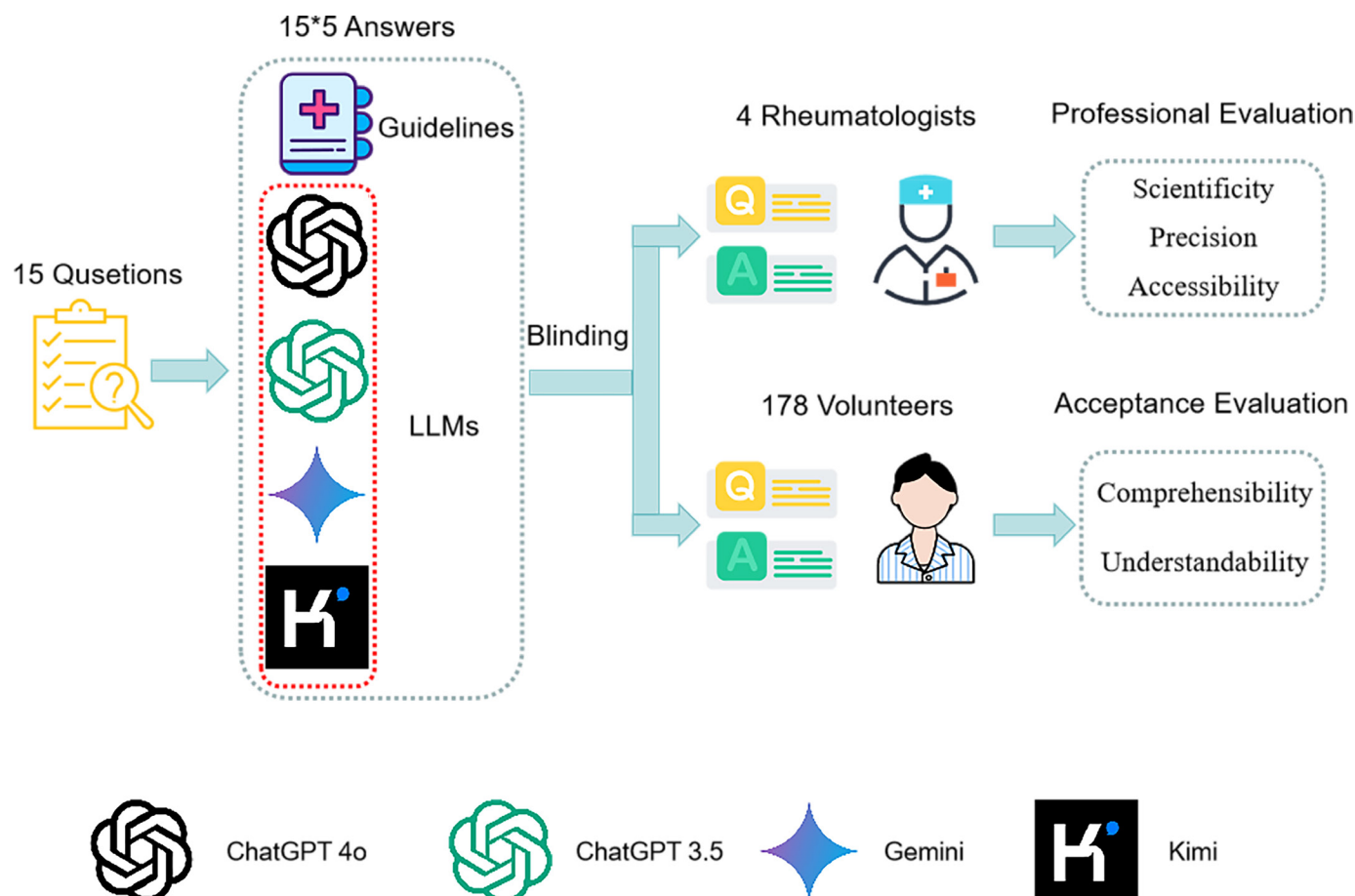
## METHODS

A cross-sectional, single-blind study design was employed, involving 182 volunteers from different regions of China. The study was conducted in Chinese to ensure that all participants could understand and respond to the questionnaires and instructions. Volunteers were asked to complete two online surveys according to the

requirements, and the survey data were subsequently collected and analysed (figure 1).

## Questionnaire design

Two questionnaires were used in this study. Questionnaire 1 was designed to evaluate the scientificity, precision and accessibility of the content of the answers provided by different LLMs to medical questions. Questionnaire 2 was used to evaluate patients' acceptance of the answers provided by different LLMs to medical questions. Both questionnaires included 15 AS/SpA-related questions as the basic questions (online supplemental documents). These 15 questions were derived from the 'Practical Guidelines for Patients with AS/SpA' (hereafter referred to as the 'Guidelines'),<sup>39</sup> which were developed to help patients better understand and implement the principles of AS/SpA diagnosis and treatment and participate in disease management. Given the complexity and diversity of rheumatic diseases, including all conditions in our study would have required an extensive number of questions, making the feasibility of the study challenging. Therefore, we chose to focus on a single disease for our initial research. Among the conditions considered, AS/SpA was selected because it is one of the few diseases for which a comprehensive patient self-management guideline is available. These Guidelines formed the basis for our study design and provided a structured framework for evaluating the performance of LLMs in patient education. Although there are several international and national guidelines available for the diagnosis and treatment of AS/SpA, these guidelines are primarily aimed at healthcare professionals. This Guidelines was chosen because it is highly regarded for its scientific accuracy, practicality and comprehensiveness, making it a valuable reference for patient education in clinical practice. According to the Guidelines, these 15 questions were selected through a systematic review of published guidelines, research papers and systematic reviews in the field of AS/SpA, as well as face-to-face interviews with 52 rheumatologists and other multidisciplinary experts and patients with AS/SpA. A total of 34 questions were selected, and a questionnaire survey was completed by presidents or patient representatives of AS/SpA patient associations from 19 countries and 6 regions in Asia, America and Europe. Finally, 15 questions were selected according to the ranking. These 15 questions cover multiple aspects, including disease management, treatment options and quality of life, and are representative. For each of the 15 basic questions, our two questionnaires provided five answers, one of which was from the original description of the Guidelines, and the other four were from the four LLMs (OpenAI's ChatGPT4o, OpenAI's ChatGPT3.5, Google's Gemini 1.5 Flash and Moonshot's Kimi), each providing a distinct answer option for evaluation. Questionnaire 1 evaluated the scientificity, precision and accessibility of the content of each answer to the 15 questions; Questionnaire 2 required participants to select the answer that was easy



**Figure 1** Research roadmap. A cross-sectional, single-blind study design was employed. 15 questions related to ankylosing spondylitis (AS)/spondyloarthritis (SpA) were developed. Answers to these questions were provided by both the 'Clinical Practice Guidelines for Patients with AS/SpA' and four large language models (LLMs): ChatGPT-4o, ChatGPT-3.5, Gemini1.5 Flash and Kimi. Based on the questions and corresponding answers, two questionnaires were constructed. Questionnaire 1 was completed by four volunteer physicians using a 5-point Likert scale to evaluate the professionalism of the answers. Questionnaire 2, consisting of multiple-choice questions, was completed by 178 voluntary patients with AS to assess their understanding and acceptance of the answers. Finally, data from both questionnaires were collected and analysed.

to understand and acceptable among the five answers for each question.

#### Volunteer recruitment and questionnaire collection

A total of 182 volunteers were enrolled in the study, including 4 rheumatologists with over 5 years of experience and 178 patients with a confirmed diagnosis of AS/SpA.<sup>40</sup> AS/SpA patient volunteers were recruited through the internet, with online survey links distributed to potential participants via social media groups and email. The questionnaire collection period lasted 1 month. The four rheumatologists were required to complete Questionnaire 1 to evaluate the professionalism of the answers. Doctors needed to evaluate each answer to 15 questions based on three aspects: scientificity, precision and accessibility of the content. Scientificity refers to whether the answer conforms to the theoretical basis of modern medicine, the development patterns of the disease and the evidence-based medicine of diagnosis and treatment. Precision refers to whether the concepts in the answer are clear, the expression is accurate and the word choice is appropriate. Accessibility

refers to whether complex medical knowledge can be explained in simple and easy-to-understand language, and whether the communication style can be adjusted from the patient's perspective. The evaluation used a 5-point Likert scale, ranging from 1 to 5, with 1 being the lowest and 5 being the highest.

The 178 AS/SpA volunteers completed Questionnaire 2 to evaluate the acceptability of the answers, including comprehensibility and understandability. Comprehensibility refers to the ability to understand the meaning of medical information and to relate the medical information to one's own situation. Understandability refers to the clarity and simplicity of the presentation of medical information, avoiding the use of obscure technical terms. Unlike doctors, to ensure the reliability of the questionnaire results and avoid volunteers feeling bored when faced with a large number of scientific answers, leading to biased choices, volunteers only needed to evaluate the answers to the questions they were interested in and did not need to complete the evaluation of all questions. For each question of interest, they could select one or



more answers that they considered understandable and acceptable.

It is worth noting that neither the doctors nor the AS/SpA volunteers participating in the survey were aware of the source of the answers. In other words, they did not know whether the answers came from the Guidelines or from LLMs, nor did they know which LLM was used.

### Data analysis

In the statistical analysis, we first assessed the normality of the data using the Shapiro-Wilk test. For the content of the 75 answers collected from Questionnaire 1, the average scores for scientificity, precision and accessibility were calculated for each of the four questionnaires from four doctors. Paired t-tests were used to analyse the differences in scores between the answers provided by LLMs and the answers from the Guidelines when the data were normally distributed. For the evaluation of accessibility, differences between the Guidelines versus Kimi groups were normally distributed and analysed using paired t-tests, while differences between other groups were not normally distributed and were analysed using the Wilcoxon signed-rank test. For the 178 questionnaires collected from Questionnaire 2, statistical analysis was conducted. First, the basic demographic characteristics of the volunteers were analysed. The 15 questions were categorised into four groups: disease diagnosis and assessment, daily life management, disease drug treatment and disease and fertility (box 1). The number of people interested in each category of questions was counted, and the proportion of people interested in each category was calculated.

Next, based on the AS/SpA volunteers' choices for each question, the selection situations were categorised into three types: 'Considered the Guidelines answer better' (only selected the answer from the Guidelines), 'Considered the Guidelines answer and one or more LLMs answers equally good' (selected both the answer from the Guidelines and one or more answers from the LLMs) and 'Considered the LLMs answer better' (only selected one or more answers from the LLMs). The proportion of the three categories of choices for each question and the overall proportion of the three categories of choices for the 15 questions were calculated to evaluate the acceptance of the AS/SpA volunteers towards the answers from the Guidelines and the LLMs.

Finally, a detailed analysis was conducted on the proportion of the number of choices for the answers of the four different models in the 15 questions to analyse the differences in answer quality and volunteer acceptance among the four models and to evaluate the performance of each model.

### Model access and data handling

In our study, we used specific versions of the LLMs to ensure consistency and reproducibility. The models and their versions used were: OpenAI's ChatGPT-4o, OpenAI's ChatGPT-3.5, Google's Gemini 1.5 Flash and Moonshot's

### Box 1 Classification of the 15 questions into four groups based on their respective attributes

#### Group 1: disease diagnosis and assessment

- ⇒ What is ankylosing spondylitis/spondyloarthritis (AS/SpA)?
- ⇒ Can the presence of the human leukocyte antigen (HLA)-B27 confirm the diagnosis of AS/SpA?
- ⇒ Why do patients with AS/SpA need to have their C reactive protein levels checked regularly?
- ⇒ Why do patients with AS/SpA need to have MRI scans repeated?
- ⇒ Why should patients with AS/SpA be tested for hepatitis B virus (HBV)?

#### Group 2: daily life management

- ⇒ Does smoking affect disease activity and function in patients with AS/SpA?
- ⇒ What are the benefits of exercise for patients with AS/SpA?

#### Group 3: disease drug treatment

- ⇒ What are the first-line medications for AS/SpA?
- ⇒ What are the cardiovascular effects of oral non-steroidal anti-inflammatory drugs (NSAIDs) in patients with AS/SpA?
- ⇒ What are the gastrointestinal effects of oral NSAIDs in patients with AS/SpA?
- ⇒ When are tumour necrosis factor (TNF)-alpha inhibitors used to treat AS/SpA?
- ⇒ Can patients with AS/SpA with HBV coinfection be treated with TNF-alpha inhibitors?
- ⇒ Does using TNF-alpha inhibitors increase the risk of tuberculosis infection in patients with AS/SpA?

#### Group 4: disease and fertility

- ⇒ Can HLA-B27-positive AS/SpA be passed on to the next generation?
- ⇒ What are the effects of TNF-alpha inhibitors on fertility in patients with AS/SpA?

Kimi. These models were accessed via their respective web interfaces, which allowed us to interact with the models directly through a web browser and retrieve responses to our queries.

Data storage and management were handled securely to ensure privacy and integrity. All data collected from the questionnaires and the responses generated by the LLMs were stored in an encrypted database. Access to the database was restricted to authorised personnel only, and all data transfers were conducted over secure channels.

To handle errors in calling or parsing answers, we implemented a robust error-handling mechanism. This included logging all errors encountered during the interaction with the web interfaces. If an error occurred, the specific query and error details were reviewed manually to identify and resolve the issue. This approach ensured that any transient issues were addressed, and any systemic problems were identified and corrected promptly.

### Patient and public involvement

In this study, there was no patient or public involvement. The research design, implementation and dissemination of results were all independently completed by the research team.

**Table 1** Demographic and clinical characteristics of patients

Characteristic	Value (proportion %)
Gender	
Male	142 (79.78)
Female	36 (20.22)
Age (years)	
<20	0 (0)
20–30	12 (6.75)
30–40	95 (53.37)
40–50	58 (32.58)
50–60	11 (6.18)
>60	2 (1.12)

## RESULTS

### Participant characteristics and interest distribution

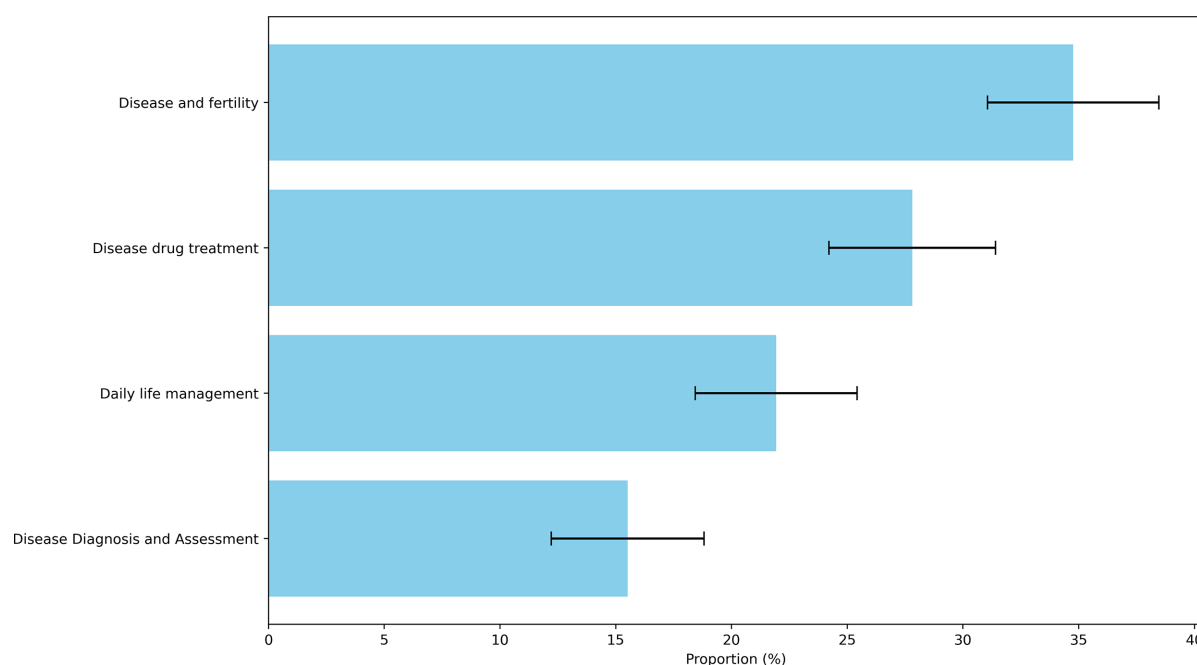
Among the 178 AS/SpA volunteers included in this study, 79.78% were male and 20.22% were female, which is generally consistent with the gender distribution of patients with AS/SpA. The largest number of AS/SpA volunteers were aged 30–40 years and 40–50 years, accounting for 53.37% and 32.58%, respectively (table 1). Based on the IP addresses of the AS/SpA volunteers who completed the questionnaire, the geographical distribution of the volunteers was analysed. It was found that AS/SpA volunteers covered 23 of the 34 provincial-level administrative regions in China, with

a geographical coverage rate of 67.6%. AS/SpA volunteers were most interested in the following categories of questions in descending order: disease and fertility (34.76%), disease drug treatment (27.81%), daily life management (21.93%) and disease diagnosis and assessment (15.51%) (figure 2).

### Evaluation of professionalism of LLM answers

The rheumatologists' evaluation of the scientificity, precision and accessibility of the answers provided by LLMs is shown in table 2. The scores for the scientificity of the content of the answers provided by the four LLMs (ChatGPT-4o:  $4.68 \pm 0.10$ , ChatGPT-3.5:  $4.43 \pm 0.10$ , Gemini 1.5 Flash:  $4.13 \pm 0.12$  and Kimi:  $4.77 \pm 0.07$ ) were very close to the score of the Guidelines ( $4.72 \pm 0.09$ ), with Kimi's score being even slightly higher than that of the Guidelines. Except for Gemini 1.5 Flash ( $p < 0.05$ ), there was no statistically significant difference between the scores of the other models and the Guidelines ( $p > 0.05$ ).

The scores for the precision of the expression of the answers provided by the four models (ChatGPT-4o:  $4.57 \pm 0.11$ , ChatGPT-3.5:  $4.38 \pm 0.10$ , Gemini 1.5 Flash:  $3.85 \pm 0.11$  and Kimi:  $4.78 \pm 0.06$ ) were very close to the score of the Guidelines ( $4.70 \pm 0.07$ ), with Kimi's score being even slightly higher than that of the Guidelines. There were statistically significant differences between the scores of ChatGPT-3.5 and Gemini 1.5 Flash and the Guidelines ( $p < 0.05$ ), but no statistically significant differences between the scores of ChatGPT-4o and Kimi and the Guidelines ( $p > 0.05$ ).



**Figure 2** Distribution of patients with ankylosing spondylitis' interest across different question categories. The 15 questions were classified into four categories. A horizontal bar chart, sorted by the proportion of patients' interest, illustrates the distribution of questions within each category. Patients show interest in the following areas with the following percentages: 'disease and fertility' at 34.76% $\pm$ 3.7%, 'disease drug treatment' at 27.81% $\pm$ 3.6%, 'daily life management' at 21.93% $\pm$ 3.5% and 'disease diagnosis and assessment' at 15.51% $\pm$ 3.3%. Error bars representing the SD are added to each bar to illustrate the variability of the data.

**Table 2** Rheumatologists' assessment of the professionalism of responses from different large language models and their comparison with guidelines

Evaluation	5-point Likert scale score	P value*
Scientificity		
Guidelines vs ChatGPT-4o	4.72±0.09 vs 4.68±0.10	0.812
Guidelines vs ChatGPT-3.5	4.72±0.09 vs 4.43±0.10	0.080
Guidelines vs Gemini 1.5 Flash	4.72±0.09 vs 4.13±0.12	0.002
Guidelines vs Kimi	4.72±0.09 vs 4.77±0.07	0.670
Precision		
Guidelines vs ChatGPT-4o	4.70±0.07 vs 4.57±0.11	0.349
Guidelines vs ChatGPT-3.5	4.70±0.07 vs 4.38±0.10	0.039
Guidelines vs Gemini 1.5 Flash	4.70±0.07 vs 3.85±0.11	0.000
Guidelines vs Kimi	4.70±0.07 vs 4.78±0.06	0.475
Accessibility		
Guidelines vs ChatGPT-4o	4.02±0.07 vs 4.67±0.12	0.006
Guidelines vs ChatGPT-3.5	4.02±0.07 vs 4.63±0.08	0.001
Guidelines vs Gemini 1.5 Flash	4.02±0.07 vs 4.38±0.12	0.026
Guidelines vs Kimi	4.02±0.07 vs 4.70±0.06	0.000

\*Evaluation of the scientificity and precision of the questions and answers. Differences between groups were normally distributed, and paired t-tests were used for analysis. For the evaluation of accessibility, differences between the Guidelines versus Kimi groups were normally distributed and analysed using paired t-tests, while differences between other groups were not normally distributed and were analysed using the Wilcoxon signed-rank test.

The scores for the accessibility of the answers provided by ChatGPT-4o, ChatGPT-3.5, Gemini 1.5 Flash and Kimi (4.67±0.12, 4.63±0.08, 4.38±0.12 and 4.70±0.06, respectively) were all higher than the score of the Guidelines (4.02±0.07), and the differences were statistically significant ( $p<0.05$ ).

### Evaluation of patient acceptance of LLM answers

#### Overall trends

On average, 54.8% of patients preferred the answers provided by LLMs, while 16.4% considered the Guidelines' answers to be better, and 28.7% considered the answers from the Guidelines and one or more LLMs to be equally good. This indicates a general preference among patients for the responses generated by LLMs over those from the Guidelines. For most questions, a higher proportion of patients with AS/SpA found the answers of LLMs more acceptable, particularly for questions related to treatment timing and the impact of treatments on fertility (figure 3A and C).

#### Model-level analyses

When comparing the four LLMs, ChatGPT-4o had the highest average selection rate of 33.68%, ranking first. It significantly outperformed other models in specific questions, such as 'What is AS/SpA?' (48.57%) and 'Does TNF-alpha inhibitor treatment increase the risk of tuberculosis infection in AS/SpA patients?' (40.68%). The Kimi model had a stable performance with an average selection rate of 24.77%, ranking second. It performed well in questions like 'What are

the cardiovascular effects of oral NSAID treatment in AS/SpA patients?' (33.33%) and 'When should AS/SpA patients start TNF-alpha inhibitor treatment?' (32.50%). The average selection rates for ChatGPT-3.5 and Gemini 1.5 Flash were 20.27% and 21.28%, respectively (figure 3B and D).

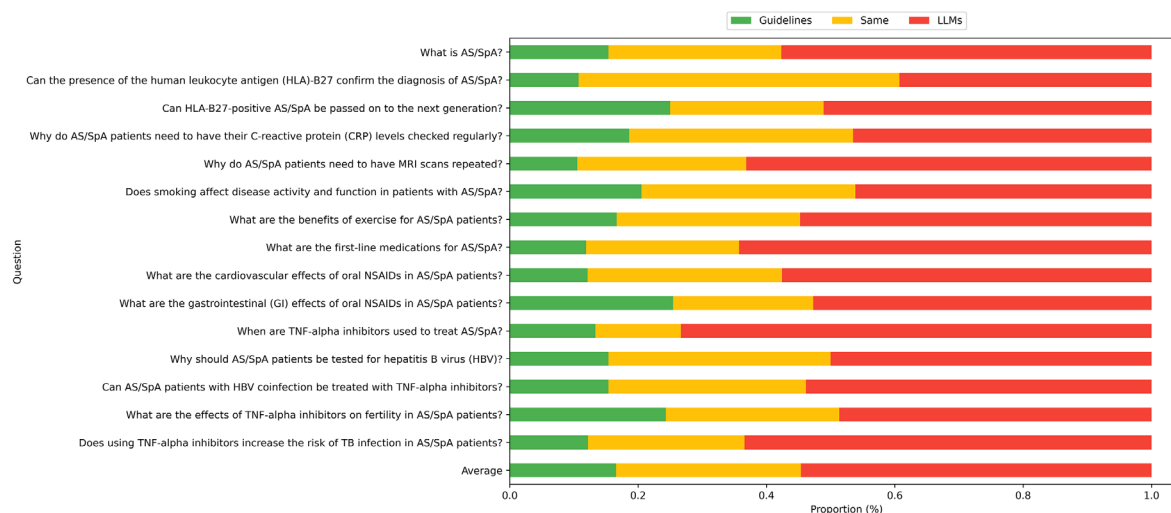
### DISCUSSION

#### Contextualisation within existing research

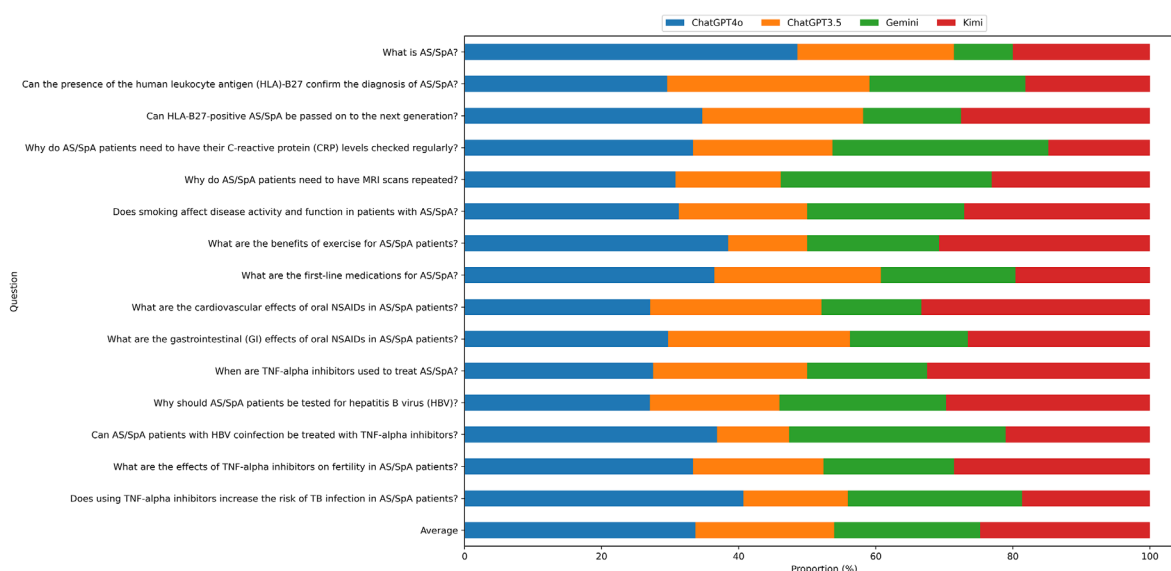
Our study contributes to the growing body of research that explores the integration of LLMs with rheumatology, particularly in the context of AS/SpA. Recent studies have demonstrated the potential of LLMs in various medical fields, including rheumatology, for tasks such as patient education, diagnosis and treatment planning. For instance, Venerito *et al* applied LLMs to assist in the diagnosis of rheumatic conditions, highlighting the models' ability to process and interpret complex clinical data.<sup>29</sup>

Our approach aligns with these studies in leveraging LLMs to enhance patient education and engagement. However, our study uniquely focuses on the specific challenges and needs of patients with AS/SpA, providing tailored educational content that addresses their concerns and questions. Unlike previous studies that may have focused on broader rheumatic conditions, our study delves into the nuances of AS/SpA, offering detailed insights into disease management, treatment options and quality of life issues specific to this patient population.

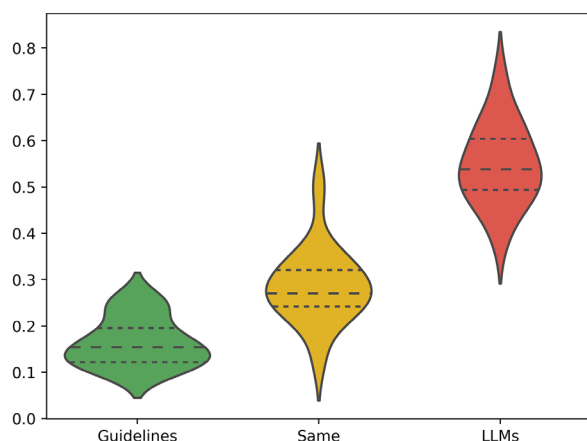
A



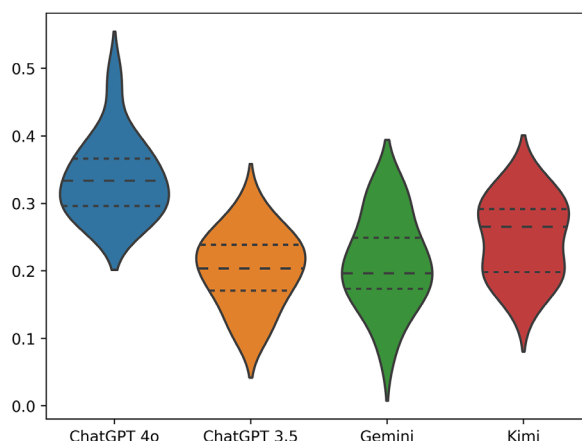
B



C



D



**Figure 3** Evaluation of patients' acceptance of large language model (LLM) responses. (A and C) A 100% stacked horizontal bar chart and a violin plot are used to visualise the proportions and distributions of patients who rated the answers from the Guidelines, LLMs or both as better, for each of the 15 questions. (B and D) A 100% stacked horizontal bar chart and a violin plot are used to visualise the proportions and distributions of patients who rated the answers from different LLMs (ChatGPT-4o, ChatGPT-3.5, Gemini 1.5 Flash and Kimi) as better, for each of the 15 questions. The median in the violin plot is marked by a long dashed line, while the shorter dashed lines delineate the IQR. AS/SpA, ankylosing spondylitis/spondyloarthritis; NSAIDs, non-steroidal anti-inflammatory drugs; TB, tuberculosis; TNF, tumour necrosis factor.

### Qualitative insights into model performance

The superior performance of specific LLMs, such as ChatGPT-4o, can be attributed to several factors related to their training data diversity and algorithmic design. ChatGPT-4o, developed by OpenAI, is trained on a vast and diverse dataset that includes a wide range of medical literature, patient forums and clinical guidelines. These diverse training data enable the model to generate responses that are not only accurate but also contextually relevant and easily understandable for patients. The model's advanced algorithmic design, which incorporates deep learning techniques such as the transformer architecture, further enhances its ability to process and generate high-quality text. ChatGPT-4o's algorithmic design emphasises the generation of clear and concise responses, making it particularly effective in providing understandable and acceptable answers to patients with AS/SpA.

### Practical implications for clinical workflows

Integrating LLMs into clinical workflows offers significant benefits, including enhanced patient education and improved efficiency in managing patient inquiries. LLMs can provide accurate, relevant and easily understandable information, which can boost patient engagement and adherence to treatment plans, leading to better health outcomes. For instance, LLMs can generate personalised educational materials that address specific patient concerns, making the information more relatable and actionable. Additionally, LLMs can assist healthcare providers by generating high-quality responses to common patient questions, reducing the burden on healthcare staff and allowing them to focus on more complex tasks, thereby improving workflow efficiency and resource allocation.

### Limitations

While our study demonstrates the potential of LLMs in improving patient education for AS/SpA, it is not without limitations. One significant limitation is the occurrence of hallucinations, where the LLMs may generate responses that are not entirely accurate or relevant to the specific context of AS/SpA. This highlights the need for continuous refinement and validation of the models to ensure the reliability and accuracy of the information provided.

Additionally, our study acknowledges the potential for biases in the LLMs' responses. These biases may stem from the training data, which may not be representative of diverse patient populations. For instance, the models may exhibit biases related to age, gender or treatment choices, potentially affecting the relevance and applicability of the generated responses for different demographic groups. Future research should focus on addressing these biases by incorporating more diverse and representative training data to enhance the models' performance across various patient demographics.

Furthermore, the study's findings are limited by the relatively small sample size and the specific geographical

region from which the participants were recruited. This may affect the generalisability of the results to other regions or populations. Future studies should consider larger and more diverse samples to validate the findings and explore the potential of LLMs in different contexts. Given that patients were recruited via the internet, self-selection bias may have been introduced, potentially affecting the generalisability of our findings. The validation of the scientific accuracy of the educational content was conducted by a relatively small sample of four rheumatologists. The limited number of experts involved in the validation process could potentially introduce bias and may affect the generalisability of the results. Future studies should consider increasing the number of experts to enhance the robustness and reliability of the validation process.

Finally, due to the voluntary nature of participation, there may be self-selection bias. For example, patients who are more interested in disease knowledge and more willing to participate in surveys may be more likely to participate, which could affect the representativeness of the sample.

In conclusion, our study highlights the promising role of LLMs in enhancing patient education for AS/SpA. By providing accurate, relevant and easily understandable information, LLMs can significantly improve patient engagement and adherence to treatment plans. However, the study also underscores the need for addressing limitations such as hallucinations and potential biases to fully realise the potential of LLMs in medical practice.

### CONCLUSION

This study demonstrates that LLMs have significant potential for application in health education for patients with AS/SpA. They can provide accurate, professional and easily understandable medical information, helping to improve patient adherence to treatment and quality of life. Although the application of LLMs in the medical field still faces certain challenges, their advantages in medical science popularisation and patient education make them a promising tool for future medical practice. Through further research and optimisation, LLMs have the potential to be widely applied in many medical fields, ultimately providing patients with higher quality medical services.

### Author affiliations

<sup>1</sup>Pazhou Lab, Guangzhou, Guangdong, China

<sup>2</sup>The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, Guangdong, China

<sup>3</sup>Department of Rheumatology, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, Guangdong, China

<sup>4</sup>School of Automation Science and Engineering, South China University of Technology, Guangzhou, China

<sup>5</sup>Research Center for Brain-Computer Interface, Pazhou Lab, Guangzhou, Guangdong, China

<sup>6</sup>Department of Rheumatology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, Guangdong, China



**Acknowledgements** We thank the patients and investigators who participated in the study as well as the authors of the 'Practical Guidelines for Patients with Ankylosing Spondylitis/ Spondyloarthritis'.

**Contributors** YR, Y-nK, S-yC, YL, JG and QL had full access to all data of this study and take responsibility for data integrity and accuracy of the analysis. Study concept and design: YR, Y-nK, S-yC, YL, JG and QL. Data acquisition: YR, S-yC, YC, FM, LX, RL, SW, HL and WX. Answer evaluation: Y-nK, JZ, XL and JW. Data analysis: YR, YL and QL. Interpretation of data: YR, YL and QL. Drafting the manuscript: YR, Y-nK, S-yC and QL. All authors read and approved the manuscript. All authors reviewed the manuscript. Guarantor is QL.

**Funding** The authors acknowledge funding received from the Shenzhen Science and Technology Program (No. JCYJ20220530145001002).

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Ethics approval** This study involves human participants. The study complied with the Declaration of Helsinki and was approved by the Ethics Committee of the Seventh Affiliated Hospital of Sun Yat-Sen University, Shenzhen, Guangdong Province, China. The ethical approval number is KY-2024-244-02. Participants gave informed consent to participate in the study before taking part.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. All data relevant to the study are included in the article or uploaded as supplementary information.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iDs

Yong Ren <http://orcid.org/0000-0002-4205-6964>  
Wenqi Xia <http://orcid.org/0009-0002-7931-8056>

## REFERENCES

- Boonen A, Chorus A, Miedema H, *et al.* Employment, work disability, and work days lost in patients with ankylosing spondylitis: a cross sectional study of Dutch patients. *Ann Rheum Dis* 2001;60:353–8.
- McGonagle D, David P, Macleod T, *et al.* Predominant ligament-centric soft-tissue involvement differentiates axial psoriatic arthritis from ankylosing spondylitis. *Nat Rev Rheumatol* 2023;19:818–27.
- Navarro-Compán V, Sepriano A, El-Zorkany B, *et al.* Axial spondyloarthritis. *Ann Rheum Dis* 2021;80:1511–21.
- Braun J, Sieper J. Ankylosing spondylitis. *Lancet* 2007;369:1379–90.
- Ciurea A, Scherer A, Weber U, *et al.* Age at symptom onset in ankylosing spondylitis: is there a gender difference? *Ann Rheum Dis* 2014;73:1908–10.
- Sieper J, Poddubnyy D. Axial spondyloarthritis. *Lancet* 2017;390:73–84.
- Ward MM, Deodhar A, Gensler LS, *et al.* 2019 Update of the American College of Rheumatology/Spondylitis Association of America/Spondyloarthritis Research and Treatment Network Recommendations for the Treatment of Ankylosing Spondylitis and Nonradiographic Axial Spondyloarthritis. *Arthritis Rheumatol* 2019;71:1599–613.
- Braun J, Davis J, Dougados M, *et al.* First update of the international ASAS consensus statement for the use of anti-TNF agents in patients with ankylosing spondylitis. *Ann Rheum Dis* 2006;65:316–20.
- Hu Y, Lai Y, Liao F, *et al.* Assessing Accuracy of CHATGPT on Addressing *Helicobacter pylori* Infection-Related Questions: A National Survey and Comparative Study. *Helicobacter* 2024;29:e13116.
- Ramiro S, Nikiphorou E, Sepriano A, *et al.* ASAS-EULAR recommendations for the management of axial spondyloarthritis: 2022 update. *Ann Rheum Dis* 2023;82:19–34.
- Zochling J, van der Heijde D, Burgos-Vargas R, *et al.* ASAS/EULAR recommendations for the management of ankylosing spondylitis. *Ann Rheum Dis* 2006;65:442–52.
- Darve A, Deodhar A. Treatment of axial spondyloarthritis: an update. *Nat Rev Rheumatol* 2022;18:205–16.
- Coulter EH, McDonald MT, Cameron S, *et al.* Physical activity and sedentary behaviour and their associations with clinical measures in axial spondyloarthritis. *Rheumatol Int* 2020;40:375–81.
- Dagfinrud H, Kvien TK, Hagen KB. Physiotherapy interventions for ankylosing spondylitis. *Cochrane Database Syst Rev* 2008;2008:CD002822.
- Fabre S, Molto A, Dadoun S, *et al.* Physical activity in patients with axial spondyloarthritis: a cross-sectional study of 203 patients. *Rheumatol Int* 2016;36:1711–8.
- O'Dwyer T, O'Shea F, Wilson F. Exercise therapy for spondyloarthritis: a systematic review. *Rheumatol Int* 2014;34:887–902.
- Passalent LA, Soever LJ, O'Shea FD, *et al.* Exercise in ankylosing spondylitis: discrepancies between recommendations and reality. *J Rheumatol* 2010;37:835–41.
- Ward MM, Deodhar A, Gensler LS, *et al.* 2019 Update of the American College of Rheumatology/Spondylitis Association of America/Spondyloarthritis Research and Treatment Network Recommendations for the Treatment of Ankylosing Spondylitis and Nonradiographic Axial Spondyloarthritis. *Arthritis Care Res (Hoboken)* 2019;71:1285–99.
- Levitova A, Hulejova H, Spiritovic M, *et al.* Clinical improvement and reduction in serum calprotectin levels after an intensive exercise programme for patients with ankylosing spondylitis and non-radiographic axial spondyloarthritis. *Arthritis Res Ther* 2016;18:275.
- Huo B, Cacciamani GE, Collins GS, *et al.* Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med* 2023;29:2988.
- Liang H, Tsui BY, Ni H, *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* 2019;25:433–8.
- Moor M, Banerjee O, Abad ZSH, *et al.* Foundation models for generalist medical artificial intelligence. *Nature New Biol* 2023;616:259–65.
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al.* Large language models in medicine. *Nat Med* 2023;29:1930–40.
- Omar M, Watad A, McGonagle D, *et al.* The role of deep learning in diagnostic imaging of spondyloarthropathies: a systematic review. *Eur Radiol* 2024.
- Omar M, Naffaa ME, Glicksberg BS, *et al.* Advancing rheumatology with natural language processing: insights and prospects from a systematic review. *Rheumatol Adv Pract* 2024;8:rkae120.
- Mannstadt I, Mehta B. Large language models and the future of rheumatology: assessing impact and emerging opportunities. *Curr Opin Rheumatol* 2024;36:46–51.
- Venerito V, Gupta L. Large language models: rheumatologists' newest colleagues? *Nat Rev Rheumatol* 2024;20:75–6.
- Omar M, Agbareia R, Klang E, *et al.* Large Language Models in Rheumatologic Diagnosis: A Multimodal Performance Analysis. *J Rheumatol* 2024.;jrheum.
- Venerito V, Bilgin E, Iannone F, *et al.* AI am a rheumatologist: a practical primer to large language models for rheumatologists. *Rheumatology (Oxford)* 2023;62:3256–60.
- Goodman RS, Patrinely JR Jr, Osterman T, *et al.* On the cusp: Considering the impact of artificial intelligence language models in healthcare. *Med* 2023;4:139–40.
- Sezgin E, Sirrianni J, Linwood SL. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. *JMIR Med Inform* 2022;10:e32875.
- Steimetz E, Minkowitz J, Gabutan EC, *et al.* Use of Artificial Intelligence Chatbots in Interpretation of Pathology Reports. *JAMA Netw Open* 2024;7:e2412767.
- Kerbage A, Kassab J, El Dahdah J, *et al.* Accuracy of ChatGPT in Common Gastrointestinal Diseases: Impact for Patients and Providers. *Clin Gastroenterol Hepatol* 2024;22:1323–5.
- Lu MY, Chen B, Williamson DFK, *et al.* A multimodal generative AI copilot for human pathology. *Nature New Biol* 2024;634:466–73.
- Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* 2018;320:2199–200.

- 36 Mehandru N, Miao BY, Almaraz ER, *et al.* Evaluating large language models as agents in the clinic. *NPJ Digit Med* 2024;7:84.
- 37 Hager P, Jungmann F, Holland R, *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30:2613–22.
- 38 Wankmüller S. A comparison of approaches for imbalanced classification problems in the context of retrieving relevant documents for an analysis. *J Comput Soc Sci* 2023;6:91–163.
- 39 Xie Y, Yang KH, Lyu Q, *et al.* Practice guideline for patients with ankylosing spondylitis/spondyloarthritis. *Zhonghua Nei Ke Za Zhi* 2020;59:511–8.
- 40 van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheum* 1984;27:361–8.