

Original Paper

Quality and Readability of ChatGPT and Gemini Responses to Patient Questions About Systemic Isotretinoin: Comparative Evaluation Study

Elif Bal Avcı¹, MD; Buğra Burç Dağtaş², MD; Dilara İlhan Erdil³, MD; Ozan Erdem⁴, MD; Andaç Salman⁵, Prof Dr; Selçuk Korkmaz⁶, Prof Dr; Sezgi Sarıkaya Solak¹, Prof Dr

¹Department of Dermatology, Faculty of Medicine, Trakya University, Edirne, Türkiye

²Department of Dermatology, Istanbul Training and Research Hospital, Istanbul, Türkiye

³Department of Dermatology, Etlik City Hospital, Ankara, Türkiye

⁴Department of Dermatology, Medeniyet University Faculty of Medicine, Göztepe Prof. Dr. Suleyman Yalcin City Hospital, Istanbul, Türkiye

⁵Department of Dermatology, School of Medicine, Acıbadem Mehmet Ali Aydınlar University, Istanbul, Türkiye

⁶Department of Biostatistics and Medical Informatics, Trakya University Faculty of Medicine, Edirne, Türkiye

Corresponding Author:

Elif Bal Avcı, MD
Department of Dermatology
Faculty of Medicine, Trakya University
Trakya University Balkan Campus, Iskender
Edirne
Türkiye
Phone: 90 5054511134
Email: drelifbalavci@gmail.com

Abstract

Background: The use of AI in health care is rapidly expanding due to its accessibility, with growing reliance among both physicians and patients.

Objective: This study evaluated ChatGPT (OpenAI) and Gemini (Google) responses to patient questions on systemic isotretinoin.

Methods: A total of 40 commonly anticipated patient questions were developed based on online sources and patient education materials and categorized into 5 groups. Responses were generated using the January 2025 free versions of ChatGPT (GPT-4o) and Gemini 2.0 Flash. A total of 5 experienced dermatologists evaluated the responses in a blinded manner using a 5-point Likert scale assessing accuracy, clarity, completeness, usefulness, and overall quality score (Global Quality Score [GQS]). Readability was measured using Flesch-Kincaid Grade Level (FKGL), Flesch Reading Ease, and Simple Measure of Gobbledygook (SMOG) index scores. Mixed-effects modeling was performed to adjust for reviewer and question variability, and intraclass correlation coefficients (ICCs) were calculated to assess interrater reliability.

Results: ChatGPT achieved significantly higher ratings in accuracy, completeness, usefulness, and GQS, while clarity showed a nonsignificant trend. When questions were analyzed by subtype, no statistically significant differences were observed between models across evaluation parameters. Mixed-effects modeling confirmed higher odds of superior ratings for ChatGPT across domains (odds ratios [ORs] 1.82-5.88; all $P < .001$). However, interrater reliability was lower for ChatGPT (global ICC 0.401) compared with Gemini (global ICC 0.677). Readability analyses demonstrated that ChatGPT required a higher reading level (FKGL 12.6 vs 11.2; SMOG 11.3 vs 10.0).

Conclusions: ChatGPT outperformed Gemini in terms of data accuracy and quality, while Gemini offered superior readability and consistency. Both models exceeded recommended readability thresholds, highlighting a gap between informational quality and linguistic accessibility in AI-generated health communication.

JMIR Dermatol 2026;9:e99780; doi: [10.2196/99780](https://doi.org/10.2196/99780)

Keywords: artificial intelligence; AI; large language models; patient education; readability; systemic isotretinoin

Introduction

Systemic isotretinoin is among the most effective treatments for moderate-to-severe acne [1]. However, its use involves significant adverse effects, strict monitoring, and complex counseling. As a result, patients frequently seek detailed information on efficacy, safety, laboratory monitoring, teratogenic risks, and long-term outcomes [2]. As health information becomes increasingly digitalized, patients are increasingly using online sources to supplement information provided by physicians [3]. AI fulfills diverse roles in dermatology, such as facilitating diagnosis, monitoring disease progression, guiding treatment selection, providing patient information, and supporting medical education [4,5]. Numerous studies across various medical specialties have used AI language models to evaluate responses to patient questions. However, standardized guidelines for performance assessment are lacking, and the methodological details of these studies often remain unclear [6]. Several studies have assessed the performance of large language models (LLMs) in responding to patient questions about acne and systemic isotretinoin, with mixed results for accuracy and quality [7-11]. Due to isotretinoin's teratogenicity, monitoring needs, and medicolegal considerations, it is essential that the information provided is both accurate and easily understandable.

The study aimed to comparatively evaluate ChatGPT and Gemini in their responses to patient questions regarding systemic isotretinoin, assessing performance across multiple quality domains, interrater reliability, and readability metrics.

Methods

Study Design

This cross-sectional comparative study assessed the quality of responses generated by 2 LLMs to patient-oriented questions concerning systemic isotretinoin. The evaluation criteria included accuracy, clarity, completeness, usefulness, Global Quality Score (GQS), and readability.

Question Development

Forty patient-oriented questions were developed based on information from patient information brochures and online patient forum queries related to isotretinoin. To prevent confusion between topical and systemic formulations, the standardized term "oral isotretinoin" was included in each question. The questions were categorized into 5 domains: general information, efficacy and treatment process, side effects, follow-up and precautions, and cosmetic concerns. Six experienced dermatologists (EBA, BBD, OE, DIE, AS, and SSS) reviewed and validated all questions to ensure content relevance and representativeness prior to inclusion. The complete list of the 40 isotretinoin-related questions evaluated in this study is provided in [Multimedia Appendix 1](#).

Evaluation of AI Models and Query Procedures

Free access versions of Gemini 2.0 Flash (Google Inc) and GPT-4o (OpenAI) were evaluated through their standard web-based interfaces to reflect real-world patient use. All queries were conducted in January 2025. The models were not accessed via APIs; therefore, API-level parameters, including temperature, were not accessible or configurable by the investigators. Each question was entered verbatim, without additional prompts or role instructions. Queries were submitted without logging into a user account, using incognito browsing mode, and each question was asked in a separate browser session to minimize the influence of prior conversational context. Each question was submitted only once to each model, and the first generated response was used for evaluation, with no regeneration or repeated querying, to simulate a typical real-world patient interaction. Responses were recorded verbatim without modification.

Blinded Expert Assessment (Accuracy, Clarity, Completeness, Usefulness, and the GQS)

The questions were entered into the 2 chatbots by 1 dermatologist (EBA) to generate the responses, and the resulting outputs were independently evaluated by 5 experienced dermatologists (≥ 5 years of experience; BBD, OE, DIE, AS, and SSS) in a blinded fashion. Interrater agreement was assessed using the intraclass correlation coefficient (ICC). Responses were scored on a 5-point Likert scale (1=very poor, 5=excellent) across 5 domains: accuracy, clarity, completeness, usefulness, and the GQS. A standardized scoring rubric was provided to evaluators prior to assessment. Evaluations were performed by 5 experienced dermatologists, informed by contemporary acne treatment guidelines and clinical practice standards. Readability was assessed using Readability Formulas [12]. The following indices were calculated for each response: Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), and Simple Measure of Gobbledygook (SMOG) index. Source citation was not included as a comparative parameter because the reference provision differed structurally between models: Gemini frequently generated citations, while ChatGPT did not consistently provide explicit references.

Statistical Analysis

Data were analyzed using R (version 4.3.2; R Foundation for Statistical Computing). Descriptive statistics are presented as mean, median, and range (minimum-maximum), as appropriate. Ratings assigned to ChatGPT and Gemini responses to the same questions by the same reviewer panel were compared using the Wilcoxon signed-rank test for paired samples. These comparisons were performed for the pooled dataset and separately within each of 5 predefined clinical domains (general information, efficacy and treatment process, side effects, follow-up and precautions, and cosmetic concerns), resulting in 30 pairwise comparisons (5 metrics across 6 analyses). Within each analysis, Bonferroni correction was applied across the 5 metric-level

comparisons (adjusted $P = \min[\text{raw } P \times 5, 1.0]$) to control the family-wise error rate. A Bonferroni-adjusted $P < .05$ was considered statistically significant. To account for the hierarchical structure of the ratings and repeated assessments by evaluators and across questions, model performance for accuracy, clarity, completeness, usefulness, and GQS was additionally evaluated using ordinal mixed-effects regression models, with evaluator and question included as random intercepts. Effect estimates are reported as odds ratios (ORs) with 95% CIs. Interrater reliability was assessed using 2-way mixed-effects ICCs for absolute agreement. All tests were 2-sided, with a significance level set at $P < .05$.

Ethical Considerations

This study did not involve human participants, animal subjects, or patient data. The analyzed data consisted solely of responses generated by publicly available AI chatbots to questions prepared by the authors. Therefore, ethics committee approval was not required.

Results

Expert Assessment of AI Responses

ChatGPT demonstrated consistent numerical outperformance across all assessed domains (Table 1).

In the domain-specific analysis, no statistically significant differences between ChatGPT and Gemini were detected across the 5 predefined domains (general information, efficacy and treatment process, side effects, follow-up and precautions, and cosmetic concerns) for any of the assessed metrics after Bonferroni adjustment (all adjusted $P > .05$). Although ChatGPT showed numerically higher mean scores across all metrics in the cosmetic concerns domain, including accuracy (4.60 vs 3.80) and completeness (4.60 vs 3.60), these differences were not statistically significant after adjustment. The complete results of the domain-specific comparisons are presented in Table 2.

Table 1. Performance comparison of ChatGPT and Gemini across key quality metrics.

Metric	ChatGPT, mean (SD; range)	ChatGPT, median (IQR)	Gemini, mean (SD; range)	Gemini, median (IQR)	<i>P</i> value	Adjusted <i>P</i> value
Accuracy	4.37 (0.71; 3-5)	5 (4-5)	4.20 (0.65; 3-5)	4 (4-5)	.004	.02
Clarity	4.42 (0.72; 3-5)	5 (4-5)	4.27 (0.62; 3-5)	4 (4-5)	.01	.07
Completeness	4.39 (0.75; 3-5)	5 (4-5)	3.77 (0.75; 1-5)	4 (3-4)	<.001	<.001
Usefulness	4.42 (0.70; 3-5)	5 (4-5)	4.04 (0.69; 2-5)	4 (4-4)	<.001	<.001
GQS ^a	4.42 (0.71; 3-5)	5 (4-5)	4.05 (0.70; 2-5)	4 (4-5)	<.001	<.001

^aGQS: Global Quality Score.

Table 2. Domain-specific comparison of performance metrics between ChatGPT and Gemini across 5 predefined domains.

Domain and metric	ChatGPT, mean (SD; range)	ChatGPT, median (IQR)	Gemini, mean (SD; range)	Gemini, median (IQR)	<i>P</i> value	Adjusted <i>P</i> value
General information						
Accuracy	4.00 (0.00; 4-4)	4 (4-4)	4.00 (0.00; 4-4)	4 (4-4)	— ^a	—
Clarity	4.00 (0.00; 4-4)	4 (4-4)	4.33 (0.58; 4-5)	4 (4-4.5)	>.99	>.99
Completeness	3.67 (0.58; 3-4)	4 (3.5-4)	3.33 (0.58; 3-4)	3 (3-3.5)	>.99	>.99
Usefulness	4.00 (0.00; 4-4)	4 (4-4)	4.00 (0.00; 4-4)	4 (4-4)	—	—
GQS ^b	4.00 (0.00; 4-4)	4 (4-4)	4.00 (0.00; 4-4)	4 (4-4)	—	—
Efficacy and treatment process						
Accuracy	3.89 (0.33; 3-4)	4 (4-4)	3.56 (0.53; 3-4)	4 (3-4)	.15	.75
Clarity	4.00 (0.71; 3-5)	4 (4-4)	4.00 (0.50; 3-5)	4 (4-4)	>.99	>.99
Completeness	3.78 (0.44; 3-4)	4 (4-4)	3.44 (1.01; 1-4)	4 (3-4)	.37	>.99
Usefulness	3.89 (0.33; 3-4)	4 (4-4)	3.67 (0.71; 2-4)	4 (4-4)	.35	>.99
GQS	3.78 (0.44; 3-4)	4 (4-4)	3.56 (0.73; 2-4)	4 (3-4)	.35	>.99
Side effects						
Accuracy	3.82 (0.40; 3-4)	4 (4-4)	3.91 (0.30; 3-4)	4 (4-4)	.77	>.99
Clarity	3.91 (0.30; 3-4)	4 (4-4)	4.09 (0.70; 3-5)	4 (4-4.5)	.48	>.99
Completeness	3.91 (0.54; 3-5)	4 (4-4)	3.55 (0.52; 3-4)	4 (3-4)	.20	>.99
Usefulness	3.91 (0.30; 3-4)	4 (4-4)	3.82 (0.40; 3-4)	4 (4-4)	.77	>.99
GQS	4.00 (0.00; 4-4)	4 (4-4)	3.91 (0.54; 3-5)	4 (3-4)	.77	>.99
Follow-up and precautions						
Accuracy	4.00 (0.43; 3-5)	4 (4-4)	4.00 (0.00; 4-4)	4 (4-4)	>.99	>.99

Domain and metric	ChatGPT, mean (SD; range)	ChatGPT, median (IQR)	Gemini, mean (SD; range)	Gemini, median (IQR)	<i>P</i> value	Adjusted <i>P</i> value
Clarity	4.50 (0.52; 4-5)	4.5 (4-5)	4.50 (0.52; 4-5)	4.5 (4-5)	>.99	>.99
Completeness	4.25 (0.45; 4-5)	4 (4-4.25)	3.92 (0.29; 3-4)	4 (4-4)	.07	.36
Usefulness	4.25 (0.45; 4-5)	4 (4-4.25)	4.08 (0.29; 4-5)	4 (4-4)	.42	>.99
GQS	4.50 (0.52; 4-5)	4.5 (4-5)	4.33 (0.49; 4-5)	4 (4-5)	.53	>.99
Cosmetic concerns						
Accuracy	4.60 (0.55; 4-5)	5 (4-5)	3.80 (0.45; 3-4)	4 (4-4)	.07	.36
Clarity	4.40 (0.55; 4-5)	4 (4-5)	3.80 (0.45; 3-4)	4 (4-4)	.15	.75
Completeness	4.60 (0.55; 4-5)	5 (4-5)	3.60 (0.89; 2-4)	4 (4-4)	.09	.45
Usefulness	4.60 (0.55; 4-5)	5 (4-5)	3.80 (0.45; 3-4)	4 (4-4)	.07	.36
GQS	4.60 (0.55; 4-5)	5 (4-5)	3.80 (0.45; 3-4)	4 (4-4)	.07	.36

^a*P* value was not calculated because there was no variation in the corresponding scores.

^bGQS: Global Quality Score.

Ordinal mixed-effects models revealed that ChatGPT was significantly more likely than Gemini to receive higher ratings across all metrics (all $P < .001$). ChatGPT outperformed Gemini in accuracy (OR 1.94, 95% CI 1.28-2.94), clarity (OR 1.82, 95% CI 1.19-2.76), and GQS (OR 3.55, 95% CI 2.32-5.43). The most substantial advantages were observed in completeness (OR 5.88, 95% CI 3.84-9.02) and usefulness (OR 3.74, 95% CI 2.44-5.74). Representative examples of responses generated by ChatGPT and Gemini are provided in [Multimedia Appendix 2](#).

Interrater Reliability

Global ICC estimates indicated “moderate” consistency for Gemini (ICC 0.677) and “poor” consistency for ChatGPT (ICC 0.401; [Table 3](#)). Although both models exceeded chance agreement ($P < .05$), Gemini demonstrated significantly more stable evaluator consensus.

Table 3. Comparison of interrater reliability (intraclass correlation coefficient [ICC]) between ChatGPT and Gemini across key evaluation domains.

Metric	ChatGPT, ICC (95% CI)	Gemini, ICC (95% CI)	Interpretation (ChatGPT)	Interpretation (Gemini)
Accuracy	0.435 (0.101-0.671)	0.554 (0.291-0.740)	Poor	Moderate
Clarity	0.389 (0.028-0.644)	0.629 (0.410-0.784)	Poor	Moderate
Completeness	0.379 (0.012-0.638)	0.527 (0.248-0.725)	Poor	Moderate
Usefulness	0.393 (0.035-0.647)	0.593 (0.353-0.763)	Poor	Moderate
GQS ^a	0.462 (0.145-0.687)	0.636 (0.421-0.788)	Poor	Moderate
Global ICC	0.401 (0.048-0.651)	0.677 (0.486-0.812)	Poor	Moderate

^aGQS: Global Quality Score

Readability

Gemini demonstrated superior accessibility across multiple readability metrics ([Table 4](#)). Significant reductions in grade-level requirements (mean 11.2 vs 12.6; adjusted $P = .002$) and SMOG index scores (adjusted $P = .001$) were

observed for Gemini compared with ChatGPT. Although Gemini also achieved higher raw FRE scores, this difference did not reach statistical significance after Bonferroni adjustment ($P = .09$).

Table 4. Comparison of readability indices between ChatGPT and Gemini.

Index	ChatGPT, mean (SD; range)	ChatGPT, median (IQR)	Gemini, mean (SD; range)	Gemini, median (IQR)	<i>P</i> value	Adjusted <i>P</i> value
FRE ^a	37.9 (10.26; 17-72)	38.00 (31.00-42.25)	42.03 (12.62; 20-65)	42.5 (32.00-51.75)	.03	.09
FKGL ^b	12.62 (1.95; 7.7-16.59)	12.44 (11.48-13.51)	11.16 (2.13; 7.15-15.14)	10.805 (32.00-51.75)	<.001	.002
SMOG ^c	11.31 (1.65; 7.78-15.37)	11.31 (10.40-12.10)	10.03 (1.63; 7.24-13.62)	9.865 (8.79-11.12)	<.001	.001

^aFRE: Flesch Reading Ease.

^bFKGL: Flesch-Kincaid Grade Level.

^cSMOG: Simple Measure of Gobbledygook.

Discussion

In this study, ChatGPT demonstrated significantly higher performance across evaluation metrics, including accuracy, completeness, clarity, and overall quality, while Gemini showed greater interrater reliability and superior readability. Accordingly, the choice of LLM may depend on whether the priority is high-quality medical content or improved patient comprehension.

Recent evaluations indicate that LLMs generate more contextually appropriate medical responses than traditional internet-based resources [13]. ChatGPT has demonstrated potential as a public dermatology resource, particularly for common dermatological diseases, with 88% of its responses rated as appropriate [14]. However, the clinical utility of LLMs remains constrained by their propensity to generate fictitious citations [15,16]. Their inability to account for individual patient factors and age-related nuances may also lead to misleading guidance, potentially resulting in adverse outcomes based on unverified data [11].

The literature further demonstrates that LLM accuracy and comprehensiveness ratings vary considerably across dermatologic diseases and topics [11]. For example, ChatGPT's accuracy has ranged from 36% to 90% across various topics [6,14,17,18]. ChatGPT showed higher accuracy in dermatological topics with well-established evidence, such as melanoma and nonmelanoma skin cancers, rosacea, and alopecia. In contrast, lower accuracy was observed in areas requiring individualized approaches or where scientific consensus remains limited, including the effects of diet on skin diseases, nonpharmacological treatments for eczema and psoriasis, antiaging interventions, skincare routines, and wound healing [14]. In a study examining isotretinoin-related queries, ChatGPT's accuracy varied by domain: daily life achieved 72% reliability, while the treatment process yielded 64% with outdated or incorrect data [7]. In contrast, our findings showed no significant differences across query categories.

Our findings contradict previous research indicating that Gemini provided more accurate answers than ChatGPT on isotretinoin questions [8], as well as studies asserting Gemini's superiority in comprehensiveness in melanoma-related queries [18]. Recent evidence regarding Mohs surgery indicates that GPT-4 outperforms Gemini (Google Bard) in clinical relevance and accuracy, providing answers that closely align with official American College of Mohs Surgery recommendations [19], suggesting that differences in topic may significantly influence the performance of LLMs. Furthermore, such discrepancies may result from the absence of standardized evaluation criteria and differences among LLM models and versions.

Consistent with prior research on isotretinoin, which reported a low Fleiss κ ($\kappa=0.045$) [7], our study also demonstrated similarly weak interrater agreement. Consistent with our findings, Gemini demonstrated higher interrater agreement than ChatGPT in a study of common dermatological conditions [20]. This low interrater agreement may be

partially attributed to heterogeneity in physicians' clinical communication styles and expectations in patient education. To address inconsistent evaluations, it has been proposed that LLM outputs be assessed using standardized, guideline-based frameworks [21]. Nevertheless, the complex dynamics of physician-patient interactions and the subjective nature of clinical judgment continue to influence evaluation outcomes.

Previous studies suggest that AI-generated health information often exceeds recommended readability levels, potentially limiting accessibility for individuals with low health literacy [21-23]. Recent studies have shown that successive iterations of LLMs exhibit progressive improvement in generating more readable and accessible content [24,25]. In our study, both ChatGPT and Gemini produced patient information at readability levels exceeding recommended thresholds for public education materials. Consistent with our results, some studies report higher readability for Gemini [18], whereas others find no significant readability differences between Gemini and ChatGPT [8]. High readability demands, incomplete incorporation of updated approaches, and limited actionable guidance preclude the use of LLM outputs as standalone patient education tools [7,8]. For high-risk medications such as isotretinoin, patients should be explicitly informed that LLM-generated outputs constitute preliminary information and do not substitute for professional medical guidance. Given the substantial patient concerns about isotretinoin [26-28], clear and accessible educational materials may mitigate misconceptions, reduce treatment resistance, and improve adherence.

Additionally, ChatGPT can be directed to create patient brochures at predefined reading levels, thereby improving accessibility for diverse patient populations [18,29]. Effective prompt engineering enables LLMs to deliver clearer, more targeted responses [8,30-32]. Incorporating an audience-adaptive response model in which users identify as health care professionals or laypersons prior to output generation could facilitate flexible adjustment of terminology and grade-level complexity.

Gemini consistently advised consultation with a health care professional, whereas this was not observed in our study with ChatGPT, despite prior reports of such disclaimers [33, 34]. This suggests that Gemini may offer a comparatively safer profile for patient safety.

The strengths of this study include methodological transparency, a comprehensive multidimensional evaluation approach, a comparative design, and the provision of actionable recommendations. This study has several limitations. The selected questions may not accurately represent real-world patient inquiries. The study was conducted exclusively in English, and language-dependent variability in AI model training data and source retrieval may limit generalizability across different languages. Moreover, despite blinded assessment, clinicians familiar with AI language models may have inferred the model's identity, potentially introducing observer bias. Another limitation is that our analysis included only ChatGPT and Gemini. Therefore, the findings may not be generalizable to other

commercially available or open-source LLMs. Given the rapid evolution of LLMs, the findings of this study reflect the performance of GPT-4o and Gemini 2.0 Flash at the time of evaluation. As newer model iterations are released with enhanced reasoning and language capabilities, ongoing comparative studies are required to determine whether these advancements yield measurable improvements in clinical quality, reliability, and readability. Furthermore, because LLMs are frequently updated, their responses and performance may change over time. Future studies should include a broader range of models and evaluate changes across model versions longitudinally.

In conclusion, there is a clear distinction between ChatGPT's technical accuracy and Gemini's linguistic accessibility. For these models to function as reliable clinical tools, transparent source attribution and patient-centered, intelligible outputs are imperative. Adopting standardized, objective evaluation frameworks is also crucial for enabling robust comparisons of outputs and tracking model performance over time, thereby mitigating interrater variability. However, the complexity and heterogeneity of patient characteristics make it unlikely that clinicians will be fully substituted for by AI, particularly when it comes to contextualizing clinical evidence for individual patients.

Acknowledgments

During the preparation of this manuscript, ChatGPT (OpenAI, GPT-5 series) was used solely for English language editing, including grammar, spelling, and clarity improvements. Different versions within the GPT-5 series may have been used at different stages of manuscript preparation, as automatic model selection was enabled. The authors reviewed and edited all AI-assisted revisions and take full responsibility for the final content of the manuscript.

Funding

This study received no external funding.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request and with appropriate justification.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of questions.

[\[DOCX File \(Microsoft Word File\), 17 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Example questions and answers.

[\[DOCX File \(Microsoft Word File\), 34 KB-Multimedia Appendix 2\]](#)

References

1. Reynolds RV, Yeung H, Cheng CE, et al. Guidelines of care for the management of acne vulgaris. *J Am Acad Dermatol*. May 2024;90(5):1006.e1-1006.e30. [doi: [10.1016/j.jaad.2023.12.017](https://doi.org/10.1016/j.jaad.2023.12.017)] [Medline: [38300170](https://pubmed.ncbi.nlm.nih.gov/38300170/)]
2. Sadeghzadeh-Bazargan A, Ghassemi M, Goodarzi A, Roohaninasab M, Najar Nobari N, Behrangi E. Systematic review of low-dose isotretinoin for treatment of acne vulgaris: focus on indication, dosage, regimen, efficacy, safety, satisfaction, and follow up, based on clinical studies. *Dermatol Ther*. Jan 2021;34(1):e14438. [doi: [10.1111/dth.14438](https://doi.org/10.1111/dth.14438)] [Medline: [33085149](https://pubmed.ncbi.nlm.nih.gov/33085149/)]
3. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. Apr 25, 2023;11:1166120. [doi: [10.3389/fpubh.2023.1166120](https://doi.org/10.3389/fpubh.2023.1166120)] [Medline: [37181697](https://pubmed.ncbi.nlm.nih.gov/37181697/)]
4. Sohail N, Puyana C, Zimmerman L, Tsoukas MM. Artificial intelligence in dermatology: bridging the gap in patient care and education. *Clin Dermatol*. Sep 2024;42(5):434-436. [doi: [10.1016/j.clindermatol.2024.06.009](https://doi.org/10.1016/j.clindermatol.2024.06.009)] [Medline: [38936639](https://pubmed.ncbi.nlm.nih.gov/38936639/)]
5. Gabashvili IS. ChatGPT in dermatology: a comprehensive systematic review. *medRxiv*. Preprint posted online on Jun 12, 2023. [doi: [10.2196/preprints.48979](https://doi.org/10.2196/preprints.48979)]
6. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform*. Mar 2024;151:104620. [doi: [10.1016/j.jbi.2024.104620](https://doi.org/10.1016/j.jbi.2024.104620)] [Medline: [38462064](https://pubmed.ncbi.nlm.nih.gov/38462064/)]
7. Küçük R, Güngör CB, Topal İO, Küçük RB, Karadağ AS. Accuracy, readability, understandability, and actionability of ChatGPT responses to patient questions on systemic isotretinoin treatment. *Indian Dermatol Online J*. Jan 1, 2026;17(1):27-32. [doi: [10.4103/idoj.idoj_1243_24](https://doi.org/10.4103/idoj.idoj_1243_24)] [Medline: [41437721](https://pubmed.ncbi.nlm.nih.gov/41437721/)]
8. Soysal MÇ. Performance of artificial intelligence large language models (LLMs) in answering frequently asked questions about isotretinoin. *Cutan Ocul Toxicol*. 2026;45(1):58-63. [doi: [10.1080/15569527.2025.2601639](https://doi.org/10.1080/15569527.2025.2601639)] [Medline: [41410102](https://pubmed.ncbi.nlm.nih.gov/41410102/)]

9. Frederickson KL, Gui H, Barbieri JS, Daneshjou R. Artificial intelligence use in acne diagnosis and management-a scoping review. *Int J Dermatol*. Mar 2026;65(3):437-443. [doi: [10.1111/ijd.70110](https://doi.org/10.1111/ijd.70110)] [Medline: [41199593](https://pubmed.ncbi.nlm.nih.gov/41199593/)]
10. Deoghare S. An interesting conversation with ChatGPT about acne vulgaris. *Indian Dermatol Online J*. Nov 7, 2023;15(1):137-140. [doi: [10.4103/idoj.idoj_77_23](https://doi.org/10.4103/idoj.idoj_77_23)] [Medline: [38283020](https://pubmed.ncbi.nlm.nih.gov/38283020/)]
11. Lakdawala N, Channa L, Gronbeck C, et al. Assessing the accuracy and comprehensiveness of ChatGPT in offering clinical guidance for atopic dermatitis and acne vulgaris. *JMIR Dermatol*. Nov 14, 2023;6:e50409. [doi: [10.2196/50409](https://doi.org/10.2196/50409)] [Medline: [37962920](https://pubmed.ncbi.nlm.nih.gov/37962920/)]
12. Readability Formulas. URL: https://readabilityformulas.com/#google_vignette [Accessed 2026-10-11]
13. 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*. Jun 1, 2023;183(6):589-596. [doi: [10.1001/jamainternmed.2023.1838](https://doi.org/10.1001/jamainternmed.2023.1838)] [Medline: [37115527](https://pubmed.ncbi.nlm.nih.gov/37115527/)]
14. Ferreira AL, Chu B, Grant-Kels JM, Ogunleye T, Lipoff JB. Evaluation of ChatGPT dermatology responses to common patient queries. *JMIR Dermatol*. Nov 17, 2023;6:e49280. [doi: [10.2196/49280](https://doi.org/10.2196/49280)] [Medline: [37976093](https://pubmed.ncbi.nlm.nih.gov/37976093/)]
15. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. Mar 19, 2023;11(6):887. [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
16. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nat Med*. Aug 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
17. D'Agostino M, Feo F, Martora F, et al. ChatGPT and dermatology. *Ital J Dermatol Venerol*. Oct 2024;159(5):566-571. [doi: [10.23736/S2784-8671.24.07928-3](https://doi.org/10.23736/S2784-8671.24.07928-3)] [Medline: [39039954](https://pubmed.ncbi.nlm.nih.gov/39039954/)]
18. Kamminga NC, Kievits JE, 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*. Jan 24, 2025;192(2):306-315. [doi: [10.1093/bjd/ljae377](https://doi.org/10.1093/bjd/ljae377)] [Medline: [39365602](https://pubmed.ncbi.nlm.nih.gov/39365602/)]
19. 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*. May 2024;90(5):1078-1080. [doi: [10.1016/j.jaad.2024.01.037](https://doi.org/10.1016/j.jaad.2024.01.037)] [Medline: [38296195](https://pubmed.ncbi.nlm.nih.gov/38296195/)]
20. Podder I, Pipil N, Dhabal A, Mondal S, Pienyii V, Mondal H. Evaluation of artificial intelligence-based chatbot responses to common dermatological queries. *Jordan Med J*. 2024;58(3):271-278. [doi: [10.35516/jmj.v58i2.2960](https://doi.org/10.35516/jmj.v58i2.2960)]
21. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw Open*. Feb 3, 2025;8(2):e2457879. [doi: [10.1001/jamanetworkopen.2024.57879](https://doi.org/10.1001/jamanetworkopen.2024.57879)] [Medline: [39903463](https://pubmed.ncbi.nlm.nih.gov/39903463/)]
22. Chen AI, Ferris LK, Nambudiri VE, Piette EW. ChatRx: ChatGPT's potential to educate patients on medication adverse effects. *J Am Acad Dermatol*. Mar 2024;90(3):669-670. [doi: [10.1016/j.jaad.2023.11.008](https://doi.org/10.1016/j.jaad.2023.11.008)] [Medline: [37952568](https://pubmed.ncbi.nlm.nih.gov/37952568/)]
23. Andrikyan W, Sametinger SM, Kosfeld F, et al. Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients. *BMJ Qual Saf*. Jan 28, 2025;34(2):100-109. [doi: [10.1136/bmjqs-2024-017476](https://doi.org/10.1136/bmjqs-2024-017476)] [Medline: [39353736](https://pubmed.ncbi.nlm.nih.gov/39353736/)]
24. Lambert R, Choo ZY, Gradwohl K, Schroedl L, Ruiz De Luzuriaga A. Assessing the application of large language models in generating dermatologic patient education materials according to reading level: qualitative study. *JMIR Dermatol*. May 16, 2024;7:e55898. [doi: [10.2196/55898](https://doi.org/10.2196/55898)] [Medline: [38754096](https://pubmed.ncbi.nlm.nih.gov/38754096/)]
25. Rames MM, O'Hern K, Rames JD, et al. ChatGPT as a resource for patient education in cosmetic dermatological procedures: a boon or a bane? *J Cosmet Dermatol*. Mar 2024;23(3):1085-1086. [doi: [10.1111/jocd.16166](https://doi.org/10.1111/jocd.16166)] [Medline: [38168902](https://pubmed.ncbi.nlm.nih.gov/38168902/)]
26. Al-Harbi M. Concerns and awareness of acne patients about isotretinoin in Qassim region of Saudi Arabia. *Int J Health Sci (Qassim)*. Jan 2010;4(1):47-51. [Medline: [21475525](https://pubmed.ncbi.nlm.nih.gov/21475525/)]
27. Jakobi AU, Bircher AJ, Pagnamenta A, Terrani I. Isotretinoin concerns in Switzerland: a student-based transversal study. *J Clin Med*. Mar 7, 2025;14(6):1801. [doi: [10.3390/jcm14061801](https://doi.org/10.3390/jcm14061801)] [Medline: [40142609](https://pubmed.ncbi.nlm.nih.gov/40142609/)]
28. Kara Polat A, Akin Belli A, Ergun EZ, et al. Knowledge levels and concerns about oral isotretinoin treatment in the parents of adolescent acne patients. *Dermatol Ther*. Jul 2020;33(4):e13669. [doi: [10.1111/dth.13669](https://doi.org/10.1111/dth.13669)] [Medline: [32459383](https://pubmed.ncbi.nlm.nih.gov/32459383/)]
29. Breneman A, Trager MH, Gordon ER, Samie FH. Readability rescue: large language models may improve readability of patient education materials. *Arch Dermatol Res*. Oct 10, 2024;316(9):669. [doi: [10.1007/s00403-024-03441-y](https://doi.org/10.1007/s00403-024-03441-y)] [Medline: [39387945](https://pubmed.ncbi.nlm.nih.gov/39387945/)]
30. Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinform*. Nov 22, 2023;25(1):bbad493. [doi: [10.1093/bib/bbad493](https://doi.org/10.1093/bib/bbad493)]

31. Russe MF, Reiser M, Bamberg F, Rau A. Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. *Rofo*. Nov 2024;196(11):1166-1170. [doi: [10.1055/a-2264-5631](https://doi.org/10.1055/a-2264-5631)] [Medline: [38408477](https://pubmed.ncbi.nlm.nih.gov/38408477/)]
32. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res*. Oct 4, 2023;25:e50638. [doi: [10.2196/50638](https://doi.org/10.2196/50638)] [Medline: [37792434](https://pubmed.ncbi.nlm.nih.gov/37792434/)]
33. Dulai A, Tarmaster A, Sivamani R. Assessment of an artificial intelligence model's capability of responding to queries in the realm of integrative dermatology. *JOID*. Jan 2025;1(1). [doi: [10.64550/joid.0hy8wz51](https://doi.org/10.64550/joid.0hy8wz51)]
34. Reynolds K, Tejasvi T. Potential use of ChatGPT in responding to patient questions and creating patient resources. *JMIR Dermatol*. Mar 6, 2024;7:e48451. [doi: [10.2196/48451](https://doi.org/10.2196/48451)] [Medline: [38446541](https://pubmed.ncbi.nlm.nih.gov/38446541/)]

Abbreviations

FKGL: Flesch-Kincaid Grade Level

FRE: Flesch Reading Ease

GQS: Global Quality Score

ICC: intraclass correlation coefficient

LLM: large language model

OR: odds ratio

Edited by John Meisenheimer; peer-reviewed by Ashley Snyder, Mei Chen; submitted 04.May.2026; final revised version received 09.Sep.2026; accepted 28.Sep.2026; published 30.Oct.2026

Please cite as:

Bal Avcı E, Dağtaş BB, İlhan Erdil D, Erdem O, Salman A, Korkmaz S, Sarıkaya Solak S

Quality and Readability of ChatGPT and Gemini Responses to Patient Questions About Systemic Isotretinoin: Comparative Evaluation Study

JMIR Dermatol 2026;9:e99780

URL: <https://derma.jmir.org/2026/1/e99780>

doi: [10.2196/99780](https://doi.org/10.2196/99780)

© Elif Bal Avcı, Buğra Burç Dağtaş, Dilara İlhan Erdil, Ozan Erdem, Andaç Salman, Selçuk Korkmaz, Sezgi Sarıkaya Solak. Originally published in *JMIR Dermatology* (<http://derma.jmir.org>), 30.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Dermatology*, is properly cited. The complete bibliographic information, a link to the original publication on <http://derma.jmir.org>, as well as this copyright and license information must be included.