Original Article

Evaluation of the Accuracy, Reliability, Quality, and Readability of Artificial Intelligence Chatbots-Generated Responses to Acne-Related Questions

Ecem Bostan[1], Mahmut Talha Uçar[2], Elif Dönmez[3]

[1]Department of Dermatology and Venereology, Ankara Medipol University Faculty of Medicine, Ankara, Türkiye
[2]Department of Public Health, University of Health Sciences Hamidiye Faculty of Medicine, İstanbul, Türkiye
[3]Department of Oncology Nursing, University of Health Sciences Hamidiye Faculty of Nursing, İstanbul, Türkiye

Abstract

Aim: Since artificial intelligence (AI) has entered our lives, it has been widely used in daily medical practice to determine accurate diagnoses, predict prognosis, and inform about various treatment modalities. Acne vulgaris is one of the most frequently encountered skin problems in dermatology. Patients with acne can consult AI. The aim of the present study was to evaluate the accuracy, reliability, quality, and readability of AI-generated responses to frequently asked acne-related questions.

Materials and Methods: To evaluate the accuracy, reliability, quality, and readability of AI-generated responses to acne-related queries, a multi-domain assessment approach involving four validated tools [modified DISCERN, Global Quality Scale (GQS), Flesch Reading Ease score (FRES), and 5-point Likert scale] was used.

Results: Among the three generative AI chatbots, DeepSeek achieved the highest mean FRES, followed by ChatGPT-4.0 and ChatGPT-4.5. For modified DISCERN scores, ChatGPT-4.5 achieved the highest mean score, followed by ChatGPT-4.0 and DeepSeek, indicating superior information quality in ChatGPT-4.5 responses. The mean FRES was highest for DeepSeek among the three generative AI chatbots, whereas ChatGPT-4.5 had the highest mean modified DISCERN score. This suggests that ChatGPT-4.5 responses have higher informational quality. In terms of accuracy, ChatGPT-4.5 again achieved the highest mean score. ChatGPT-4.5 scored the highest GQS, slightly above ChatGPT-4.0, with DeepSeek scoring the lowest.

Conclusion: These results highlight that ChatGPT-4.5 generally provided more accurate, higher-quality responses, whereas DeepSeek offered superior readability according to the Flesch Reading Ease metric.

Keywords: Acne vulgaris, artificial intelligence, patient education as topic

Introduction

Generative artificial intelligence (AI) models such as ChatGPT, Gemini, and DeepSeek are now widely used in our daily lives to gather information on various subjects. Generative AI can learn from substantial amounts of data and generate new content such as text, images, music, and video.[1] Therefore, chatbots have emerged as popular and preferred tools for patients to seek medical advice and counseling before consulting a physician.

Patients with restricted access to medical care may utilize chatbots for frequently encountered dermatologic conditions such as acne, atopic dermatitis, psoriasis, and rosacea. A recent study that investigated the accuracy and sufficiency of ChatGPT, Google Bard, and Bing in answering questions about

Adress for correspondence: Ecem Bostan, Assoc. Prof., MD, Department of Dermatology and Venereology, Ankara Medipol University Faculty of Medicine, Ankara, Türkiye
Email: bostanecem@gmail.com
ORCID ID: 0000-0002-8296-4836

Submission: 05-Oct-2025
Acceptance: 06-Nov-2025 **Web Publication:** 27-Nov-2025

Access this article online

Quick Response Code:

Website: www.turkjdermatol.com

DOI: 10.4274/tjd.galenos.2025.06977

How to cite this article: Bostan E, Uçar MT, Dönmez E. Evaluation of the accuracy, reliability, quality, and readability of artificial intelligence chatbots-generated responses to acne-related questions. Turk J Dermatol. 2025;19(4):235-243.

common dermatological disorders showed that ChatGPT's responses to these questions were the most accurate and the most convenient.[2] The same study also found that ChatGPT and BingAI exhibited superior diagnostic performance, and these conversational chatbots emphasized the importance of consulting a physician for their medical conditions.[2] Image-based AI algorithms were developed to assess acne severity and identify acne morphologies; they successfully classified patients with acne.[3]

Widespread use of chatbots to gather medical information about different health conditions may give rise to significant ethical problems when false or inconvenient medical knowledge, especially about treatment modalities, is transferred to users. Therefore, the establishment of generative AI tools that can provide accurate, reliable, and readable responses to users, especially regarding medical problems, is of considerable importance.

In the present study, the accuracy, reliability, quality, and readability of AI-generated responses to the most commonly asked acne vulgaris-related questions were evaluated.

Materials and Methods

The reliability, quality, readability, and accuracy of AI-generated responses to acne-related queries (Table 1) were evaluated using a comprehensive, multi-domain assessment framework comprising four validated instruments.

Table 1. Acne-related questions retrieved from Quora and asked to ChatGPT-3.5, ChatGPT-4, and DeepSeek

	Questions
1	How can I deal with acne?
2	How can I get rid of pimples and scars?
3	How will a dermatologist help with my acne problem?
4	How well does accutane work for acne?
5	How do I deal with my adult acne?
6	Do milk and dairy products cause acne? Why?
7	Does laser treatment really get rid of acne scars permanently?
8	What is the best effective way to get rid of pimples due to hormonal imbalance?
9	How can you prevent breakouts?
10	What are the best creams to remove acne scars?
11	Why do antibiotics cause acne?
12	What is the best skincare routine for acne?
13	What is the best treatment for acne scars?
14	Does sunlight help with acne? Why?
15	What does tretinoin do for acne?
16	What is the most recommended face wash to get rid of acne?
17	Does acne eventually go away without treatment?
18	Can acne scars and redness be removed with natural remedies?
19	Why do pimples (acne) form?
20	Will my acne scars go away?

To collect representative patient questions, the keyword "acne" was searched on the Quora platform, one of the most active patient-driven discussion forums where individuals openly share their dermatological concerns in everyday language. This platform was preferred because its publicly available user-generated content reflects natural phrasing and real-world health literacy, providing an authentic basis for evaluating chatbot performance in patient communication contexts.

Analytics regarding response volume, user engagement, and upvotes were used to rank 670 questions by popularity. After the exclusion of irrelevant or inappropriate entries (n = 8), 662 questions remained. Among these, the 40 most frequently discussed were reviewed collaboratively by a board-certified dermatologist and a public health researcher. Through this multidisciplinary evaluation, twenty clinically relevant and commonly asked questions were identified for inclusion in the analysis (Figure 1).

AI-generated responses to these questions were independently obtained from ChatGPT-4.0, ChatGPT-4.5, and DeepSeek. Both the dermatologist and the public health researcher subsequently assessed each response. The dermatologist focused on the medical accuracy and clinical relevance of the information provided, while the public health researcher evaluated the reliability, quality, and readability of the texts from a health-communication perspective. The primary outcome of the study was the overall accuracy and reliability of AI-generated responses, as measured by the mDISCERN, GQS, and accuracy assessment tools. The secondary outcomes included readability [Flesch Reading Ease scores (FRES)] and inter-rater reliability [Cronbach's alpha and intraclass correlation coefficients (ICC)].

All responses were generated between 21 April and 5 May 2024, representing a time-specific snapshot of chatbot performance. Ethical approval was not required because the study utilized publicly accessible online data without involving human participants, patient records, or identifiable personal information. Accordingly, the study meets institutional criteria for exemption from human-subjects ethics review.

Reliability Assessment (Modified DISCERN)

The reliability of each response was assessed by the dermatologist using the modified DISCERN instrument[4] (Supplementary File 1), which comprises eight items that evaluate the clarity of aims, achievement of objectives, relevance, citation of sources, timing of publication, balance and impartiality, provision of supplementary resources, and

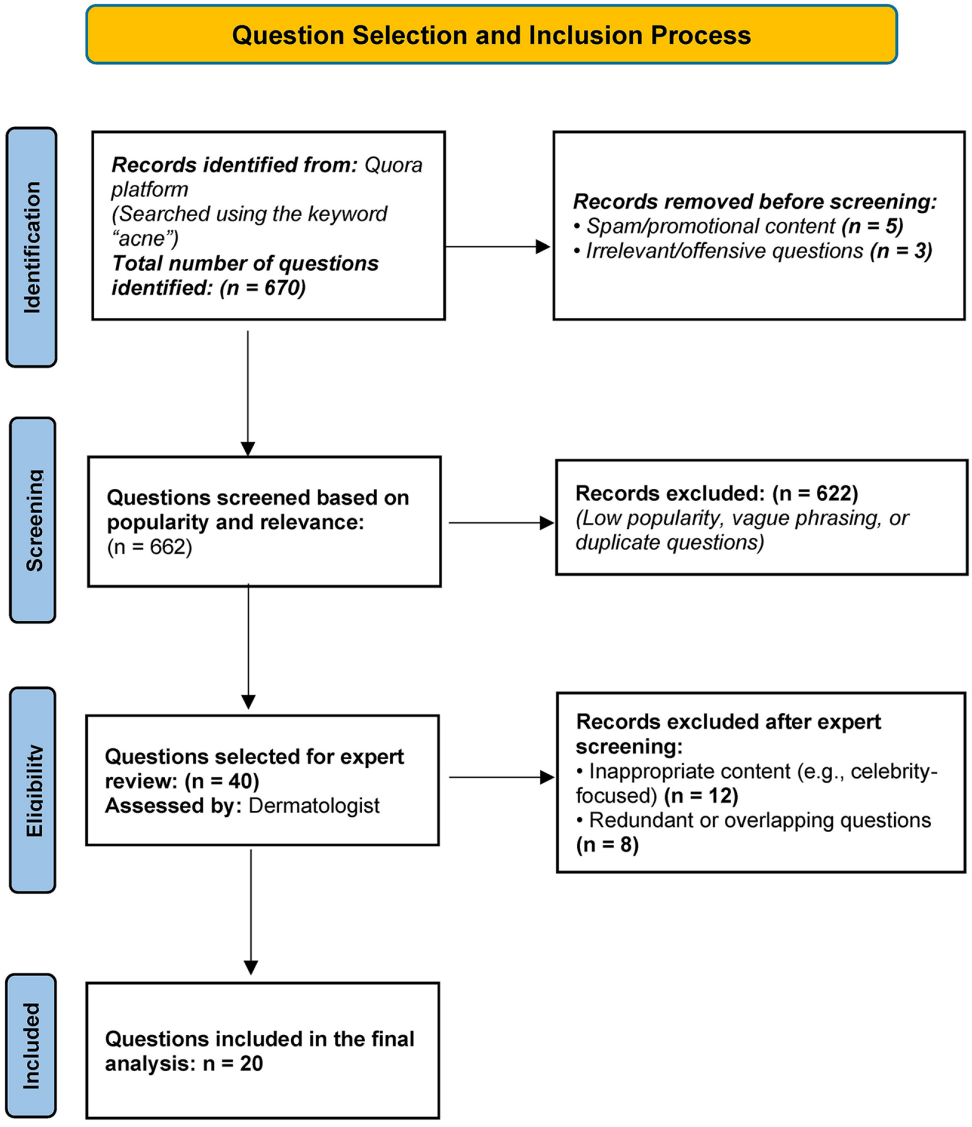

Figure 1. PRISMA-based flow diagram showing the identification, screening, eligibility assessment, and inclusion of acne-related questions collected

acknowledgement of uncertainty. Each item was rated on a 5-point Likert scale (1 = low, 5 = high), with total scores ranging from 8 to 40.[4] Higher scores indicate greater reliability and information integrity.

Quality Assessment (Global Quality Scale)

Overall quality was also rated by the dermatologist using the Global Quality Scale (GQS)[5] (Supplementary File 2), a validated 5-point instrument designed to assess the coherence, comprehensiveness, and patient-centered utility of online health information. A score of 1 reflected poor quality and minimal usefulness, whereas a score of 5 indicated excellent content flow and substantial patient benefit.[5]

Readability Assessment (Flesch Reading Ease Score)

Readability was assessed by a public health researcher using the FRES, which evaluates the ease of comprehension based on average sentence length and the average number of syllables per word. Scores range from 0 to 100, with higher scores indicating easier readability. The FRES for each response was calculated using a standardized online tool (https://readabilityformulas.com),[6] and interpreted according to established classification thresholds: very easy (90-100), easy (80-89), fairly easy (70-79), standard (60-69), fairly difficult (50-59), difficult (30-49), and very difficult (0-29).[6]

Accuracy Assessment

The accuracy of each AI-generated response was evaluated using a five-point Likert scale adapted from previous studies

assessing the quality of medical information generated by large language models.[7-9] This method has been widely adopted in the recent literature to assess the factual accuracy and clinical consistency of AI-generated health content.[7-9] Scores ranged from 1 to 5, where:

1 – indicated completely incorrect or misleading information;

2 – represented mostly incorrect content with minor correct elements;

3 – reflected a balance of correct and incorrect information;

4 – denoted mostly correct information with minor inaccuracies or omissions;

5 – indicated completely accurate information consistent with current dermatological guidelines and evidence-based practice.

Each response was independently rated by two evaluators with clinical expertise in dermatology and public health. Discrepancies in scoring were resolved through discussion and consensus.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics (Version 29.0). Descriptive statistics were presented as mean ± standard deviation. The Kruskal-Wallis test was used to compare the three AI models (ChatGPT-4.0, ChatGPT-4.5, and DeepSeek) across four evaluation domains: reliability (mDISCERN), quality (GQS), readability (FRES), and accuracy. Where significant differences were found, pairwise comparisons were conducted using the Mann-Whitney U test with Bonferroni correction. Effect sizes were calculated using eta-squared (η^2) for Kruskal-Wallis analyses and rank-biserial correlation (r) for pairwise Mann-Whitney U tests to quantify the magnitude of differences. Statistical significance was set at $P < 0.05$.

Results

Among the chatbot models, DeepSeek had the highest mean FRES (44.50±14.16), followed by ChatGPT-4.0 (42.40±11.39) and ChatGPT-4.5 (23.70±9.27). This suggests that DeepSeek and ChatGPT-4.0 produced responses that were more readable than those of ChatGPT-4.5. Regarding modified DISCERN scores, ChatGPT-4.5 had the highest mean (30.75±3.40), followed by ChatGPT-4.0 (28.55±2.93) and DeepSeek (25.40±3.36). This suggests that ChatGPT-4.5 responses were of a higher quality. In terms of GQS, ChatGPT-4.5 scored highest (4.15±0.81), slightly above ChatGPT-4.0 (4.10±0.71); DeepSeek had the lowest score (3.70±0.73). When accuracy was evaluated, ChatGPT-4.5 showed the highest mean accuracy (4.25±0.55), followed by ChatGPT-4.0 (4.05±0.51), and DeepSeek (3.95±0.51). This indicates relatively consistent accuracy across the models. The summary of mDISCERN, GQS, and readability scores is shown in Table 2, whereas the radar plots of responses generated by ChatGPT4.0, ChatGPT4.5, and DeepSeek for reliability, quality, readability, and accuracy across 20 acne-related questions are depicted in Figure 2.

A Kruskal-Wallis test was conducted to compare the performance of ChatGPT-4.0, ChatGPT-4.5, and DeepSeek across four evaluation metrics: GQS, mDISCERN, FRES, and accuracy assessment.

Global Quality Score

No statistically significant difference was found among the three models in terms of GQS [χ^2 (2) = 4.746, P = 0.093]. However, ChatGPT-4.5 had the highest mean rank (34.58), followed by ChatGPT-4.0 (32.80) and DeepSeek (24.13).

Modified DISCERN Score

A significant difference was observed among the models [χ^2 (2) = 19.961, $P < 0.001$]. ChatGPT-4.5 showed the highest mean rank (42.33), followed by ChatGPT-4.0 (31.38); DeepSeek had the lowest (17.80), indicating that its information quality scores were significantly lower.

Flesch Reading Ease Score

The difference among the models was statistically significant [χ^2 (2) = 24.703, $P < 0.001$]. DeepSeek achieved the highest readability rank (39.63), closely followed by ChatGPT-4.0 (37.15), while ChatGPT-4.5 ranked lowest (14.73), suggesting that ChatGPT-4.5 responses were more difficult to read.

Table 2. The summary of mDISCERN, GQS, and Readability scores

Model*	mDISCERN (mean ± SD)	GQS (mean ± SD)	Flesch Reading Ease score (mean ± SD)
ChatGPT-4.0	28.55±2.93	4.25±0.44	41.8±11.67
ChatGPT-4.5	30.7±3.28	4.25±0.44	39.4±15.82
DeepSeek	25.05±4.77	3.75±0.85	45.55±10.87

*Effect sizes (η^2): mDISCERN = 0.32, FRES = 0.38, GQS = 0.03, Accuracy = 0.01
GQS: Global Quality Scale, FRES: Flesch Reading Ease score, SD: Standard deviation

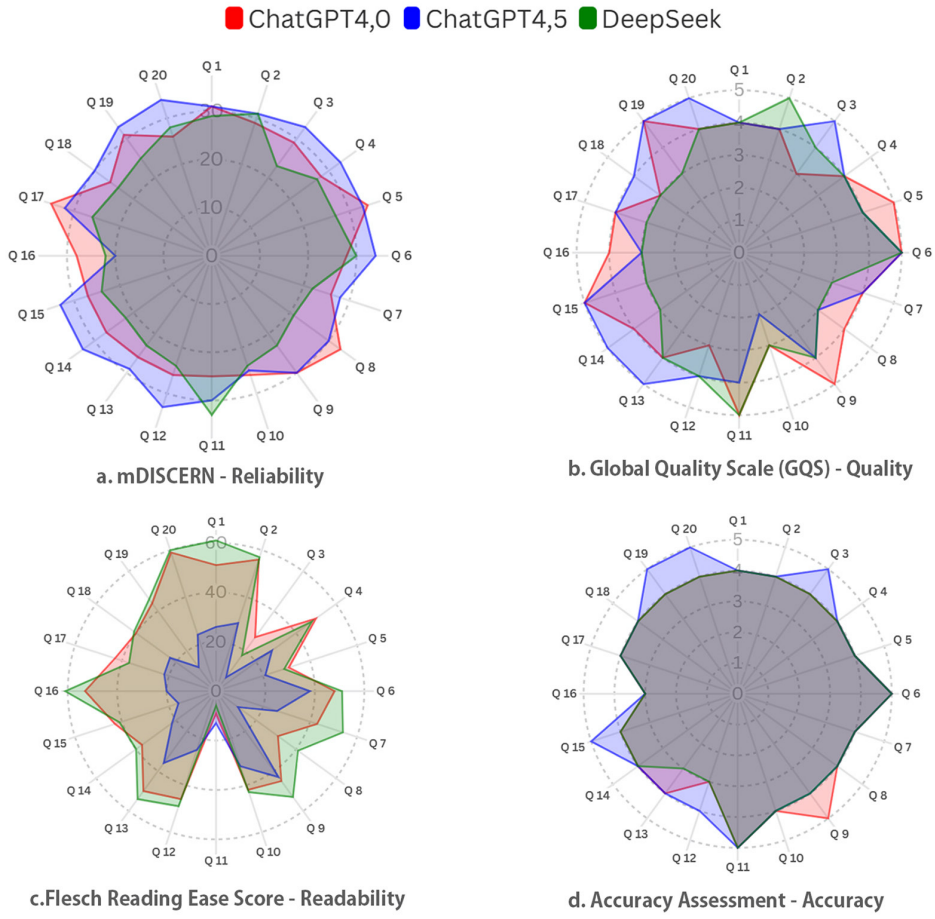

Figure 2. Radar plots of chatbot responses for reliability, quality, readability and accuracy across 20 acne-related questions

Accuracy Assessment

The models did not differ significantly in terms of accuracy [χ^2 (2) = 3.361, P = 0.186]. ChatGPT-4.5 ranked highest (34.88), followed by ChatGPT-4.0 (29.60) and DeepSeek (27.03).

Post-hoc Comparisons for mDISCERN

Pairwise comparisons using the Mann-Whitney U test with Bonferroni correction (adjusted α = 0.0167) revealed significant differences among the three chatbot models.

ChatGPT-4.5 vs. ChatGPT-4.0: ChatGPT-4.5 yielded significantly higher mDISCERN scores than ChatGPT-4.0 (U = 109.5; Z = -2.47; P = 0.014).

Although this value was statistically significant at the conventional 0.05 level, it narrowly missed significance following Bonferroni correction.

ChatGPT-4.0 vs. DeepSeek: ChatGPT-4.0 demonstrated significantly higher scores than DeepSeek (U = 92.0, Z = -2.94, P = 0.003); this difference remained significant after correction.

In the comparison between ChatGPT-4.5 and DeepSeek, A marked difference was observed in favour of ChatGPT-4.5, which substantially outperformed DeepSeek (U = 54.0; Z = –3.97; P < 0.001), even after adjustment for multiple comparisons.

These findings support the superior information quality of ChatGPT-4.5, particularly in comparison to DeepSeek.

Post-hoc Comparisons for Readability

Pairwise comparisons based on the FRES revealed significant differences between models, as determined by Mann-Whitney U tests with Bonferroni adjustment (α = 0.0167):

ChatGPT-4.0 vs. ChatGPT-4.5:

ChatGPT-4.0 generated significantly more readable responses than ChatGPT-4.5 (U = 38.5, Z = -4.37, P < 0.001).

ChatGPT-4.0 vs. DeepSeek:

No statistically significant difference in readability was observed between ChatGPT-4.0 and DeepSeek (U = 171.5, Z = -0.77, P = 0.440).

ChatGPT-4.5 vs. DeepSeek:

DeepSeek responses were significantly more readable than those of ChatGPT-4.5 (U = 46.0; Z = -4.17; $P < 0.001$).

Collectively, these findings highlight the relatively poor readability of ChatGPT-4.5 responses compared to both ChatGPT-4.0 and DeepSeek, with DeepSeek showing the highest readability overall.

For GQS, the internal consistency was poor, with a Cronbach's alpha of 0.388. The single-measure ICC was 0.106 [95% confidence interval (CI): -0.052 to 0.353], indicating low reliability between evaluators. The average-measure ICC was 0.262 (95% CI: -0.175 to 0.621), and the results were not statistically significant ($P = 0.097$). Regarding mDISCERN, the internal consistency was moderate (Cronbach's alpha = 0.542). The single-measure ICC was 0.264 (95% CI: 0.013-0.554) and the average-measure ICC was 0.518 (95% CI: 0.038-0.788), suggesting a fair level of agreement. These results were statistically significant ($P = 0.020$). The FRESs showed high internal consistency (Cronbach's alpha = 0.865). The single-measure ICC was 0.353 (95% CI: 0.004-0.674) and the average-measure ICC was 0.620 (95% CI: 0.012-0.861); both estimates were statistically significant ($P < 0.001$), indicating good inter-rater reliability. For accuracy assessment, the inter-rater reliability was high. Cronbach's alpha was 0.843, indicating excellent internal consistency across evaluators. The single-measure ICC was 0.602 (95% CI, 0.348-0.800), and the average-measure ICC was 0.819 (95% CI, 0.616-0.923), both were statistically significant ($P < 0.001$). These results confirm a strong absolute agreement between the models in terms of response accuracy. The summary of reliability metrics is shown in Table 3.

Discussion

Findings of this study indicate that ChatGPT-4.5 generally provided more accurate, higher-quality responses to acne-related queries, while DeepSeek provided superior readability as measured by the Flesch Reading Ease metric. To our knowledge, the present study is the first investigation to evaluate and compare three generative AI tools. We believe that the preliminary findings of our study will stimulate further investigations into the accuracy, reliability, quality, and readability of conversational AI-generated responses to questions about general skin problems.

As generative AI tools are now being increasingly used in our daily lives for purposes such as gathering information about various subjects, creating images, video, or text, or simply chatting, patients might find it easier to consult generative AI tools about their health problems. When prompt face-to-face dermatologic care is difficult to access, conversational AI programs that can compile information from large, complex datasets may be a satisfactory alternative.[10] Several studies have been conducted recently to evaluate the accuracy, credibility, and comprehensiveness of the information generated by conversational AI programs.[10-13] In a recent study by Gawey et al.,[12] the readability of ChatGPT-retrieved responses to the most frequently-asked questions about hidradenitis suppurativa (HS) were compared with the readability of the information provided by HS Foundation, HS Patient Guide and HS related websites. In this study, ChatGPT's responses were found to have a higher mean readability grade compared with other HS-related sources, even though FRES was significantly lower for ChatGPT than for other HS-related sources.[12] These findings underline the fact that the higher reading level of ChatGPT may impair the users' perception. Although comprehensibility is essential for readers to understand the information presented by AI tools, it is not the only criterion for appraising data generated by generative AI programs. Another study by Kamminga et al.[11] which compared the responses of large language models (ChatGPT-3.5, ChatGPT-4 and Gemini) and Dutch patient information resources (PIRs) to melanoma-related questions in terms of medical accuracy, readability, completeness and personalization and reproducibility, showed that ChatGPT-related answers had the highest accuracy whereas Gemini-generated responses were the best in readability, completeness and personalization. The same study also revealed that the best-performing large language models surpassed gold-standard PIRs on personalization and completeness, but not on accuracy and readability.[11] These results suggest that even though large language models demonstrated promising results,

Table 3. The summary of reliability metrics

Evaluation criteria	Cronbach's alpha	Single ICC*	CI** (single ICC)	Avg ICC	CI (Avg ICC)	*P*
GQS	0.388	0.106	-0.052-0.353	0.262	-0.17-0.621	0.097
mDISCERN	0.542	0.264	0.013-0.554	0.518	0.038-0.788	0.02
Flesch Reading Ease	0.865	0.353	0.004-0.674	0.62	0.012-0.861	< 0.001
Accuracy assessment	0.843	0.602	0.348-0.800	0.819	0.616-0.923	< 0.001

*ICC indicates intraclass correlation coefficient
**CI indicates confidence interval
GQS: Global Quality Scale

fortification and surveillance of accuracy and reproducibility are still needed.[11] In our study, among the three generative AI models, ChatGPT-4.5-derived responses had the highest quality and correctness, according to investigators' assessment, whereas DeepSeek was the easiest to read. This outcome underscores that different large language models have varying strengths and weaknesses. There appears to be a need for the standardization, personalization, and consolidation of large language models.

Recently, an investigation by Boostani et al.[13] evaluated the performance of GPT-4o and Gemini Flash 2.0 in diagnosing acne and rosacea-related clinical photographs. The outcomes of this study showed that GPT-4o demonstrated higher accuracy than Gemini in diagnosing rosacea and acne, but subtyping performance was markedly lower.[13] The considerably diagnostic accuracy (93%) of GPT-4o for acne and rosacea emphasizes the potential and competence of large language models in diagnosing skin diseases.[13] Furthermore, the performance of different ChatGPT versions in the dermatology specialty examination was assessed.[14-16] In one of these studies,[16] GPT-4 was found to obtain an overall accuracy of 75% on 250 randomly chosen dermatology board-style questions whereas in another investigation,[15] ChatGPT-4 performed better with an overall accuracy of 90% when compared to the performance (63%) of ChatGPT-3.5. Even though we did not investigate the performance of generative AI tools on the dermatology specialty examination, we also found that ChatGPT-4.5 ranked highest (34.88), followed by ChatGPT-4.0 (29.60), and DeepSeek (27.03) when the accuracy of the answers to acne vulgaris-related questions was assessed. Collectively, these results suggest that AI might become an essential adjunct for improving dermatology education and facilitating patient care and communication in the coming years.[17] Patients might find it easier to consult AI about the causes, prognoses and treatment options for different health problems, since gaining access to timely in-person medical care is not always feasible. However, ethical conflicts that may arise from the use of AI chatbots as the primary source of consultation for various health problems remain to be elucidated.

In addition to satisfying users' various skin-related problems and assisting with the dermatology speciality examination, AI has also gained prominence in cosmetic dermatology.[18-20] In a clinical study by Cazzaniga et al.,[21] artificial neural network models were used to estimate the clinical response to excimer laser therapy in vitiligo patients. Furthermore, the use of robot-assisted hair removal laser systems has been proven to be efficacious and safe.[22,23] An inception-based convolutional neural network has also been used to detect facial wrinkles and aid in deciding whether the forehead region needs filler injections; this model demonstrated an accuracy of 85.3%.[24] These studies once again highlight that integrating AI into aesthetic dermatology will most likely provide a more standardized and personalized approach to treatment for cosmetic interventions. AI seems to be a promising, complementary tool that enables the unification of the physician's ingenuity with the use of large amounts of evidence-based data.

This study has several notable strengths. First, it represents one of the earliest comparative analyses of generative AI chatbots—specifically ChatGPT-4.0, ChatGPT-4.5, and DeepSeek—in the context of acne vulgaris, a frequently encountered dermatological condition. The study's novelty and focused scope provide valuable insights into the evolving role of AI in patient education and dermatologic self-care. Second, the methodological rigor of the study is underscored by the use of four validated tools—modified DISCERN, GQS, FRES, and a 5-point Likert scale—offering a multidimensional evaluation of AI-generated content in terms of reliability, quality, readability, and accuracy. The inclusion of both a board-certified academic dermatologist and a public health researcher as independent evaluators further strengthens the validity and clinical relevance of the findings. Additionally, appropriate statistical analyses, including Kruskal-Wallis tests and Bonferroni-corrected post-hoc comparisons, were conducted to ensure the robustness of inter-model comparisons. These features collectively enhance the reliability and applicability of the study's results.

Study Limitations

Several limitations of the present study must also be acknowledged. The scope of the study was limited to acne vulgaris; therefore, the findings may not be generalizable to other dermatologic or systemic medical conditions. Furthermore, chatbot responses were retrieved and evaluated at a single point in time, representing a snapshot of model performance. Because generative AI tools are frequently updated, their future outputs may differ from those analyzed in this study. Although validated instruments were employed and inter-rater reliability statistics were used to mitigate this bias, some degree of subjectivity in evaluators' scoring cannot be entirely excluded. Additionally, the study focused exclusively on English-language content and relied on questions sourced from a single online platform (Quora), which may introduce language- and platform-related biases and limit the cross-cultural applicability of the findings. Despite these limitations, the study provides an important foundation for future research and contributes to the growing discourse on the integration of AI in dermatologic education and patient care.

Conclusion

With the ongoing involvement of AI in our daily lives, there is growing interest in incorporating AI into medicine. AI is now widely used in dermatology, facilitating the diagnosis of various skin diseases and providing detailed information on prognosis and treatment options. Our study also showed that generative AI programmes appear to be effective in answering acne-related questions and building bridges between patients and physicians, although there seems to be a need to strengthen several parameters (reliability, accuracy, and readability) across generative AI tools.

Ethics

Ethics Committee Approval: Not applicable.

Informed Consent: Not applicable.

Footnotes

Authorship Contributions

Concept: M.T.U., E.D., Design: M.T.U., E.D., Data Collection or Processing: E.B., M.T.U., E.D., Analysis or Interpretation: E.B., M.T.U., Literature Search: E.B., Writing: E.B., M.T.U., E.D.

Conflict of Interest: The authors declared that they have no conflict of interest.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Currie GM, Hawk KE, Rohren EM. Generative artificial intelligence biases, limitations and risks in nuclear medicine: an argument for appropriate use framework and recommendations. Semin Nucl Med. 2025;55(3):423-436.
2. Chau CA, Feng H, Cobos G, Park J. The comparative sufficiency of ChatGPT, Google Bard, and Bing AI in answering diagnosis, treatment, and prognosis questions about common dermatological diagnoses. JMIR Dermatol. 2025;8:e60827.
3. Seité S, Khammari A, Benzaquen M, Moyal D, Dréno B. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. Exp Dermatol. 2019;28:1252-1257.
4. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105-111.
5. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102:2070-2077.
6. Readability Formulas [Internet]. 2025 Jul 5 [cited 2025 Nov 10]. Available from: https://readabilityformulas.com
7. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024;24(1):211.
8. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, Lagana G, Guenza G, Agosta E, Vinjolli F, Hoxha M, D'Amelio C, Favaretto N, Chisci G. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med. 2024;13(3):735.
9. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, Chang S, Berkowitz S, Finn A, Jahangir E, Scoville E, Reese T, Friedman D, Bastarache J, van der Heijden Y, Wright J, Carter N, Alexander M, Choe J, Chastain C, Zic J, Horst S, Turker I, Agarwal R, Osmundson E, Idrees K, Kieman C, Padmanabhan C, Bailey C, Schlegel C, Chambless L, Gibson M, Osterman T, Wheless L. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. JAMA Network Open. 2023;6(10):e2336483.
10. Lakdawala N, Channa L, Gronbeck C, Lakdawala N, Weston G, Sloan B, Feng H. Assessing the accuracy and comprehensiveness of ChatGPT in offering clinical guidance for atopic dermatitis and acne vulgaris. JMIR Dermatol. 2023;6:e50409.
11. Kamminga NCW, Kievits JEC, Plaisier PW, Burgers JS, van der Veldt AM, van den Brand JAGJ, Mulder M, Wakkee M, Lugtenberg M, Nijsten T. 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.
12. Gawey L, Dagenet CB, Tran KA, Park S, Hsiao JL, Shi V. Readability of information generated by ChatGPT for hidradenitis suppurativa. JMIR Dermatol. 2024;7:e55204.
13. Boostani M, Bánvölgyi A, Goldust M, Cantisani C, Pietkiewicz P, Lőrincz K, Holló P, Wikonkál NM, Paragh G, Kiss N. Diagnostic performance of GPT-4o and Gemini flash 2.0 in acne and rosacea. Int J Dermatol. 2025;64(10):1881-1882.
14. Samman L, Akuffo-Addo E, Rao B. The Performance of artificial intelligence Chatbot (GPT-4) on image-based dermatology certification board exam Questions. J Cutan Med Surg. 2024;28(5):507-508.
15. Passby L, Jenko N, Wernham A. Performance of ChatGPT on specialty certificate examination in dermatology multiple-choice questions. Clin Exp Dermatol. 2024;49(7):722-727.
16. Elias ML, Burshtein J, Sharon VR. OpenAI's GPT-4 performs to a high degree on board-style dermatology questions. Int J Dermatol. 2024;63(1):73-78.
17. Diamond C, Rundle CW, Albrecht JM, Nicholas MW. Chatbot utilization in dermatology: a potential amelioration to burnout in dermatology. Dermatol Online J. 2022;28(6).
18. Elder A, Ring C, Heitmiller K, Gabriel Z, Saedi N. The role of artificial intelligence in cosmetic dermatology-Current, upcoming, and future trends. J Cosmet Dermatol. 2021;20(1):48-52.
19. Kania B, Montecinos K, Goldberg DJ. Artificial intelligence in cosmetic dermatology. J Cosmet Dermatol. 2024;23(10):3305-3311.
20. Gold MH, Goldust M. Synergy of artificial intelligence and laser tech in cosmetic dermatology. J Cosmet Dermatol. 2025;24(3):e16799.
21. Cazzaniga S, Sassi F, Mercuri SR, Naldi L. Prediction of clinical response to excimer laser treatment in vitiligo by using neural network models. Dermatology. 2009;219(2):133-137.
22. Lim HW, Lee DH, Cho M, Park S, Koh W, Kim Y, Chung JH, Kim S. Comparison of efficacy between novel robot-assisted laser hair removal and physician-directed hair removal. Photomed Laser Surg. 2015;33(10):509-516.
23. Lim HW, Park S, Noh S, Lee DH, Yoon C, Koh W, Kim Y, Chung JH, Kim HC, Kim S. A study on the development of a robot-assisted automatic laser hair removal system. Photomed Laser Surg. 2014;32(11):633-641.
24. Alrabiah A, Alduailij M, Crane M. Computer-based approach to detect wrinkles and suggest facial fillers. Int J Adv Comput Sci Appl. 2019;10(9):319-325.

Supplementary File 1. mDISCERN criteria scoring
mDISCERN criteria total score (8-40 points)
1. Are the aims clear? 1-5 point
2. Does it achieve its aims? 1-5 point
3. Is it relevant? 1-5 point
4. Is it clear what sources of information were used to compile the publication (other than the author or producer)? 1-5 point
5. Is it clear when the information used or reported in the publication was produced? 1-5 point
6. Is it balanced and unbiased? 1-5 point
7. Does it provide details of additional sources of support and information? 1-5 point
8. Does it refer to areas of uncertainty? 1-5 point

Supplementary File 2. Global Quality index scoring	
Global Quality index scoring	**Score**
Poor quality, poor flow of the site, most information missing, not at all useful for patients	**1**
Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients	**2**
Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients	**3**
Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients	**4**
Excellent quality and excellent flow, very useful for patients	**5**