Original Paper

Evaluating GPT-4 Responses on Scars or Keloids for Patient Education: Large Language Model Evaluation Study

Mingjun Rao, MD; Tang Xiujun, MSc; Wang Haoyu, BA

Department of Plastic Surgery, Guizhou Provincial People's Hospital, Guiyang, China

Corresponding Author:

Tang Xiujun, MSc
Department of Plastic Surgery
Guizhou Provincial People's Hospital
83 Zhongshan East Road, Nanming District
Guiyang 550002
China
Phone: 86 15343315902
Email: xiujunsszx@163.com

Abstract

Background: Scars and keloids impose significant physical and psychological burdens on patients, often leading to functional limitations, cosmetic concerns, and mental health issues such as anxiety or depression. Patients increasingly turn to online platforms for information; however, existing web-based resources on scars and keloids are frequently unreliable, fragmented, or difficult to understand. Large language models such as GPT-4 show promise for delivering medical information, but their accuracy, readability, and potential to generate hallucinated content require validation for patient education applications.

Objective: This study aimed to systematically evaluate GPT-4's performance in providing patient education on scars and keloids, focusing on its accuracy, reliability, readability, and reference quality.

Methods: This study involved collecting 354 questions from Reddit communities (r/Keloids, r/SCAR, and r/PlasticSurgery), covering topics including treatment options, pre- and postoperative care, and psychological impacts. Each question was input into GPT-4 in independent sessions to mimic real-world patient interactions. Responses were evaluated using multiple tools: the Patient Education Materials Assessment Tool-Artificial Intelligence for understandability and actionability, DISCERN-AI for treatment information quality, the Global Quality Scale for overall information quality, and standard readability metrics (Flesch Reading Ease score, and Gunning Fog Index). Three plastic surgeons used the Natural Language Assessment Tool for Artificial Intelligence to rate the accuracy, safety, and clinical appropriateness, while the Reference Evaluation for Artificial Intelligence tool validated references for reference hallucination, relevance, and source quality. We conducted the same analysis to assess the quality of GPT-4–generated content in response to questions from 3 medical websites.

Results: GPT-4 demonstrated high accuracy and reliability. The Patient Education Materials Assessment Tool-Artificial Intelligence showed 75.5% understandability, DISCERN-AI rated responses as "good" (26.3/35), and the Global Quality Scale score was 4.28 of 5. Surgeons' evaluations averaged 3.94 to 4.43 out of 5 across dimensions (accuracy 3.9, SD 0.7; safety 4.3, SD 0.8; clinical appropriateness 4.4, SD 0.5; actionability 4.1, SD 0.8; and effectiveness 4.1, SD 0.8). Readability analyses indicated moderate complexity (Flesch Reading Ease Score: 50.13; Gunning Fog Index: 12.68), corresponding to a 12th-grade reading level. Reference Evaluation for Artificial Intelligence identified 11.8% (383/3250) hallucinated references, while 88.2% (2867/3250) of references were real, with 95.1% (2724/2867) from authoritative sources (eg, government guidelines and the literature). The overall results about questions from medical websites were consistent with the answers to Reddit questions.

Conclusions: GPT-4 has serious potential as a patient education tool for scars and keloids, offering reliable and accurate information. However, improvements in readability (to align with sixth to eighth grade standards) and reduction of reference hallucinations are essential to enhance accessibility and trustworthiness. Future large language model optimizations should prioritize simplifying medical language and strengthening reference validation mechanisms to maximize clinical utility.

JMIR Med Inform 2026;14:e78838; doi: 10.2196/78838

Keywords: scar; keloid; GPT-4; patient education; generative AI; generative artificial intelligence; readability; large language model

Introduction

Scars and keloids are common skin healing outcomes [1], often causing discomfort during the proliferative phase [2]. When located on visible areas such as the face, they can severely impact a patient's appearance, leading to psychological distress such as low self-esteem, anxiety, and depression, which may further hinder social interactions and career development [3]. Scars near joints may cause functional limitations, while perineal scars can result in long-term complications such as dyspareunia and dysmenorrhea [4,5]. Many patients urgently seek to improve both the aesthetic and functional aspects of scars or keloids. However, treatments often require multimodal approaches over weeks to months, making high patient adherence crucial. Consequently, access to accurate, timely, and comprehensive management information is critical for patients to better understand treatment expectations, options, prognosis, and potential complications [6-9].

Currently, patients increasingly rely on internet-based health information [10]. This trend faces multifaceted challenges, including information overload, variability in source credibility and content accuracy, and the health literacy required to understand the contents [11,12]. Notably, many online resources on scars or keloids are often unreliable, fragmented, or difficult to understand, failing to tackle the fundamental needs of patients with scars or keloids [13].

In recent years, artificial intelligence (AI) tools powered by large language models (LLMs), such as GPT-4 (used by ChatGPT), have demonstrated significant potential in delivering medical information [14]. GPT-4's capacity to generate natural language responses through interactive conversions could aid users in understanding intricate medical concepts, treatment, and management strategies, positioning it as a potentially valuable alternative to traditional search engines for accessing knowledge associated with scars or keloids [15-17].

It is reported that 52% of US adults have used LLMs, and GPT-4, as a leading LLM, receives over 5 billion monthly visits. In total, 39% of LLM users have used LLMs for health care queries [18]. Despite the increasing use of LLMs in health care [19-21], there remains a research gap, and it is currently unclear whether GPT-4 can generate high-quality patient education content related to scars and keloids. Thus, we conducted this study to comprehensively assess the use of GPT-4 in keloid and scar patient education by performing a multidimensional evaluation (encompassing accuracy, reliability, readability, and reference quality) of GPT-4 responses.

Methods

Study Objective

This study aimed to investigate the potential of GPT-4 to provide reliable, accurate, readable, and actual medical information for patients with scars or keloids. To achieve this, we used GPT-4 (OpenAI) to evaluate its accuracy, reliability, readability, and hallucinations in answering questions related to treatments of scars or keloids.

Question Collection

The research questions were manually collected by the authors from Reddit. First, the authors analyzed all posts on the "Hot" page (the most popular and recently active posts) of the r/Keloids subreddit, all posts on the r/SCAR subreddit, and all posts on the r/PlasticSurgery subreddit as of April 6, 2025. We extracted the main text of each post that included the keywords "Scar" or "Keloid" and organized the data using Microsoft Excel. Second, one author (MR) excluded the posts that contained no questions, were duplicates, or had repeated questions. In addition, the same author (MR) performed the initial classification of the questions. To ensure reliability and minimize bias, another author (WH) independently reviewed the process. Consistency between the two authors' classifications was confirmed through discussion. This data collection approach has been adopted in previous Reddit-based research [22]. Furthermore, we adopted 49 questions about keloids or scars from 3 medical websites.

Ethical Considerations

The institutional review board of the People's Hospital of Guizhou Province, affiliated with Guizhou University, deemed this study exempt from ethics approval.

Quality Assessment

Overview

Each question was input individually into GPT-4. Following previous research protocols, a new chat session was initiated for each question to avoid context contamination and to simulate real-world patient interactions [23]. Consistent with real-world activities using GPT-4, no specialized prompt instructions were appended to the question inputs. The contents generated by ChatGPT-4 were evaluated using a modified version of existing health information quality assessment tools.

Patient Education Materials Assessment Tool-AI Tool

The Patient Education Materials Assessment Tool (PEMAT) was used to assess the understandability and actionability of ChatGPT-generated content. The original PEMAT includes 17 items for understandability and 7 for actionability. Since all contents generated by ChatGPT are pure text, the PEMAT was simplified to 8 items for understandability and 3 items for

actionability (Multimedia Appendix 1). Each item received 1 point if it met the standard, and scores were reported as percentages. A score of 70% or greater was recorded as a "pass" per PEMAT guidelines [24].

DISCERN-AI Tool

The DISCERN standard, a previously validated tool to help health care consumers and professionals evaluate the quality of treatment information, was adapted for ChatGPT-generated content. Since all contents generated by ChatGPT are pure text, 7 items (questions 3-9 from the 15-item DISCERN tool) were selected and scored on a 1 to 5 scale (Multimedia Appendix 2). Each output was rated as follows: very poor (7-12 points), poor (13-17 points), fair (18-23 points), good (24-28 points), and excellent (29-35 points) [24,25].

Global Quality Scale

The Global Quality Scale (GQS) is a 5-point Likert scale used to evaluate information quality and the flow and ease of use of information. The scores range from 1 (low quality) to 5 (high quality), while scores of 4 or 5 indicated high-quality outputs, a score of 3 was considered moderate quality and scores of 1 or 2 were categorized as low quality.

Readability Assessment

The readability of the ChatGPT-generated content was evaluated using several established readability formulas, including Flesch Reading Ease score, Gunning Fog Index, Flesch-Kincaid Grade Level, Coleman-Liau Index, and Simple Measure of Gobbledygook (SMOG). Each output was copied into Microsoft Word and analyzed via the Readable website [26]. The Flesch Reading Ease score ranges from 0 to 100, and higher scores indicate greater readability. A score between 60 and 70 corresponds to reading levels of grades 8 and 9 and is generally understandable by the average adult. The Gunning Fog Index and Flesch-Kincaid Grade Level are used to estimate sentence complexity; the scores represent the years of formal education required to understand the contents. For example, a score of 12 implies the output is suitable for readers at the 12th-grade level. The Coleman-Liau Index is similar to the Gunning Fog Index and Flesch-Kincaid Grade Level but uses character counts instead of syllables, making it more suitable for languages where syllable counts may not accurately reflect complexity. The SMOG Index measures syllable density, often used to assess health information materials. A score of 12 indicates that the material is suitable for readers at the 12th-grade level or higher.

Natural Language Assessment Tool for Artificial Intelligence

Three experienced plastic surgeons independently reviewed each GPT-4–generated content using a specially developed Natural Language Assessment Tool for Artificial Intelligence (NLAT-AI) [24]. Using this tool, we assessed accuracy, safety, appropriateness, actionability, and effectiveness. Each output was rated using a 5-point Likert scale (1=strongly disagree, 5=strongly agree; Multimedia Appendix 3). All results were summarized using descriptive statistics.

Reference Evaluation for AI

Given known issues of LLM hallucination (ie, generating plausible but nonexistent references), a brief evaluation tool, Reference Evaluation for AI, was developed to analyze references provided in ChatGPT-generated content [27]. Each reference was verified through direct links or a Google search. The tool assessed (1) reference hallucination (whether references were real or fabricated), (2) relevance and consistency between references and AI output, and (3) source quality (based on the authority of the issuing institution or organization, such as government guidelines, health care organizations, or scientific research; Multimedia Appendix 4).

Results

Question Collection and Classification

A total of 507 posts were identified and analyzed (posts from the r/Keloids subreddit: n=193, 38.1%; posts from the r/Keloids subreddit: n=211, 41.6%; and posts from the r/Scar subreddit: n=103, 20.3%). After removing posts that merely shared information or were duplicates, 354 unique questions were obtained. The questions were categorized into 16 groups based on their contents (Table 1). Furthermore, we obtained 49 questions from 3 medical websites that included 38 unique questions (Table S1 in Multimedia Appendix 5).

Table 1. Questions on scars or keloids from Reddit (N=354).

Question group	Questions, n (%)
Questions on other respects	28 (7.9)
Questions on other treatments for scars or keloids	4 (1.1)
Questions on common treatments for scars or keloids	46 (13)
Questions on trauma-related scars or keloids	16 (4.5)
Questions on psychological issues caused by scars or keloids	9 (2.5)
Questions on at-home scar or keloid care	3 (0.8)
Questions on preoperative scar or keloid consultation	37 (10.5)
Questions on postoperative scar or keloid consultation	55 (15.5)

Question group	Questions, n (%)
Questions on selection of treatments for scars or keloids	80 (22.6)
Questions on impact of scars or keloids on daily life	2 (0.6)
Questions on scar or keloid symptoms	7 (2)
Questions on scar camouflage	6 (1.7)
Questions on the impact of nutrition on scars or keloids	3 (0.8)
Questions on choosing physicians for scar or keloid treatment or related costs	32 (9)
Questions on old scars	14 (4)
Questions on scar or keloid prevention	12 (3.4)

Evaluation of GPT-4–Generated Content

GPT-4 generated content that provided a wide range of medically accurate information. Using the PEMAT-AI, DISCERN-AI, and GQS patient education material evaluation tools, the output of GPT-4 was assessed, with all tools indicating high scores. The overall understandability score using PEMAT-AI easily surpassed the 70% threshold for acceptability (mean 75.5%, SD 12.2%). The DISCERN-AI tool resulted in an overall rating of "good" quality (mean 26.3, SD 3.4), with all 16 groups of questions rated as "good." The GQS score averaged 4.3 out of 5 (SD 0.8), categorizing the outputs as high quality. More details are shown in Table S1 in Multimedia Appendix 6. Intraclass correlation coefficient (ICC) for PEMAT-AI, DISCERN-AI, and GQS were 0.73, 0.69, and 0.78, respectively (Table S2 in Multimedia Appendix 5). The results of the ICC demonstrated high reliability of the evaluation tools.

Plastic Surgeons' Evaluation via the NLAT-AI Tool

Using the NLAT-AI tool, 3 independent plastic surgeons evaluated the GPT-4–generated content. All dimensions of the contents received scores above the neutral midpoint of 3 on a 5-point Likert scale. The overall average scores for each dimension were as follows: accuracy 3.9 (SD 0.7), safety 4.3 (SD 0.8), appropriateness 4.4 (SD 0.5), actionability 4.1 (SD 0.8), and effectiveness 4.1 (SD 0.8). More detailed descriptive statistics for each question are presented in Table S2 in Multimedia Appendix 6. Internal validity tests showed an ICC of 0.76 (Table S2 in Multimedia Appendix 5), indicating high reliability.

Readability Assessment

The results of the readability assessments indicated that the GPT-4–generated content was "difficult to read." The average Flesch Reading Ease score was 50.1 (SD 8.1), which is considered moderately difficult. The Gunning Fog Index averaged 12.7 (SD 3.3), and the Flesch-Kincaid Grade Level was 12.4 (SD 2.5), indicating that the text was at a high school level (approximately suitable for individuals aged 16-17 years). The Coleman-Liau Index averaged 12.8 (SD 2.6), and the SMOG Index averaged 11.3 (SD 3.16). More detailed evaluation results are shown in Table S3 in Multimedia Appendix 6.

Reference Evaluation for AI Assessment

Most of the references provided in GPT-4's output effectively supported the content. A total of 88.2% (2867/3250) of the references were from actual sources (actual websites or academic papers), while 383 hallucinated references were identified. Among these 2867 real references, 2746 (95.8%) references effectively supported the content. In addition, a total of 95.1% (2724/2867) of the real references were from authoritative sources (government guidelines, health care organizations, or scientific research). More detailed evaluation results are shown in Table S4 in Multimedia Appendix 6.

The Assessment of Questions From Websites

The evaluation results of GPT-4 responses to website-sourced questions were broadly consistent with those from Reddit-derived questions across all assessments (Tables S3-S6 in Multimedia Appendix 1).

Discussion

Principal Findings

This is the first study to assess the overall quality of ChatGPT responses to real-world questions from Reddit about keloids or scars. The results revealed that the content generated by GPT-4 was generally comprehensive and aligned with current medical guidelines and the literature. Using several assessment tools, as well as plastic surgeons' evaluations, the scores were robust, and the plastic surgeons' evaluations were largely positive. The overall results indicate that GPT-4–generated content is reliable, accurate, safe, and actionable, despite there being room for improvement in terms of readability and hallucination.

Over 80% of dermatology outpatients obtain medical information through social media or the internet, with 47% considering it an important source of information [28]. Although patients have access to a wealth of information, studies evaluating the quality of online health information have identified significant deficiencies [29]. As for scars and keloids, the information available to patients contains a lot of low-quality content. A previous study assessing 88 websites related to "burn scars" showed that most of the

commercial websites provided information of moderate to poor quality [13]. In contrast, LLMs provide a broad range of fundamentally accurate information and real-time dynamic interactions compared to traditional webpages [30,31]. As a leading LLM, GPT-4 exhibits certain advantages over other LLMs and has demonstrated top-tier performance across diverse evaluations in health care. In answering questions from the American Board of Surgery In-Training Examination, GPT-4 achieved an accuracy rate comparable to that of Copilot, while significantly outperforming Gemini [32]. In other fields of clinical medicine, GPT-4 also attained superior performance relative to other LLMs [33,34]. However, in a substantial number of evaluative scenarios, the performance of GPT-4 did not yield statistically significant differences when compared with Copilot or Gemini. Collectively, the performance of GPT-4 currently represents the best capability of LLMs.

In our study, experienced plastic surgeons evaluated the outputs of GPT-4, confirming that the contents were reliable and accurate. The accuracy of GPT-4 in patient education has also been studied in other clinical contexts (eg, rhinoplasty, sleep apnea, and prostate cancer) where it demonstrated high accuracy and strong reliability [24,35,36]. Such high accuracy and reliability suggest that LLMs such as GPT-4 can effectively address clinical questions from patients with scars or keloids, serving as a valuable auxiliary tool in clinical medicine.

Despite GPT-4's significant potential in responding to keloid or scar patient queries, its outputs commonly had high reading difficulty. Our study revealed that the average reading level of GPT-4–generated content was at a high school level. The results suggest that ChatGPT does not always meet the comprehension needs of all patients. The relatively low readability of GPT-4 can hinder accessibility for certain socioeconomic populations with limited health literacy [37]. Among the latest generation of young adults in the United States, up to 13% have not graduated from high school. This rate reaches 20% among people of color (including African Americans and Native Americans) [38], who are also identified as high-risk groups for developing malignant scars [39]. Due to poor readability, GPT-4 has apparent barriers in its application among these populations [40,41]. To enhance the utility of LLMs for populations with lower educational attainment, it is recommended that developers consider training specialized LLMs based on datasets with good readability [42]. Biomedical text can be simplified through hyperparameter substitution techniques, improving patient understanding [43]. In addition, structured prompting can also contribute to enhancing readability [44].

Moreover, our study also revealed the existence of hallucination, where GPT-4 cited nonexistent references or websites. Fabricated references not only mislead readers and distort their understanding of keloid or scar but also—given the presence of numerous seemingly authoritative yet false information sources—may lead patients to overtrust the content generated by GPT-4 [45,46]. Given the presence of hallucinations, specific clinical diagnosis and treatment must rely on clinicians; LLMs can only serve as auxiliary tools. To address the hallucination issue in LLMs, it is recommended that developers effectively apply retrieval-augmented generation to retrieve documents from an external corpus (such as academic library systems), as this can significantly reduce the hallucinations [47-49]. Integrating external, structured knowledge sources (such as knowledge graphs, databases, or other domain-specific resources) into LLMs can also help ensure that LLMs produce responses with fewer hallucinations [50]. Furthermore, prompt engineering can mitigate hallucination by improving the reasoning capabilities [51].

GPT-4 can provide comprehensive and generally accurate information, which can further assist patients with keloids or scars in accessing timely and precise information. However, current LLMs exhibit limitations, such as hallucinations and relatively low readability; therefore, they are not recommended as the sole source of information for patients. Limited by the lack of clinical background in current LLMs; the insufficient ability to process audio, image, and video information; limited ability to access academic libraries; and the noninterpretability of black box algorithms, current LLMs still require further development to be adapted for applications in health care [52]. The AI agent, as a promising approach, can extend the capabilities of LLMs by enabling them to use external tools, plan and execute multistep tasks, as well as interact dynamically [53]. Multimodal LLM is promising to process text (eg, clinical notes and user-input questions), medical images (eg, photos and computed tomography scans), and videos (eg, treatment procedures) provided by patients, which will more effectively assist patients and health care providers in clinical practice about keloid and scar management [54].

Limitations

Most of the questions collected from Reddit were posts from patients who had not yet sought medical care. Consequently, the questions posed may be biased toward pretreatment information needs, as fewer questions were reported during the treatment phase. This may compromise the generalizability of GPT-4's evaluation across different patient care stages. In addition, Reddit users are concentrated in the age group of 18 to 49 years, with an average age of 23 years, and the majority are aged under 30 years. Thus, the data collected from Reddit clearly fails to represent the middle-aged and older population [55]. Relying solely on Reddit posts for data collection introduces demographic selection bias.

In terms of assessment tools, the qualitative assessment conducted by experienced plastic surgeons was inherently at risk of bias, given the surgeons' attitudes toward the use of GPT-4. Nevertheless, they provided valuable insights owing to their in-depth understanding of scar and keloid education materials. Furthermore, exploratory assessment tools (DISCERN-AI, PEMAT-AI, and NLAT-AI) were used in this study, while their validity requires further testing. LLMs differ from traditional printed educational materials in that their responses to repeated queries of the same question are generated instantaneously and may vary. Currently, existing

assessment tools lack the ability to detect such variability in LLM outputs when the same question is posed multiple times [56]. Furthermore, content generated by LLMs is often conveyed with excessive certainty, as these models lack the ability to accurately express information involving uncertainties. Providing definitive answers to such uncertain content may mislead patients, yet current assessment scales fail to evaluate this critical limitation [57,58]. Further research is needed to develop specific tools to enable more robust evaluation of LLM output quality.

Conclusions

Our analysis found that GPT-4 provided high-quality responses to real-world questions related to scars and keloids, suggesting its potential as a useful patient education tool in scar and keloid treatment. The GPT-4 outputs were generally reliable and accurate but need improvement, primarily in readability and hallucinations.

Funding

This study is funded by the Talent Fund of Guizhou Provincial People's Hospital (awarded to MR; [2023]-30).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Patient Education Materials Assessment Tool for Artificial Intelligence for evaluating the understandability (8 items) and actionability (3 items) of artificial intelligence–generated patient education text.
[DOCX File (Microsoft Word File), 19 KB-Multimedia Appendix 1]

Multimedia Appendix 2

DISCERN-AI tool (7 core items) for assessing artificial intelligence–generated treatment information quality, with 1-3 scoring for each item (relevance, source clarity, date transparency, balance, additional support, uncertainty acknowledgment, and overall quality).
[DOCX File (Microsoft Word File), 21 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Natural Language Assessment Tool for Artificial Intelligence assessment framework: 5 domains (accuracy, safety, appropriateness, actionability, and effectiveness).
[DOCX File (Microsoft Word File), 24 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Reference Evaluation for AI assessment criteria: 3-item tool for validating artificial intelligence–generated references (real or fabricated, content support, and authoritative source status) on large language model reference hallucinations.
[DOCX File (Microsoft Word File), 21 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Supplementary tables for health care website–derived scar or keloid questions: includes 38 unique questions from 3 medical websites (Table S1), intraclass correlation coefficient values for assessment tools (Table S2), and artificial intelligence output evaluation scores (Patient Education Materials Assessment Tool for Artificial Intelligence, DISCERN-AI, Global Quality Scale, Natural Language Assessment Tool for Artificial Intelligence, readability, and reference quality) for website questions (Tables S3-S6).
[DOCX File (Microsoft Word File), 24 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Supplementary tables for Reddit-derived scar or keloid questions: includes subcategory-specific artificial intelligence output evaluation scores for all 16 question groups (Patient Education Materials Assessment Tool for Artificial Intelligence, DISCERN-AI, or Global Quality Scale: Table S1; Natural Language Assessment Tool for Artificial Intelligence: Table S2; and readability metrics: Table S3) and subcategory-specific reference evaluation results (Reference Evaluation for AI) for 3250 total cited references (Table S4), plus overall aggregate scores for all assessments.
[DOCX File (Microsoft Word File), 25 KB-Multimedia Appendix 6]

References

1. Stoica AE, Grumezescu AM, Hermenean AO, Andronescu E, Vasile BS. Scar-free healing: current concepts and future perspectives. Nanomaterials (Basel). Oct 31, 2020;10(11):2179. [doi: 10.3390/nano10112179] [Medline: 33142891]
2. Xiao Y, Sun Y, Zhu B, et al. Risk factors for hypertrophic burn scar pain, pruritus, and paresthesia development. Wound Repair Regen. Mar 2018;26(2):172-181. [doi: 10.1111/wrr.12637] [Medline: 29719102]
3. Hsieh JC, Maisel-Campbell AL, Joshi CJ, Zielinski E, Galiano RD. Daily quality-of-life impact of scars: an interview-based foundational study of patient-reported themes. Plast Reconstr Surg Glob Open. Apr 2021;9(4):e3522. [doi: 10.1097/GOX.0000000000003522] [Medline: 33868874]
4. Schouten H, Nieuwenhuis M, van der Schans C, Niemeijer A, van Zuijlen P. Considerations in determining the severity of burn scar contractures with focus on the knee joint. J Burn Care Res. Jul 5, 2023;44(4):810-816. [doi: 10.1093/jbcr/irad016] [Medline: 36752774]
5. Woodward AP, Matthews CA. Outcomes of revision perineoplasty for persistent postpartum dyspareunia. Female Pelvic Med Reconstr Surg. Mar 2010;16(2):135-139. [doi: 10.1097/SPV.0b013e3181cc8702] [Medline: 22453161]
6. Waibel JS, Waibel H, Sedaghat E. Scar therapy of skin. Facial Plast Surg Clin North Am. Nov 2023;31(4):453-462. [doi: 10.1016/j.fsc.2023.06.005] [Medline: 37806679]
7. Gomolin T, Cline A, Ginsberg D, Safai B. Scar tissue I wish you saw: patient expectations regarding scar treatment. J Cosmet Dermatol. Sep 2021;20(9):2739-2742. [doi: 10.1111/jocd.13945] [Medline: 33434326]
8. Cho SB, Ryu DJ, Lee SJ, et al. Scar characteristics and treatment expectations: a survey of 589 patients. J Cosmet Laser Ther. Dec 2009;11(4):224-228. [doi: 10.3109/14764170903341723] [Medline: 19951193]
9. Andrews N, Jones LL, Moiemen N, et al. Below the surface: parents' views on the factors that influence treatment adherence in paediatric burn scar management - a qualitative study. Burns. May 2018;44(3):626-635. [doi: 10.1016/j.burns.2017.09.003] [Medline: 29031888]
10. Jiang S, Beaudoin CE. Health literacy and the internet: an exploratory study on the 2013 HINTS survey. Comput Human Behav. May 2016;58:240-248. [doi: 10.1016/j.chb.2016.01.007]
11. Battineni G, Baldoni S, Chintalapudi N, et al. Factors affecting the quality and reliability of online health information. Digit Health. 2020;6:2055207620948996. [doi: 10.1177/2055207620948996] [Medline: 32944269]
12. Khaleel I, Wimmer BC, Peterson GM, et al. Health information overload among health consumers: a scoping review. Patient Educ Couns. Jan 2020;103(1):15-32. [doi: 10.1016/j.pec.2019.08.008] [Medline: 31451363]
13. Bohacek L, Gomez M, Fish JS. An evaluation of internet sites for burn scar management. J Burn Care Rehabil. 2003;24(4):246-251; [doi: 10.1097/01.BCR.0000075844.04297.D9] [Medline: 14501424]
14. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed. Mar 2024;245:108013. [doi: 10.1016/j.cmpb.2024.108013] [Medline: 38262126]
15. Neha F, Bhati D, Shukla DK, Amiruzzaman M. ChatGPT: transforming healthcare with AI. AI. 2024;5(4):2618-2650. [doi: 10.3390/ai5040126]
16. Jowsey T, Stokes-Parish J, Singleton R, Todorovic M. Medical education empowered by generative artificial intelligence large language models. Trends Mol Med. Dec 2023;29(12):971-973. [doi: 10.1016/j.molmed.2023.08.012] [Medline: 37718142]
17. Iqbal U, Lee LTJ, Rahmanti AR, Celi LA, Li YCJ. Can large language models provide secondary reliable opinion on treatment options for dermatological diseases? J Am Med Inform Assoc. May 20, 2024;31(6):1341-1347. [doi: 10.1093/jamia/ocae067] [Medline: 38578616]
18. Lee R. Close encounters of the AI kind: the increasingly human-like way people are engaging with language models. Elon University; URL: https://imaginingthedigitalfuture.org/reports-and-publications/close-encounters-of-the-ai-kind/close-encounters-of-the-ai-kind-main-report [Accessed 2025-11-26]
19. Gencer G, Gencer K. Large language models in healthcare: a bibliometric analysis and examination of research trends. J Multidiscip Healthc. 2025;18:223-238. [doi: 10.2147/JMDH.S502351] [Medline: 39844924]
20. Kumar D, Sood SK, Rawat KS. Empowering elderly care with intelligent IoT-driven smart toilets for home-based infectious health monitoring. Artif Intell Med. Oct 2023;144:102666. [doi: 10.1016/j.artmed.2023.102666] [Medline: 37783534]
21. Kumar D, Rawat KS, Sood SK. Revolution of artificial intelligence and IOT in healthcare: a keyword co-occurrence network analysis using CiteSpace. In: Shukla AK, Thakur DG, Arabkoohsar A, editors. Recent Advances in Mechanical Engineering. 2023:231-237. [doi: 10.1007/978-981-99-2349-6_20] ISBN: 9789819923496
22. Wang J, Patel P, Jagdeo J. An analysis of keloid patient questions on Reddit. Wound Repair Regen. 2024;32(2):164-170. [doi: 10.1111/wrr.13160] [Medline: 38372454]
23. Campbell DJ, Estephan LE, Sina EM, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid. Mar 2024;34(3):371-377. [doi: 10.1089/thy.2023.0491] [Medline: 38010917]

24. Gibson D, Jackson S, Shanmugasundaram R, et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. J Med Internet Res. Aug 14, 2024;26:e55939. [doi: 10.2196/55939] [Medline: 39141904]
25. Cassidy JT, Baker JF. Orthopaedic patient information on the world wide web: an essential review. J Bone Joint Surg Am. Feb 17, 2016;98(4):325-338. [doi: 10.2106/JBJS.N.01189] [Medline: 26888683]
26. Readable. Our Readability Checker helps you to communicate clearly. URL: https://readable.com [Accessed 2026-02-24]
27. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. Feb 2023;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
28. AlGhamdi KM, Almohideb MA. Internet use by dermatology outpatients to search for health information. Int J Dermatol. Mar 2011;50(3):292-299. [doi: 10.1111/j.1365-4632.2010.04705.x] [Medline: 21342162]
29. Zhang Y, Kim Y. Consumers' evaluation of web-based health information quality: meta-analysis. J Med Internet Res. Apr 28, 2022;24(4):e36463. [doi: 10.2196/36463] [Medline: 35482390]
30. Monib WK, Qazi A, Mahmud MM. Exploring learners' experiences and perceptions of ChatGPT as a learning tool in higher education. Educ Inf Technol. Jan 2025;30(1):917-939. [doi: 10.1007/s10639-024-13065-4]
31. Jin Y, Liu H, Zhao B, Pan W. ChatGPT and mycosis- a new weapon in the knowledge battlefield. BMC Infect Dis. Oct 27, 2023;23(1):731. [doi: 10.1186/s12879-023-08724-9] [Medline: 37891532]
32. Sanli AN, Tekcan Sanli DE, Karabulut A. Can American Board of Surgery in Training Examinations be passed by large language models? Comparative assessment of Gemini, Copilot, and ChatGPT. Am Surg. Nov 2025;91(11):1923-1929. [doi: 10.1177/00031348251341956] [Medline: 40353502]
33. Gill GS, Tsai J, Moxam J, Sanghvi HA, Gupta S. Comparison of Gemini Advanced and ChatGPT 4.0's performances on the Ophthalmology Resident Ophthalmic Knowledge Assessment Program (OKAP) examination review question banks. Cureus. Sep 2024;16(9):e69612. [doi: 10.7759/cureus.69612] [Medline: 39421095]
34. Tan YZ, Nah SA, Saw SN, Rajandram R, Ong TA. Evaluating the performance of artificial intelligence chatbots in answering urology questions derived from guidelines or board examinations: a systematic review. Urol Sci. 2025:10. [doi: 10.1097/us9.0000000000000089]
35. Lim B, Seth I, Kah S, et al. Using generative artificial intelligence tools in cosmetic surgery: a study on rhinoplasty, facelifts, and blepharoplasty procedures. J Clin Med. Oct 14, 2023;12(20):6524. [doi: 10.3390/jcm12206524] [Medline: 37892665]
36. Incerti Parenti S, Bartolucci ML, Biondi E, et al. Online patient education in obstructive sleep apnea: ChatGPT versus Google search. Healthcare (Basel). Sep 5, 2024;12(17):1781. [doi: 10.3390/healthcare12171781] [Medline: 39273804]
37. Toiv A, Saleh Z, Ishak A, et al. Digesting digital health: a study of appropriateness and readability of ChatGPT-generated gastroenterological information. Clin Transl Gastroenterol. Nov 1, 2024;15(11):e00765. [doi: 10.14309/ctg.0000000000000765] [Medline: 39212302]
38. Veronique I. Report on the condition of education 2024. Institute of Education Sciences; 2024. URL: https://nces.ed.gov/pubs2024/2024144.pdf
39. Oei F, Putra IB, Jusuf NK. The relationship between skin color and keloid. Bali Med J. 2021;10(2):835-838. [doi: 10.15562/bmj.v10i2.2619]
40. DeTemple DE, Meine TC. Comparison of the readability of ChatGPT and Bard in medical communication: a meta-analysis. BMC Med Inform Decis Mak. Sep 1, 2025;25(1):325. [doi: 10.1186/s12911-025-03035-2] [Medline: 40890707]
41. Whittaker P, Sun M. Quality and readability of chatbot responses to patient questions: a systematic cross-sectional meta-synthesis. Health Informatics J. 2025;31(4):14604582251388879. [doi: 10.1177/14604582251388879] [Medline: 41106853]
42. Li M, Zhang Y, Li Z, et al. From quantity to quality: boosting LLM performance with self-guided data selection for instruction tuning. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024:7602-7635; Mexico City, Mexico. [doi: 10.18653/v1/2024.naacl-long.421]
43. Swanson K, He S, Calvano J, et al. Biomedical text readability after hypernym substitution with fine-tuned large language models. PLOS Digit Health. Apr 2024;3(4):e0000489. [doi: 10.1371/journal.pdig.0000489] [Medline: 38625843]
44. Daulat SR, Dholaria N, Burnet G, et al. Prompt engineering and follow-up questioning improves the readability of spine surgery questions in large language models. World Neurosurg. Nov 2025;203:124423. [doi: 10.1016/j.wneu.2025.124423] [Medline: 40889596]

45. Aljohani M, Hou J, Kommu S, Wang X. A comprehensive survey on the trustworthiness of large language models in healthcare. Presented at: Proceedings of the 2025 Findings of the Association for Computational Linguistics EMNLP'25; Nov 4-9, 2025:6720-6748; Suzhou, China. [doi: 10.18653/v1/2025.findings-emnlp.356]
46. Kim Y, Jeong H, Chen S, et al. Medical hallucination in foundation models and their impact on healthcare. Health Systems and Quality Improvement. Preprint posted online on 2025. [doi: 10.1101/2025.02.28.25323115]
47. Zhang W, Zhang J. Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics. 2025;13(5):856. [doi: 10.3390/math13050856]
48. Bhattacharya R. Strategies to mitigate hallucinations in large language models. AMA. 2024;10(1):62. [doi: 10.69554/NXXB8234]
49. Wołk K. Evaluating retrieval-augmented generation variants for clinical decision support: hallucination mitigation and secure on-premises deployment. Electronics (Basel). 2025;14(21):4227. [doi: 10.3390/electronics14214227]
50. Agrawal G, Kumarage T, Alghamdi Z, Liu H. Can knowledge graphs reduce hallucinations in LLMs?: a survey. Presented at: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 16-21, 2024:3947-3960; Mexico City, Mexico. [doi: 10.18653/v1/2024.naacl-long.219]
51. Zhang H, Deng H, Ou J, Feng C. Mitigating spatial hallucination in large language models for path planning via prompt engineering. Sci Rep. 2025;15(1):8881. [doi: 10.1038/s41598-025-93601-5]
52. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). Jan 21, 2025;5(1):26. [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
53. Ferrag MA, Tihanyi N, Debbah M. From LLM reasoning to autonomous ai agents: a comprehensive review. arXiv. Preprint posted online on Apr 28, 2025. [doi: 10.48550/arXiv.2504.19678]
54. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. Sep 25, 2024;26:e59505. [doi: 10.2196/59505] [Medline: 39321458]
55. Reddit user age, gender, & demographics. Exploding Topics. 2025. URL: https://explodingtopics.com/blog/reddit-users [Accessed 2025-11-26]
56. Calloway C. Why do different LLMs give different answers to the same question? Model uncertainty and variability in LLM-based intrusion detection systems ranking. Norfolk State University; 2025. URL: https://digitalcommons.odu.edu/cgi/viewcontent.cgi?article=1123&context=covacci-undergraduateresearch [Accessed 2024-05-20]
57. Yona G, Aharoni R, Geva M. Can large language models faithfully express their intrinsic uncertainty in words? Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Nov 12-16, 2024:632-702; Miami, FL. [doi: 10.18653/v1/2024.emnlp-main.443]
58. Fu T, Conde J, Martínez G, Grandury M, Reviriego P. Multiple choice questions: reasoning makes large language models (LLMS) more self-confident even when they are wrong. arXiv. Preprint posted online on Jan 16, 2025. [doi: 10.48550/arXiv.2501.09775]

Abbreviations

AI: artificial intelligence
GQS: Global Quality Scale
ICC: intraclass correlation coefficient
LLM: large language model
NLAT-AI: Natural Language Assessment Tool for Artificial Intelligence
PEMAT: Patient Education Materials Assessment Tool
SMOG: Simple Measure of Gobbledygook

Edited by Arriel Benis; peer-reviewed by Bárbara Aline Ferreira Assunção, Fumitoshi Fukuzawa, Jörg Klewer, Keshav Singh Rawat; submitted 10.Jun.2025; accepted 29.Dec.2025; published 27.Feb.2026

Please cite as:
Rao M, Xiujun T, Haoyu W
Evaluating GPT-4 Responses on Scars or Keloids for Patient Education: Large Language Model Evaluation Study
JMIR Med Inform 2026;14:e78838
URL: https://medinform.jmir.org/2026/1/e78838
doi: 10.2196/78838