Clinical, Cosmetic and Investigational Dermatology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluating Large Language Models for Accuracy and Completeness of Vitiligo Patient Education: A Comparative Analysis

Jieyan Su[1], Xi Yang[1], Xiangying Li[1], Jiaxuan Chen[2], Caixin Jiang[2], Yi Wang[1], Le Zhuang[1,*], Hang Li[3–5,*]

[1]Department of Dermatology, Central Hospital Affiliated to Shandong First Medical University, Jinan, People's Republic of China; [2]School of Clinical Medicine, Shandong Second Medical University, Weifang, People's Republic of China; [3]Department of Dermatology, Peking University First Hospital, Beijing, People's Republic of China; [4]National Clinical Research Center for Skin and Immune Diseases, Beijing, People's Republic of China; [5]NMPA Key Laboratory for Quality Control and Evaluation of Cosmetics, Beijing, People's Republic of China

*These authors contributed equally to this work

Correspondence: Hang Li, Peking University First Hospital, No. 8, Xishiku Street, Xicheng District, Beijing, 100034, People's Republic of China, Tel +8613693058190, Fax +8601083572350, Email drlihang@126.com; Le Zhuang, Central Hospital Affiliated to Shandong First Medical University, No. 105, Jiefang Road, Jinan, 250013, People's Republic of China, Tel +8615966301378, Fax +86053155739999, Email zhuangle@sdu.edu.cn

Background: Vitiligo causes significant psychological stress, creating a strong demand for accessible educational resources beyond clinical settings. This demand remains largely unmet. Large language models (LLMs) have the potential to bridge this gap by enhancing patient education. However, uncertainties exist regarding their ability to accurately address individualized patient inquiries and whether comprehension capabilities vary between LLMs.

Purpose: This study aims to evaluate the applicability, accuracy, and potential limitations of OpenAI o1, DeepSeek-R1, and Grok 3 for vitiligo patient education.

Methods: Three dermatology experts first developed sixteen vitiligo-related questions based on common patient concerns, which were categorized as descriptive or recommendatory with basic and advanced levels. The responses from the three LLMs were then evaluated by three vitiligo-specialized dermatologists for accuracy, comprehensibility, and relevance using a Likert scale. Additionally, three patients rated the comprehensibility of the responses, and a readability analysis was performed.

Results: All three LLMs demonstrated satisfactory accuracy, comprehensibility, and completeness, although their performance varied. They achieved 100% accuracy in responding to basic descriptive questions but exhibited inconsistency when addressing complex recommendatory queries, particularly regarding treatment recommendations for specific populations. Pairwise comparisons indicated that DeepSeek-R1 outperformed OpenAI o1 in accuracy scores ($p = 0.042$), while no significant difference was observed compared to Grok 3 ($p = 0.157$). Readability assessments revealed elevated reading difficulty across all models, with DeepSeek-R1 exhibiting the lowest readability (mean Flesch Reading Ease score of 19.7; pairwise comparisons showed DeepSeek-R1 scores were significantly lower than those of OpenAI o1 and Grok 3, both $p < 0.01$), potentially reducing accessibility for diverse patient populations.

Conclusion: Reasoning-LLMs demonstrate high accuracy in responding to simple vitiligo-related questions, but the quality of treatment recommendations declines as question complexity increases. Current models exhibit errors in providing vitiligo treatment advice, necessitating enhanced filtering mechanisms by developers and mandatory human oversight for medical decision-making.

Plain Language Summary: This study looked at how well three advanced chatbots—OpenAI o1, DeepSeek R1, and Grok 3—answer questions about vitiligo, a skin condition that causes patches of skin to lose color. Vitiligo can be stressful, and patients often need clear, accurate information at home. We tested these chatbots to see if they could provide reliable and easy-to-understand answers. Three skin experts created 16 questions about vitiligo, covering basic facts and treatment advice. The chatbots' answers were rated by experts and patients for accuracy, clarity, and relevance. All three chatbots did well overall, scoring high on accuracy and completeness, especially for simple questions. DeepSeek R1 was the most accurate, while OpenAI o1 and Grok 3 were easier to read. However, the chatbots sometimes gave wrong advice, especially about treatments for specific groups such as children or pregnant

Received: 10 July 2025
Accepted: 30 September 2025
Published: 23 October 2025

women. Some answers could be hard to read for some users. The study shows that these chatbots can help educate people about vitiligo, especially in areas with limited access to doctors. But they are not perfect and cannot replace expert medical advice. Improvements are needed to make their answers more accurate and easier to understand. In the future, chatbots could support doctors by providing patients with reliable information, but they should not be the main source of medical guidance.

Keywords: large language models, ChatGPT, DeepSeek, Grok, vitiligo, patient education

Introduction

Vitiligo is a prevalent chronic acquired condition characterized by the loss of skin pigmentation.[1] Over half of individuals with vitiligo experience psychological disorders, which significantly impair their quality of life.[2,3] Patients and their families frequently express questions and concerns about the disease. Currently, patient education predominantly relies on online educational websites, support groups, community activities, and social media platforms. However, these approaches are limited by inconsistent information quality and challenges in ensuring accuracy. Alternative methods, such as face-to-face consultations with physicians, institutional brochures, and lectures, are constrained by geographic and temporal limitations, rendering them less accessible as on-demand educational resources. These shortcomings underscore the urgent need for dependable educational tools beyond clinical settings.[4] Large Language Models (LLMs) have gained attention across various medical domains due to their versatility, driven by extensive parameter spaces that enhance their utility in medical patient education.[5] Dermatology is a medical specialty that heavily relies on visual assessment, patient communication, and complex decision-making. LLMs demonstrate exceptional performance in image recognition, aligning seamlessly with dermatology's emphasis on visual diagnosis. Early detection of skin cancer represents the most mature and widely studied application of artificial intelligence (AI) in this field. LLMs have shown impressive capabilities in providing clinical decision support and assisting with medical writing. A study by Iqbal et al[6] found that LLMs, utilizing medical databases from Taiwan and the United States, generated medication recommendations highly consistent with those of dermatologists, effectively optimizing clinical workflows. However, due to the diverse nature of dermatological conditions and individual patient variability, clinical decision-making schemes provided by ChatGPT exhibit limitations in subspecialty areas. Dunn et al[7] compared the quality of ChatGPT-generated dermatology case reports with published reports, demonstrating through blinded expert review that ChatGPT-generated texts surpassed human-written reports in both quality and readability.

Dermatology patients frequently have questions regarding their conditions, treatments, and care plans. Traditional face-to-face communication models are time-intensive for healthcare providers. The advent of LLMs offers a novel approach to patient education. Ayers et al[8] highlighted the ability of LLMs to deliver high-quality, empathetic responses to patient inquiries on social media, achieving higher patient satisfaction compared to physician responses. Similarly, Michelle et al[9] reported that patients preferred responses generated by large language models over those prepared by the American Society for Mohs Surgery.

The application of LLMs in dermatology is rapidly expanding. Several commercial AI-based dermatology diagnostic applications, such as SkinVision and MoleMapper, have entered the market. LLMs are enhancing global patient education by extending dermatological knowledge and preliminary diagnostic capabilities to resource-limited regions, thereby reducing geographical and economic barriers.[10] LLMs can generate tailored health education materials based on patients' specific conditions, cultural backgrounds, and educational levels. Through intelligent applications, patients can access real-time skin care advice and learn to accurately identify and manage common dermatological issues. The accessibility and personalization of LLMs address deficiencies in clinical patient education. Reasoning-LLMs, a new generation of LLMs, differ from traditional LLMs, which rapidly generate outputs based solely on statistical guesses of the next word. Instead, reasoning-LLMs take time to break down problems into individual steps, deriving more accurate answers through "thought" and self-verification,[11] making them highly suitable for complex, precision-demanding tasks such as medical education.

Chatbot Arena, an open platform by LMSYS, evaluates LLMs using paired, anonymized comparisons and Elo scoring based on human preferences.[12] From its leaderboard, we selected three top-ranking reasoning LLMs—OpenAI

o1, DeepSeek-R1, and Grok3—for this study. To date, no research has assessed the efficacy of LLMs in addressing vitiligo-related questions. With the increasing number and rapid evolution of LLMs, there is also a lack of studies evaluating whether significant differences exist among them in the context of vitiligo. This study aims to fill this gap by evaluating the accuracy, completeness, and comprehensibility of different LLMs, thereby assisting vitiligo patients in selecting tools for medical inquiries and enhancing patient education and dermatological support.

Materials and Methods

This cross-sectional study received approval from the Research Ethics Committee of Jinan Central Hospital (Registration No. 20250530023). All participants provided informed consent, in accordance with the Declaration of Helsinki.

Material Sources and Processing

Drawing on established guidelines,[13,14] prior literature,[15–18] and clinical expertise, a questionnaire comprising 16 vitiligo-related questions was developed by two experienced dermatologists (Hang Li and Le Zhuang), each with more than 10 years of expertise in the field of vitiligo. The questions used include 8 questions aimed at describing the topic of vitiligo and 8 questions aimed at providing advice related to vitiligo. The questions could each be divided into "basic" or "advanced" content. "Basic" included simple questions in terms of content, and "advanced" included questions with specific inquiries in the field. The expert panel consulted a group of five patients with vitiligo, recruited from the vitiligo clinic of our hospital, and two other dermatologists from the Department of Dermatology to validate the generated questions. Their feedback was incorporated to refine the questions further and ensure that the questions effectively captured the perspectives and information needs of the patient community. The questions used to prompt LLMs are listed in Box S1 in the Supporting Information.Three Large Language Models (LLMs)—OpenAI o1 (OpenAI, San Francisco, California, USA), DeepSeek R1 (DeepSeek, Hangzhou, Zhejiang, China), and Grok3 (xAI, San Francisco, California, USA)—were selected to assess the performance of AI-generated content. In February 2025, a series of questions were individually submitted by Jiaxuan Chen to three distinct chatbot platforms. To ensure response consistency and independence, each question was entered separately using the "new chat" function provided by the respective platform interfaces. Responses generated by each LLM for every question were systematically documented. Each LLM generated one detailed answer for each question, resulting in 48 answers (16 questions×3 LLMs). To maintain objectivity, all responses were converted into plain text format, and the presentation of the materials was standardized using uniform font type and size. No additional modifications to the format were applied.

Assessment of Materials

Each response was assessed from both patient and professional perspectives. A panel of three dermatologists (Xi Yang, Yi wang, and Jieyan Su), each with over five years of experience in vitiligo, who were not involved in the question generation process, assessed the accuracy, completeness, and relevance of the responses using a 5-point Likert scale. Additionally, similar to previous report,[8] three patients with vitiligo, who had not participated in question generation, contributed to the evaluation. These patients rated the comprehensibility using a similar 5-point Likert scale. Participating vitiligo patients were required to be aged 18 years or older, possess proficient English reading and comprehension skills, and successfully complete a simplified Flesch-Kincaid reading comprehension test designed for a 10th- to 12th-grade level. Likert scale definitions are provided in Table 1. To ensure that the information was presented in an accessible manner, readability was further evaluated using the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE) metrics.[19] The FRE scale, ranging from 0 to 100, measures text readability, with higher scores indicating greater ease of comprehension and lower scores reflecting increased difficulty.[20] The FKGL estimates the educational level required to understand the text, with higher scores denoting greater complexity.[21]

Data Processing

The final score for each response was determined as follows: if all three raters assigned identical scores, that score was adopted. If two raters provided the same score, their agreed-upon score was used. In cases where all three raters assigned different scores, the lowest score was selected to ensure a conservative evaluation.

Table 1 Medical Experts and Patients Evaluation Scores Criteria

Evaluation Dimension	Score	Likert Scale Definitions
Accuracy	1	The response contains serious errors or misleading information potentially harmful to patients
	2	The response includes minor errors or inaccuracies without evident harm
	3	The response is mostly accurate but contains some ambiguous or uncertain statements
	4	The response is accurate, clearly articulated, and free of obvious errors
	5	The response is highly accurate, professionally stated, and fully aligned with current medical knowledge
Completeness	1	The response is very brief, lacking key information and offering minimal assistance
	2	The response addresses some relevant content but omits significant details, limiting its utility
	3	The response covers primary content but misses some important aspects, resulting in incomplete guidance
	4	The response includes most key content, with minor omissions, and is highly useful
	5	The response is comprehensive, addressing all key information and fully answering the question
Relevance	1	The response is almost entirely unrelated to the question, lacking focus
	2	The response contains some relevant content but largely deviates from the topic
	3	The response is mostly relevant, with minor off-topic elements
	4	The response closely aligns with the question, with nearly all content pertinent despite slight irrelevancies
	5	The response is fully relevant, with all content precisely targeted to the question
Comprehensibility	1	Completely incomprehensible, with content entirely unclear
	2	Difficult to understand, requiring significant clarification
	3	Partially comprehensible, with some content clear but confusion remaining
	4	Mostly comprehensible, with minor doubts persisting
	5	Fully comprehensible, enabling accurate understanding and resolution of all doubts

Statistical Analysis

Descriptive analyses were conducted to evaluate the accuracy, completeness, relevance, and comprehensibility of the LLM responses. Pairwise comparisons within the Friedman test were conducted to assess differences between models. Inter-rater agreement was assessed using Kendall's W coefficient of concordance. The coefficient value (W) below 0.4 suggested poor agreement, and moderate agreement had the value between 0.4 and 0.59, high agreement with 0.6–0.79, excellent agreement with value ≥0.8.[22] Data analyses were performed by SPSS Statistics 27.0 (IBM Corp., Armonk, NY, USA) and GraphPad Prism 10.4.0 (GraphPad Software, San Diego, CA, USA). Data recording and organization were facilitated by Microsoft Excel 2021. $p < 0.05$ was considered significant.

Results

Categorization of the Questions and Agreement Between Reviewers

Table S1 in the Supporting Information provides the scores assigned to the responses. Inter-rater reliability, evaluated using Kendall's coefficient of concordance (W), was high for both the patient group (W = 0.653) and the physician group (W = 0.772), indicating strong agreement among evaluators. Responses with scores exceeding 4 were classified as accurate. All three LLMs—OpenAI o1, DeepSeek-R1, and Grok3—achieved 100% accuracy on descriptive questions. For recommendatory questions, accuracy rates were 75% for OpenAI o1, 87.5% for DeepSeek-R1, and 75% for Grok3.

Accuracy Assessment

Accuracy reflects the precision and verifiability of the facts, data, and knowledge presented in the responses. Median accuracy scores were 4 for OpenAI o1 and 5 for both DeepSeek-R1 and Grok3, with DeepSeek-R1's first quartile surpassing that of Grok3 (Table 2). The Friedman test revealed a statistically significant difference between OpenAI o1 and DeepSeek-R1 ($p = 0.042$), whereas no significant differences were observed between Grok3 and OpenAI o1 or between Grok3 and DeepSeek-R1 ($p = 0.157$, $p = 0.536$, respectively) (Figure 1a).

Table 2 Median Scores for Answers From Three Large Language Models

Groups	Items	OpenAI o1	DeepSeek R1	Grok3 (Beta)
Expert assessment	Accuracy, median (Q1, Q3) Completeness, median (Q1, Q3) Relevance, median (Q1, Q3)	4(4, 5) 4(4, 4.25) 5(5, 5)	5(5, 5) 5(4.75, 5) 5(5, 5)	5(4, 5) 4(4, 5) 5(5, 5)
Patient assessment	Comprehensibility, median (Q1, Q3)	5(5, 5)	5(4.75, 5)	5(5, 5)
Objective evaluation	FRE score, median (Q1, Q3) Flesch-kincaid Grade Level, mean (SD)	34.1(30.15, 41.65) 13.35(12.5, 13.95)	17.95(12.58, 23.88) 13.45(12.8, 14.7)	37.4(25.08, 46.5) 12.05(10.1, 13.73)

Abbreviations: LLMs, Large language models; FRE, Flesch Reading Ease; FKGL, Flesch-kincaid Grade Level; Q1, First quartile; Q3, Third quartile.

Specific inaccuracies were noted. DeepSeek-R1 erroneously recommended a mid-potency topical glucocorticoid (eg, triamcinolone 0.1%) for vitiligo, whereas global experts in vitiligo diagnosis and management recommend potent to very potent topical corticosteroids.[20] This error is also commonly observed among dermatologists not specializing in vitiligo. Meanwhile, OpenAI o1 and Grok3 incorrectly deemed narrowband ultraviolet B (NB-UVB) phototherapy unsafe during pregnancy, despite evidence establishing NB-UVB as a safe treatment option for various dermatological conditions during pregnancy.[23,24] These errors underscore the limitations of LLMs in addressing highly specialized queries.

Completeness Assessment

Table 2 indicates that the median completeness scores for all three models exceeded 4, reflecting their ability to deliver comprehensive analyses and integrate diagnostic and therapeutic information. Median completeness scores were 5 for

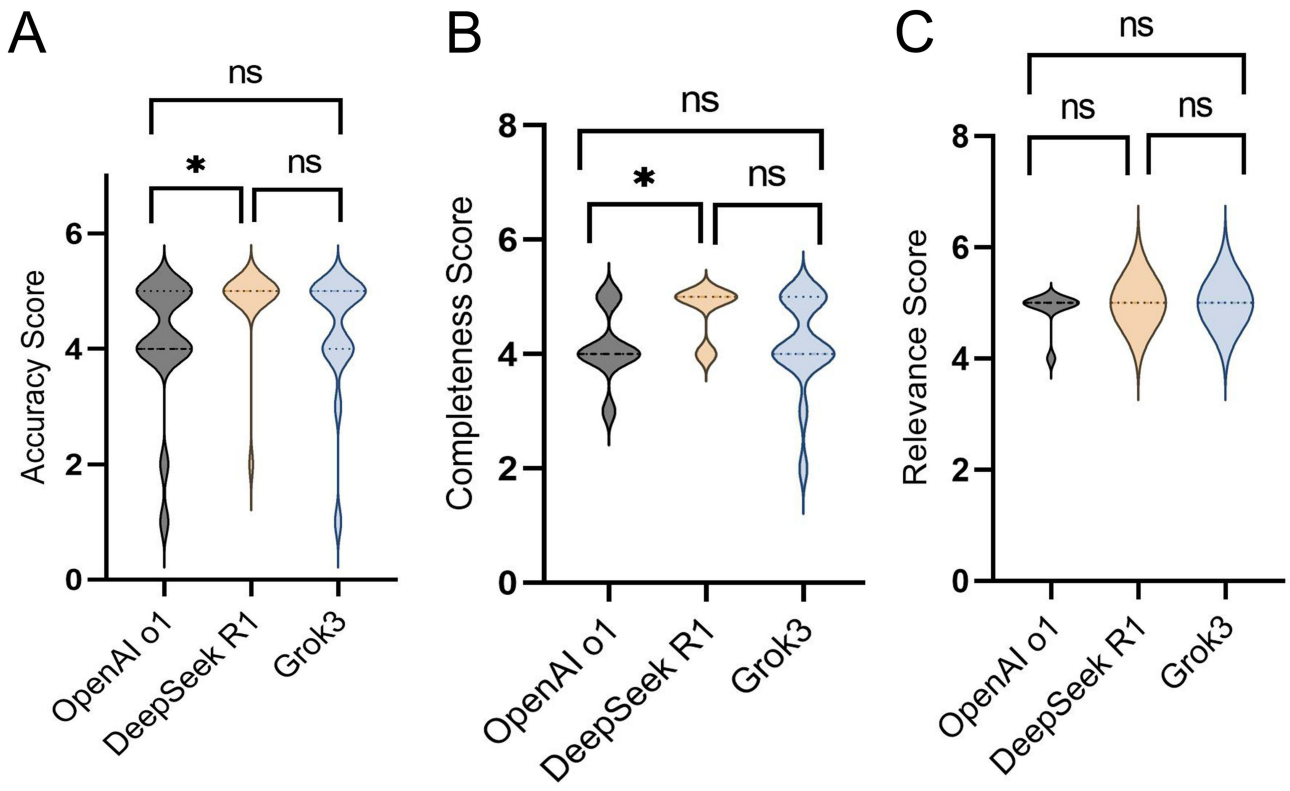

Figure 1 Comparison of Medical Expert Evaluation Scores for Vitiligo Assessment Across Large Language Models. (**A**) Accuracy scores showed a statistically significant difference between OpenAI o1 and DeepSeek-R1 (p = 0.042), but not between Grok3 and OpenAI o1 (p = 0.157) or Grok3 and DeepSeek-R1 (p = 0.536). (**B**) Completeness scores indicated a significant difference between OpenAI o1 and DeepSeek-R1 (p = 0.020), with no significant differences between Grok3 and OpenAI o1 (p = 0.052) or Grok3 and DeepSeek-R1 (p = 0.724). (**C**) Relevance scores showed no statistically significant differences among the three models (p > 0.05).

https://doi.org/10.2147/CCID.S552979

DeepSeek-R1 and 4 for both OpenAI o1 and Grok3, with Grok3's third quartile surpassing that of OpenAI o1. The Friedman test identified a statistically significant difference between OpenAI o1 and DeepSeek-R1 ($p = 0.02$), whereas no significant differences were observed between Grok3 and OpenAI o1 or between Grok3 and DeepSeek-R1 ($p = 0.052$, $p = 0.724$, respectively) (Figure 1b).

Relevance Assessment

All three LLMs achieved maximum median, first quartile, and third quartile scores of 5. No statistically significant differences were observed among the three LLMs (Friedman test, two-way analysis of variance by ranks, $p > 0.05$) (Figure 1c), indicating that their responses were consistently pertinent to the questions posed. The systematic logical reasoning capabilities of reasoning-LLMs enhance comprehension and minimize the likelihood of irrelevant outputs.

Patient Assessment

OpenAI o1 and Grok 3 achieved the highest median scores for perceived understandability (median: 5, IQR: 5–5), indicating that patients found their responses highly comprehensible. DeepSeek-R1 had a slightly lower median score of 5 (IQR: 4.75–5), suggesting marginally reduced comprehensibility. No statistically significant differences were observed among the three LLMs (OpenAI o1 vs DeepSeek R1, $p = 0.289$; OpenAI o1 vs Grok 3, $p = 0.289$; DeepSeek R1 vs Grok 3, $p = 1.000$) (Figure 2a).

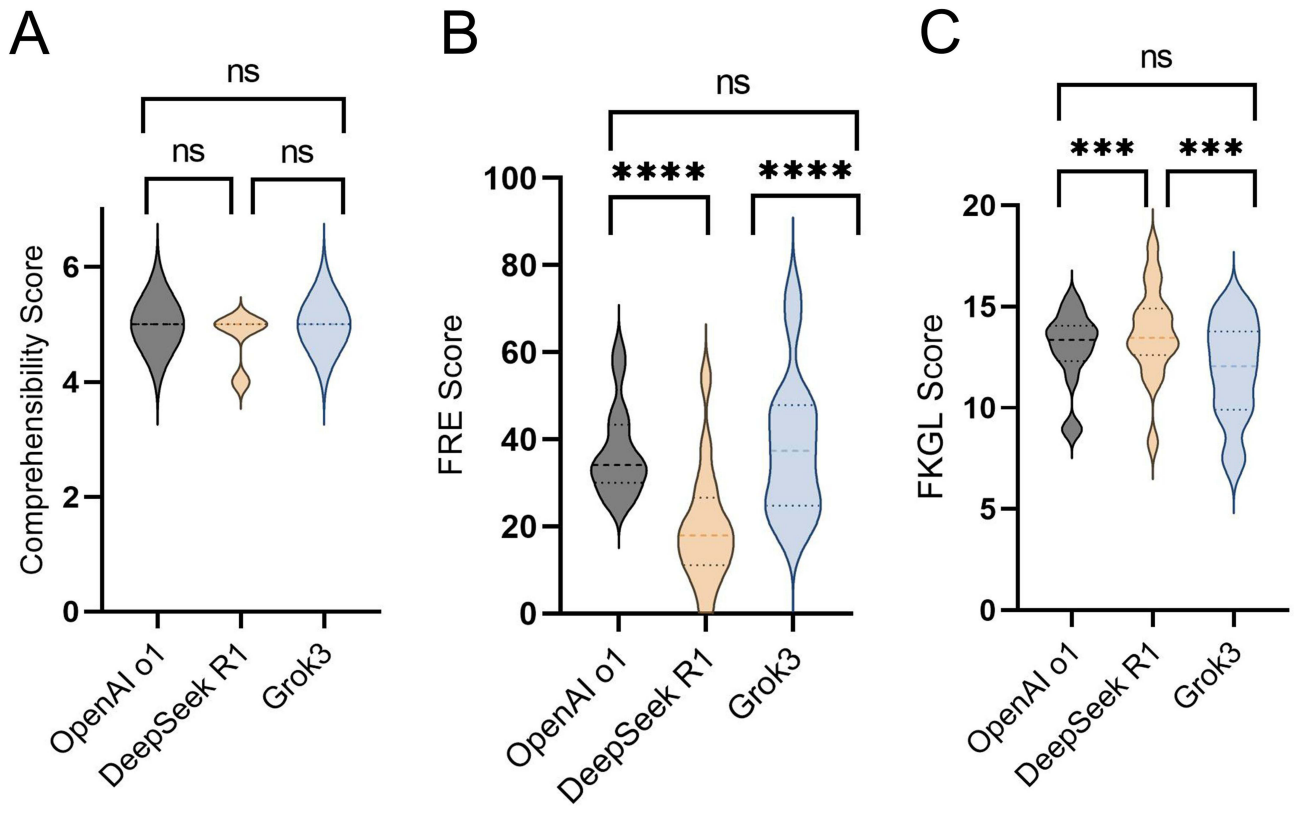

Figure 2 Comparison of Patient Evaluation and Readability Scores for Vitiligo Assessment Across Large Language Models. (**A**) Comprehensibility scores showed no statistically significant differences among the three models (OpenAI o1 vs DeepSeek-R1, $p = 0.289$; OpenAI o1 vs Grok3, $p = 0.289$; DeepSeek-R1 vs Grok3, $p = 1.000$). (**B**) Flesch Reading Ease scores revealed highly significant differences between DeepSeek-R1 and both OpenAI o1 ($p < 0.001$) and Grok3 ($p < 0.001$), with no significant difference between OpenAI o1 and Grok3 ($p = 1.000$). (**C**) Flesch-Kincaid Grade Level scores indicated significant differences between DeepSeek-R1 and both OpenAI o1 ($p = 0.030$) and Grok3 ($p = 0.020$), with no significant difference between OpenAI o1 and Grok3 ($p = 0.855$).

Readability Assessment

DeepSeek-R1 demonstrated lower readability relative to OpenAI o1 and Grok3, with a mean Flesch Reading Ease (FRE) median score of 17.95, compared to OpenAI o1 (median: 34.1) and Grok3 (median: 37.4). The Friedman test indicated highly significant differences between DeepSeek-R1 and both OpenAI o1 ($p < 0.001$) and Grok3 ($p < 0.001$), with no significant difference between OpenAI o1 and Grok3 ($p = 1.000$)(Figure 2a). Correspondingly, DeepSeek-R1's mean Flesch-Kincaid Grade Level (FKGL) median score was 13.45, higher than OpenAI o1 (median: 13.35) and Grok3 (median: 12.05). The Friedman test indicated highly significant differences between DeepSeek-R1 and both OpenAI o1 ($p = 0.03$) and Grok3 ($p = 0.02$), with no significant difference between OpenAI o1 and Grok3 ($p = 0.855$) (Figure 2b).

We constructed a summary table (Table 3) that visually compares the strengths of the three LLMs.

Comparisons Across Difficulty Levels

Analysis of responses stratified by question difficulty revealed no statistically significant differences among the three LLMs at any individual difficulty level ($p > 0.05$) (Figure 3). This suggests that no single model exhibited superior performance in addressing complex questions.

Discussion

The errors observed in the perspectives of all three large language models (LLMs) regarding the use of topical corticosteroids and phototherapy for vitiligo mirror common misconceptions held by many dermatologists not specializing in vitiligo. These inaccuracies may stem from the incorporation of erroneous information into the training datasets. Accuracy could be improved by training models with the latest clinical practice guidelines or regularly updated, domain-specific datasets.[25,26] In scenarios where questions were posed as if from a 10-year-old child, all three LLMs adjusted their readability to an FKGL of 8th grade, demonstrating commendable humanistic care. Compared to traditional LLMs, reasoning-LLMs exhibited a marked improvement in contextual awareness. Based on prior research,[26] this study designed a consensus-based scoring rule and conservative estimation for data preprocessing to minimize the impact of subjective differences among raters. Additionally, analyses using all individual scores were conducted, yielding results consistent with the consensus-based approach.Different LLMs show varied strengths in patient education for different diseases. A comparison of ChatGPT and Gemini for melanoma patient education found Gemini outperformed ChatGPT in response completeness, personalization, and readability.[27] In a study comparing ChatGPT-4 and Google Bard on common Mohs surgery patient questions, ChatGPT-4 demonstrated significantly higher accuracy and utility, matching the American College of Mohs Surgery's response accuracy.[28] For syphilis patient education, ChatGPT-4 and Claude achieved the highest accuracy, aligning with WHO standards at 92% and 89%, respectively. Perplexity and Llama 3 were less reliable, scoring between 60–70%. Errors increased in rarer conditions like tertiary and neurosyphilis, with LLMs providing outdated treatment protocols and incorrect transmission descriptions.[29]

LLMs and their applications in healthcare represent a globally significant topic in contemporary technological advancements. Patient use of AI can heighten disease awareness and promote early diagnosis.[30] Beyond serving as a source of medical information for patients, LLMs can also assist professional healthcare providers through electronic messaging.[19,31] By offering preliminary assessments and guidance on patient conditions, LLMs can support remote patient consultations, thereby increasing access to dermatological care.[32] Studies indicate widespread user enthusiasm for LLM-derived health advice, suggesting that these models can serve as an alternative information source for patients,

Table 3 Comparison of LLM Advantages

LLMs	Accuracy	Completeness	Relevance	Comprehensibility	Readability
OpenAI o1	Very good	Very good	Excellent	Excellent	Excellent
DeepSeek R1	Excellent	Excellent	Excellent	Excellent	Good
Grok3 (Beta)	Very good	Very good	Excellent	Excellent	Excellent

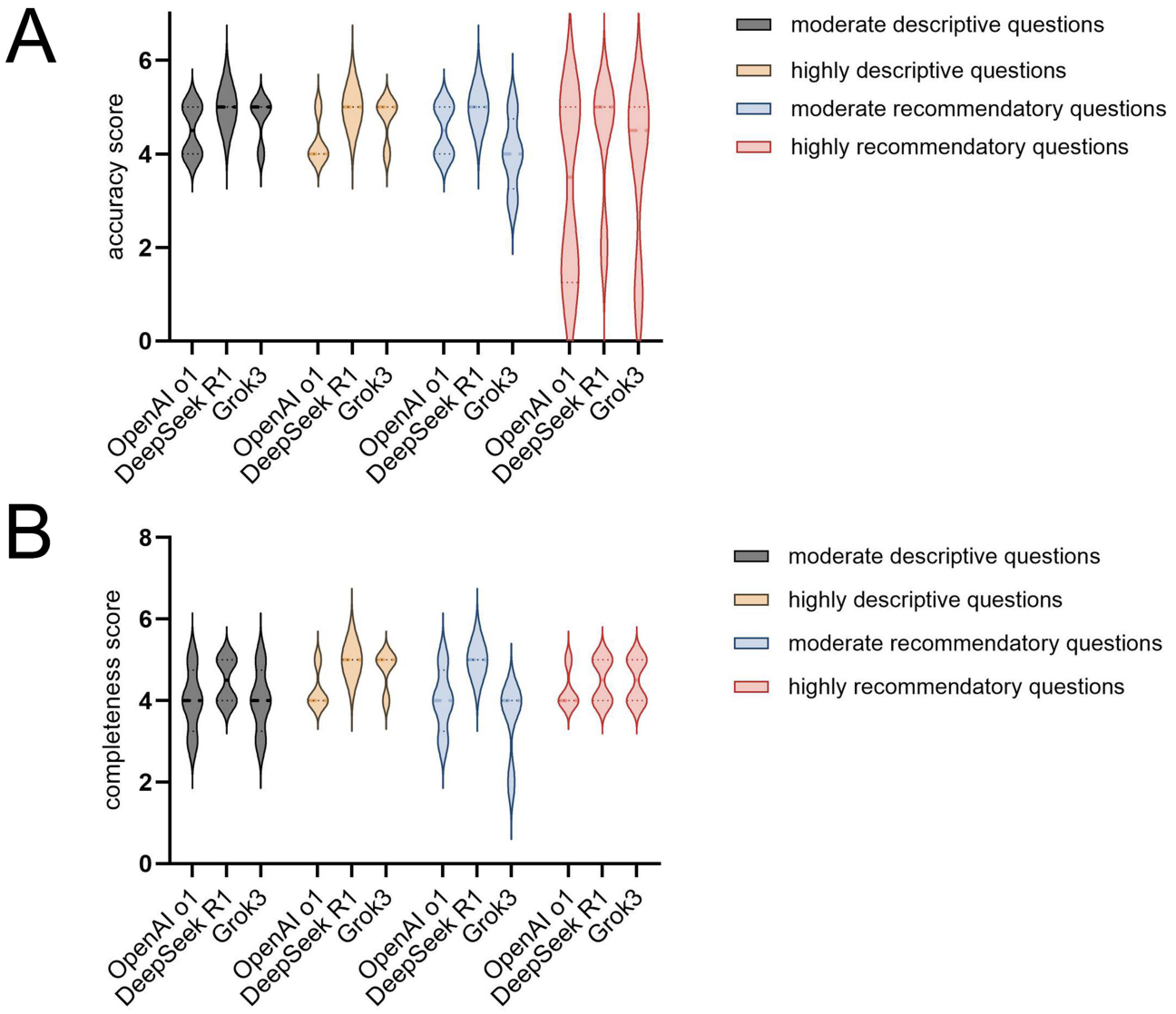

Figure 3 Performance comparison of three large language models across difficulty levels. No statistically significant differences were observed among the models at any individual level ($p > 0.05$). (**A**) Accuracy scores across difficulty levels. (**B**) Completeness scores across difficulty levels.

particularly when direct contact with a physician is unavailable.[33,34] In regions where access to specialized dermatologists is limited, LLMs can provide timely medical advice, bridging healthcare disparities in rural and underserved areas while offering initial guidance on dermatological issues.[10] OpenAI exhibits robust reasoning capabilities and multimodal support, enabling it to process both text and images. This allows it to deliver intuitive educational content for patients, such as explaining the symptoms and treatment options for vitiligo, potentially supplemented by images illustrating skin changes. Such multimodal functionality is particularly valuable for helping patients comprehend the visual manifestations of the disease and the effects of treatment. In contrast, Grok 3, developed by xAI, offers distinct advantages through its real-time data access and deep integration with the X platform. It can provide the latest updates on vitiligo treatment advancements, research developments, and relevant discussions from social media. This capability is especially beneficial for patients seeking current medical insights and community support, which are critical in managing a condition like vitiligo that may impact mental health, as social support can significantly alleviate anxiety. Meanwhile, DeepSeek-R1, an open-source and freely accessible model, has a running cost approximately 13 times lower than comparable OpenAI models.[35] Designed to prioritize deep reasoning and complex problem-solving, it demonstrated superior accuracy and completeness in this study. Optimized for Chinese-language contexts, DeepSeek-R1 enhances accessibility

and understanding for non-English-speaking vitiligo patients. To improve the readability of its English responses, training can incorporate real doctor-patient dialogue scripts to generate natural, patient-oriented language while adapting expressions for diverse cultural backgrounds.

Several limitations of this study warrant acknowledgment. First, consistent with previous exploratory studies, the relatively small sample size may have limited statistical power for detecting differences in comparative analyses, particularly when applying Bonferroni correction for multiple comparisons. Second, this study focused exclusively on text-based responses and did not incorporate images or diagnostic evaluations, potentially limiting its applicability to broader clinical contexts. Third, despite moderate-to-high inter-rater reliability, the scoring process remained inherently subjective. Future research should develop a standardized scoring system tailored for medical information to more comprehensively evaluate the feasibility of AI applications in healthcare.[36]

Conclusion

Reasoning-LLMs can provide high-quality responses to general questions about vitiligo, with their high-performance reasoning capabilities, combined with real-time data integration, offering significant advantages in delivering patient education information. Among the three models evaluated, DeepSeek-R1 achieved the highest accuracy. However, these models currently cannot offer definitive treatment recommendations and may produce errors when addressing the needs of special patient populations. These limitations highlight the need for further refinement, including improved filtering mechanisms by developers and mandatory human oversight in clinical applications. Although LLMs can support clinical decisions and enhance patient education, they should not be viewed as substitutes for dermatologists.

Data Sharing Statement

The data that support the findings of this study are available from the corresponding author, Le Zhuang, upon reasonable request.

Ethics Approval and Informed Consent

This cross-sectional study received approval from the Research Ethics Committee of Jinan Central Hospital (Registration No. 20250530023). All participants provided informed consent, in accordance with the Declaration of Helsinki.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

Research reported in this publication was supported by Shandong Province Natural Science Foundation (No. ZR2025MS1512).

Disclosure

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Ezzedine K, Eleftheriadou V, Whitton M, et al. Vitiligo. *Lancet*. 2015;386(9988):74–84.
2. Marzano AV, Alberti-Violetti S, Maronese CA, et al. Vitiligo: unmet need, management and treatment guidelines. *Dermatol Pract Concept*. 2023;13 (4S2):e2023316S.
3. Wang G, Qiu D, Yang H, et al. The prevalence and odds of depression in patients with vitiligo: a meta-analysis. *J Eur Acad Dermatol Venereol*. 2018;32(8):1343–1351. doi:10.1111/jdv.14739
4. Talsania N, Lamb B, Bewley A. Vitiligo is more than skin deep: a survey of members of the Vitiligo Society. *Clin Exp Dermatol*. 2010;35 (7):736–739. doi:10.1111/j.1365-2230.2009.03765.x

5. Benitez TM, Xu Y, Boudreau JD, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. *J Am Med Inform Assoc*. 2024;31(3):776–783. doi:10.1093/jamia/ocad252
6. Iqbal U, Lee LT, Rahmanti AR, et al. Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? *J Am Med Inform Assoc*. 2024;31(6):1341–1347. doi:10.1093/jamia/ocae067
7. Dunn C, Hunter J, Steffes W, et al. Artificial intelligence-derived dermatology case reports are indistinguishable from those written by humans: a single-blinded observer study. *J Am Acad Dermatol*. 2023;89(2):388–390. doi:10.1016/j.jaad.2023.04.005
8. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–596. doi:10.1001/jamainternmed.2023.1838
9. Robinson MA, Belzberg M, Cai C, et al. Patients prefer artificial intelligence large language model-generated responses to those prepared by the American college of mohs surgery: a double-blind comparative study using ChatGPT and Google Gemini. *JAAD Int*. 2025;21:52–54. doi:10.1016/j.jdin.2025.04.005
10. Ahmed SK, Hussein S, Aziz TA, et al. The power of ChatGPT in revolutionizing rural healthcare delivery. *Health Sci Rep*. 2023;6(11):e1684. doi:10.1002/hsr2.1684
11. Fengli X, Qianyue H, Zefang Z, et al. Towards large reasoning models: a survey of reinforced reasoning with large language models. arXiv preprint (arXiv:2501.09686v3, 2025).
12. Wei-Lin C, Lianmin Z, Ying S, et al. Chatbot arena: an open platform for evaluating LLMs by human preference. (arXiv:2403.04132, 2025).
13. Eleftheriadou V, Atkar R, Batchelor J, et al. British association of dermatologists guidelines for the management of people with vitiligo 2021. *Br J Dermatol*. 2022;186(1):18–29. doi:10.1111/bjd.20596
14. Ezzedine K, Lim HW, Suzuki T, et al. Revised classification/nomenclature of vitiligo and related issues: the vitiligo global issues consensus conference. *Pigm Cell Melanoma Res*. 2012;25(3):E1–13.
15. Read C, Wu KK, Young PM, et al. Vitiligo health education: a study of accuracy and engagement of online educational materials. *J Drugs Dermatol*. 2021;20(6):623–629.
16. Juntongjin P, Abouelsaad S, Sugkraroek S, et al. Awareness of vitiligo among multi-ethnic populations. *J Cosmet Dermatol*. 2022;21(11):5922–5930. doi:10.1111/jocd.15211
17. Speeckaert R, Van Geel N. What vitiligo patients want to know outside the dermatologist's office: an analysis of online search behaviour. *Eur J Dermatol*. 2021;31(5):667–669. doi:10.1684/ejd.2021.4141
18. Frodl A, Fuchs A, Yilmaz T, et al. ChatGPT as a source for patient information on patellofemoral surgery-A comparative study amongst laymen, doctors, and experts. *Clin Pract*. 2024;14(6):2376–2384. doi:10.3390/clinpract14060186
19. Demir S. Investigating the role of large language models on questions about refractive surgery. *Int J Med Inform*. 2025;195:105787.
20. Kincaid JP, Fishburne RP, Rogers RL, et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.
21. Lucy AT, Rakestraw SL, Stringer C, et al. Readability of patient education materials for bariatric surgery. *Surg Endosc*. 2023;37(8):6519–6525. doi:10.1007/s00464-023-10153-3
22. Zeng S, Kong Q, Wu X, et al. Artificial intelligence-generated patient education materials for helicobacter pylori infection: a comparative analysis. *Helicobacter*. 2024;29(4):e13115. doi:10.1111/hel.13115
23. van Geel N, Speeckaert R, Taieb A, et al. Worldwide expert recommendations for the diagnosis and management of vitiligo: position statement from the international vitiligo task force part 1: towards a new management algorithm. *J Eur Acad Dermatol Venereol*. 2023;37(11):2173–2184. doi:10.1111/jdv.19451
24. Alhameedy MM, Basendwh MA. Influence of narrowband ultraviolet B phototherapy on serum folate level in skin of color females: a cross-sectional study. *Int J Womens Dermatol*. 2022;8(1):e005.
25. Khene ZE, Bigot P, Mathieu R, et al. Development of a personalized chat model based on the european association of urology oncology guidelines: harnessing the power of generative artificial intelligence in clinical practice. *Eur Urol Oncol*. 2024;7(1):160–162. doi:10.1016/j.euo.2023.06.009
26. Deng L, Wang T, Yang Z, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. *Int J Surg*. 2024;110(4):1941–1950. doi:10.1097/JS9.0000000000001066
27. Kamminga NCW, Kievits JEC, Plaisier PW, et al. Do large language model chatbots perform better than established patient information resources in answering patient questions? A comparative study on melanoma. *Br. J Dermatol*. 2025;192(2):306–315.
28. Robinson MA, Belzberg M, Thakker S, et al. Assessing the accuracy, usefulness, and readability of artificial-intelligence-generated responses to common dermatologic surgery questions for patient education: a double-blinded comparative study of ChatGPT and Google Bard. *J Am Acad Dermatol*. 2024;90(5):1078–1080.
29. Ferreira LM, Nascimento JP, Souza LL, et al. Comparative analysis of language models in addressing syphilis-related queries. *Med Oral Patol Oral Cir Bucal*. 2025;30(4):e551–e560. doi:10.4317/medoral.27092
30. Xu N, Yang D, Arikawa K, et al. Application of artificial intelligence in modern medicine. *Clinical eHealth*. 2023;6:130–137.
31. Tailor PD, Dalvin LA, Starr MR, et al. A comparative study of large language models, human experts, and expert-edited large language models to neuro-ophthalmology questions. *J Neuroophthalmol*. 2025;45(1):71–77. doi:10.1097/WNO.0000000000002145
32. Ruggiero A, Martora F, Fabbrocini G, et al. The role of teledermatology during the COVID-19 pandemic: a narrative review. *Clin Cosmet Invest Dermatol*. 2022;15:2785–2793. doi:10.2147/CCID.S377029
33. Kedia N, Sanjeev S, Ong J, et al. ChatGPT and Beyond: an overview of the growing field of large language models and their use in ophthalmology. *Eye*. 2024;38(7):1252–1261. doi:10.1038/s41433-023-02915-z
34. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (Turing) test: survey study. *JMIR Med Educ*. 2023;9:e46939. doi:10.2196/46939
35. Dreyer J. China made waves with Deepseek, but its real ambition is AI-driven industrial innovation. *Nature*. 2025;638(8051):609–611. doi:10.1038/d41586-025-00460-1
36. Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digit Med*. 2024;7(1):82. doi:10.1038/s41746-024-01074-z