ORIGINAL RESEARCH

Bagcilar Med Bull
DOI: 10.4274/BMB.galenos.2026.80774

Comparative Evaluation of Artificial Intelligence Models in Providing Patient-Oriented Information on Scabies: Implications for Public Health and Clinical Practice

Uyuz Hastalığı Hakkında Hasta Odaklı Bilgi Sağlamada Yapay Zeka Modellerinin Karşılaştırmalı Değerlendirmesi: Halk Sağlığı ve Klinik Uygulama Açısından Çıkarımlar

Orhan Şen

Kayseri City Hospital, Department of Dermatology, Kayseri, Turkey

Abstract

Objective: Scabies is an important public health problem, where patients have difficulty accessing information due to its high contagiousness and the stigma it causes. The aim of this study was to comparatively evaluate the accuracy, reliability, and readability of responses provided by current artificial intelligence (AI) chat models (ChatGPT, Google Gemini, and Microsoft Copilot) to frequently asked questions about scabies.

Method: In this cross-sectional study, conducted in January 2026, the 20 most frequently searched Turkish-language questions about scabies in Turkey were identified using Google Trends data. Responses generated by AI models were analyzed using the DISCERN instrument, the global quality score (GQS), an accuracy score (1-6 Likert scale), and the Ateşman readability formula. Word counts were also compared. Statistical analyses were performed using the Friedman test for related samples, with Wilcoxon signed-rank post-hoc tests and Bonferroni correction.

Results: Google Gemini demonstrated significantly superior performance to ChatGPT and Microsoft Copilot across all evaluated scores: DISCERN (median: 5.0), GQS (median: 5.0), and accuracy (median: 6.0) ($p<0.05$). Microsoft Copilot received the lowest accuracy scores. In a readability analysis, Google Gemini produced the most

Öz

Amaç: Uyuz, hastaların yüksek bulaşıcılığı ve buna bağlı damgalanma nedeniyle bilgiye erişimde güçlük yaşadığı önemli bir halk sağlığı sorunudur. Bu çalışmanın amacı, uyuz hastalığı ile ilgili sık sorulan sorulara verilen yanıtlarda, güncel yapay zeka (YZ) sohbet modellerinin (ChatGPT, Google Gemini ve Microsoft Copilot) doğruluğunu, güvenilirliğini ve okunabilirliğini karşılaştırmalı olarak değerlendirmektir.

Yöntem: Ocak 2026 tarihinde gerçekleştirilen bu kesitsel çalışmada, Google Trends verileri kullanılarak Türkiye genelinde uyuz ile ilgili en sık aranan 20 Türkçe soru belirlenmiştir. YZ modelleri tarafından üretilen yanıtlar; DISCERN ölçeği, global kalite ölçeği (GQS), doğruluk skoru (1-6 Likert ölçeği) ve Ateşman okunabilirlik formülü ile analiz edilmiştir. Ayrıca yanıtların kelime sayıları karşılaştırılmıştır. İstatistiksel analizler, ilişkili örneklemler için Friedman testi ve Bonferroni düzeltmeli post-hoc Wilcoxon işaretli sıralar testi kullanılarak gerçekleştirilmiştir.

Bulgular: Google Gemini; DISCERN (medyan: 5,0), GQS (medyan: 5,0) ve doğruluk (medyan: 6,0) puanlarının tamamında, ChatGPT ve Microsoft Copilot'a kıyasla istatistiksel olarak anlamlı düzeyde ($p<0,05$) daha üstün performans sergilemiştir. Microsoft Copilot, doğruluk açısından en düşük skorları almıştır. Okunabilirlik

Address for Correspondence: Orhan Şen, Kayseri City Hospital, Department of Dermatology, Kayseri, Turkey
E-mail: eru2017@gmail.com **ORCID:** orcid.org/0000-0002-8946-7330
Received: 13.01.2026 **Accepted:** 22.07.2026 **Epub:** 27.07.2026

Cite this article as: Şen O. Comparative evaluation of artificial intelligence models in providing patient-oriented information on scabies: implications for public health and clinical practice. Bagcilar Med Bull. [Epub Ahead of Print]

Abstract

comprehensible texts (58.8 points, "moderate difficulty") and the most comprehensive responses, with a mean length of 297 words.

Conclusion: Google Gemini emerged as the model that provided the most accurate, reliable, and patient-friendly information on scabies. While recommending such tools as supplementary sources of information, dermatologists should emphasize the risk of AI hallucinations and the need for physician oversight.

Keywords: Artificial intelligence, ChatGPT, Google Gemini, health literacy, scabies

Öz

analizinde Google Gemini (58,8 puan) "orta güçlükte" ve en anlaşılır metinleri üretirken, aynı zamanda ortalama 297 kelime ile en kapsamlı yanıtları veren model olmuştur.

Sonuç: Google Gemini, uyuz hastalığı hakkında en doğru, güvenilir ve hasta dostu (okunabilir) bilgiyi sağlayan model olarak öne çıkmaktadır. Dermatologlar, hastalarına ek bilgi kaynağı olarak bu modelleri önerirken, YZ'nin "halüsinasyon" riskini ve hekim kontrolünün önemini vurgulamalıdır.

Anahtar kelimeler: ChatGPT, Google Gemini, sağlık okuryazarlığı, uyuz, yapay zeka

Introduction

Scabies is an ectoparasitic infestation caused by *Sarcoptes scabiei var. hominis*; it is characterized by intense pruritus, is highly contagious, and remains a common condition worldwide. According to the World Health Organization, approximately 200 million people are estimated to be affected globally at any given time (1). The disease may lead to dermatological manifestations, sleep disturbances, secondary bacterial infections, and a substantial deterioration in quality of life (2,3).

The contagious nature of scabies, together with widespread misconceptions within society, imposes a significant psychosocial burden on affected individuals and contributes to feelings of stigma (4). Consequently, patients may hesitate to seek medical care, experience embarrassment when reporting their symptoms to physicians, and ultimately delay diagnosis and treatment. Such delays in the management of this highly contagious disease increase the risk of transmission within households and the community, thereby posing a substantial public health concern.

In recent years, rapid advances in artificial intelligence (AI)–based large language models (LLMs) have initiated a new era in health information–seeking behavior (5). Chatbots offer continuous (24/7) accessibility and provide a safe environment in which patients can share symptoms they perceive as private or embarrassing without fear of judgment (6-8). As a result, individuals with suspected scabies frequently turn to AI-powered chatbots before seeking in-person medical consultation.

Contemporary AI-based conversational agents are capable of responding to symptom-related inquiries and providing information regarding disease prevention, clinical course, and available treatment options (9).

Numerous studies have examined the roles, benefits, and limitations of AI in healthcare. Although AI chat models continuously expand the scope of information they provide through regular updates, the accuracy and reliability of their outputs remain a matter of ongoing debate (7,8).

In addition to accuracy, the comprehensibility of chatbot-generated responses for patients is a critical consideration. ChatGPT, Google Gemini, and Microsoft Copilot currently rank among the top three AI chatbots in terms of active user numbers and accessibility (10). Given the increasing utilization of these platforms, systematic evaluation of their accuracy, quality, readability, and overall comprehensibility is essential.

To date, no study has specifically assessed the quality, reliability, accuracy, and readability of information generated by LLMs regarding scabies—a highly contagious dermatological disease with significant public health implications. Considering the growing reliance of patients on AI-based information sources prior to seeking professional medical advice, there is a notable gap in the existing literature.

Therefore, the present study aimed to comparatively evaluate the accuracy, quality, reliability, and comprehensibility of responses generated by three widely used AI chatbots—ChatGPT, Google Gemini, and Microsoft Copilot—to the 20 most frequently searched patient questions related to scabies, as identified through search engine queries (11).

Materials and Methods

Study Design and Data Collection

This cross-sectional observational study was conducted on January 8, 2026.

Ethics Statement

This study analyzed only publicly search trend data from Google Trends and text outputs generated by publicly available AI chatbots. As the research did not involve any human participants, patient records, identifiable personal data, biological specimens, or animal subjects, formal ethics committee approval was not required. The study was conducted in compliance with the ethical principles of the Declaration of Helsinki (revised 2013).

Question Selection Protocol

To identify questions that best reflect public information-seeking behavior related to scabies, a structured, reproducible query-selection protocol was applied to the Google Trends database (Google LLC, Mountain View, CA, USA). The following predefined parameters were used for the search:

- Search term: "Uyuz" (the Turkish lay term for scabies)
- Geographic region: Turkey
- Time range: Last 5 years (January 2021-January 2026)
- Category filter: "Health"
- Search type: Web search
- Date of data extraction: January 8, 2026.

Within these parameters, the "Related Queries" section of Google Trends was examined, focusing on the "Top" subcategory, which lists the most frequently searched terms and question patterns associated with the main keyword. From the resulting list, queries were extracted in descending order of search frequency. The selection process was refined using the following predefined inclusion and exclusion criteria:

Inclusion Criteria

- Queries phrased as a question or directly representing a patient-oriented information need (e.g., transmission, symptoms, treatment, hygiene, prognosis).
- Queries written in Turkish.
- Queries clearly related to human scabies.

Exclusion Criteria:

- Duplicate queries or queries with overlapping semantic content (only the most frequently searched version was retained).
- Queries unrelated to the medical condition (e.g., idiomatic, slang, or metaphorical uses of the term "uyuz" in Turkish, which can also denote "annoying" or "unpleasant" in colloquial usage).
- Queries referring to veterinary scabies or non-medical contexts.
- Brand-, product-, or location-specific queries (e.g., pharmacy names, specific drug brands).

After applying these predefined inclusion criteria, the 20 highest-ranked queries were retained as the final question set. To enhance objectivity and minimize selection bias, the inclusion and exclusion criteria were defined before screening commenced, and only the search frequency ranking provided by Google Trends—not the author's subjective judgment of clinical relevance—was used to order the final list. The complete list of the 20 selected questions, grouped into thematic clinical categories, is provided in Table 1.

A total of 20 questions were selected to ensure sufficient content diversity while maintaining the feasibility of manual scoring and adequately addressing patients' primary concerns (Table 1).

AI Models and Standardized Querying Protocol

The selected 20 questions were submitted on the same day (January 8, 2026) to the three most up-to-date, publicly accessible AI models at that time:

1. ChatGPT (GPT-5.1 model, OpenAI, San Francisco, CA, USA),

2. Google Gemini (Gemini 3.0 model, Google LLC, Mountain View, CA, USA),

3. Microsoft Copilot (GPT-5-based model, Microsoft Corp., Redmond, WA, USA).

A standardized querying protocol was applied to minimize variability and ensure comparability across models. Each of the 20 questions was entered verbatim, in identical Turkish wording, into all three AI models without any modification, paraphrasing, or reordering. No system prompts, custom instructions, persona settings, or hidden contextual prompts were used. All queries were submitted through the publicly accessible, default web interface of each model, and all default parameters were retained (no manual adjustment of temperature, top-p, response length, or any other configurable model parameter, as these are not user-modifiable in the standard consumer interfaces of the evaluated platforms). To minimize potential bias arising from prior user data, conversational memory, or personalization features, browser cookies and cache were

Table 1. The 20 most frequently asked questions regarding scabies identified via Google Trends data

No	Category	Question
1	Definition	What is scabies and what causes it?
2	Transmission	How is scabies transmitted from person to person?
3	Transmission	Is being in the same environment as someone with scabies enough for transmission?
4	Symptoms	What are the first symptoms of scabies and when do they start?
5	Symptoms	Why does scabies itching become more severe at night?
6	Symptoms	In which body parts are scabies rashes more commonly seen?
7	Risk groups	Can scabies be transmitted from animals (cats, dogs) to humans?
8	Risk groups	Is scabies seen only in people with poor personal hygiene?
9	Diagnosis	How is scabies diagnosed? Is a blood test required?
10	Treatment	What is the medical treatment for scabies? Which creams or pills are used?
11	Treatment	Is it normal for itching to continue after treatment is applied?
12	Treatment	How is scabies treatment performed in pregnant women and infants?
13	Treatment	Can scabies be treated with natural home remedies (vinegar, sulfur, etc.)?
14	Prevention	If one family member has scabies, should the others also be treated?
15	Hygiene	How should clothes and bed linens be washed to be cleared of scabies?
16	Hygiene	How should non-washable items (carpets, sofas, shoes) be cleaned?
17	Contagiousness	How soon can one return to work or school after starting treatment?
18	Contagiousness	How long can the scabies mite survive outside the human body (on objects)?
19	Complications	What serious health problems does scabies cause if left untreated?
20	Immunity	Can a person who has had scabies once get it again?

Note: The 20 questions presented in this table represent the most frequently searched scabies-related queries identified through Google Trends (Turkey, Health category, January 2021-January 2026, "Top" related queries). The original Turkish queries were grouped into thematic clinical categories and translated for presentation. The standardized Turkish versions of these queries were submitted verbatim to all three AI models

cleared before each session; all queries were submitted in incognito or private browsing mode while the user was logged out of any associated account; and a new, independent chat session ("New Chat") was initiated for each individual question for each model, ensuring that no prior context could influence subsequent responses. For each question, only the first generated response was recorded and analyzed, and the "regenerate" function was not used to preserve the real-world user experience that patients would typically encounter when consulting these chatbots.

A total of 60 responses (20 questions ×3 models) were collected, saved in raw text format, anonymized, and numbered for blinded analysis.

Evaluation of AI-generated Responses

The 60 AI-generated responses were independently evaluated by two board-certified dermatologists, each with a minimum of five years clinical experience in dermatology. In cases of disagreement, a third expert opinion was sought. To reduce assessment bias, the reviewers were blinded to the identity of the AI model that generated each response.

The following assessment tools were used:

DISCERN Instrument

The DISCERN instrument is an internationally validated tool developed to assess the quality and reliability of consumer health information, with each item rated from 1 (low quality) to 5 (high quality). A final overall quality rating is provided on the same five-point scale. In the present study, each response was evaluated using this overall reliability and quality rating, and the resulting score was reported as the DISCERN score (12).

Global Quality Score (GQS)

The GQS is a five-point scale used to evaluate the overall flow, accessibility, and usefulness of the content for users. A score of 1 represents "poor quality/inadequate flow", whereas a score of 5 indicates "excellent quality" (13).

Accuracy Score

A six-point Likert scale was employed to assess concordance of the AI-generated responses with current medical literature and dermatology guidelines (1: Completely incorrect; 2: Predominantly incorrect; 3: Equal amounts of correct and incorrect information; 4: More correct than incorrect; 5: Nearly completely correct; 6: Completely correct) (14).

To ensure transparency and reproducibility of the accuracy assessment, the following predefined reference standards were used as the gold standard to judge AI-generated responses:

• The 2020 International Alliance for the Control of Scabies (IACS) Consensus Criteria for the diagnosis of scabies (1).

• The 2017 European guideline for the management of scabies, jointly issued by the European Academy of Dermatology and Venereology (EADV) and the International Union against Sexually Transmitted Infections (IUSTI) (15).

• The Turkish Society of Dermatology (*Türk Dermatoloji Derneği*) "Scabies Diagnosis and Treatment Guideline", a Delphi-based national consensus document developed by experts in dermatology, parasitology, entomology, pediatrics, pharmacology, and public health, and adopted as the principal national reference for the diagnosis and management of scabies in Turkey (16).

For each response, the following content domains were systematically evaluated against the above reference standards: (i) Etiology and causative organism; (ii) modes of transmission; (iii) clinical features and symptom characteristics; (iv) anatomical distribution of lesions; (v) differential diagnosis; (vi) first-line and alternative pharmacological treatments, with appropriate dosing and duration; (vii) management of close contacts and household members; (viii) environmental decontamination measures; (ix) post-scabietic itching and expected clinical course; and (x) appropriate referral and red-flag indicators.

A response was considered fully correct only when all clinically relevant statements were consistent with the predefined reference standards listed above and contained no factual errors, outdated information, or potentially misleading recommendations. Statements that were factually inaccurate, that contradicted current guidelines, or that had the potential to mislead patients (e.g., endorsement of unproven home remedies as definitive treatment) were rated as incorrect, with the final score reflecting the overall proportion of correct versus incorrect content in the response.

Readability and Text Analysis

Word counts of all responses were recorded. The readability of Turkish texts was evaluated using the Ateşman readability formula, which generates a readability score ranging from 0 to 100 based on average word length (number of syllables) and sentence length (number of words) (17).

The Ateşman readability score was calculated using the following formula:

Ateşman score =198.825-[40.175 × average word length (syllables)]-[2.610 × average sentence length (words)]

The resulting scores were classified according to the following reference ranges:

90-100: Very easy

70-89: Easy

50-69: Moderately difficult

30-49: Difficult

0-29: Very difficult.

Statistical Analysis

Statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS), version 26.0 (IBM Corp., Armonk, NY, USA). The normality of the data distribution was assessed using the Shapiro-Wilk test and histogram analysis. Numerical variables that did not follow a normal distribution were presented as the median and interquartile range (IQR).

Since the same 20 standardized questions were submitted to all three AI models, the resulting observations represented related (paired or repeated) measurements rather than independent samples. Therefore, differences in DISCERN, GQS, and accuracy scores among the three AI models were compared using the Friedman test, which is the appropriate non-parametric method for related samples. When statistically significant differences were identified, pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction for multiple comparisons (adjusted α=0.017 for three pairwise comparisons).

Inter-rater reliability between two independent evaluators was assessed using the intraclass correlation coefficient (ICC), and results were reported with 95% confidence intervals (CIs). A p-value of <0.05 was considered statistically significant for all analyses.

Results

Quality and Accuracy Analysis

When DISCERN scores, reflecting the quality of health information provided by the AI models, were evaluated, Google Gemini demonstrated the highest performance with

a median score of 5.0 (IQR: 4.75-5.00). ChatGPT followed with a median score of 4.0 (IQR: 4.00-4.00), while Microsoft Copilot exhibited the lowest performance, with a median score of 4.0 (IQR: 3.00-4.00) (Table 2). The Friedman test revealed a statistically significant difference among the three AI models [$\chi^2(2) = 21.58$, $p<0.001$].

Post-hoc pairwise comparisons using Wilcoxon signed-rank tests with Bonferroni correction showed that the DISCERN scores of Google Gemini were significantly higher than those of both ChatGPT (p=0.008) and Microsoft Copilot (p=0.001). No statistically significant difference was observed between ChatGPT and Copilot (p=0.098).

A similar pattern was observed for the GQS, which evaluates content flow and usefulness. Google Gemini achieved a significantly higher median GQS score of 5.0 (IQR: 4.75-5.00) than ChatGPT (median: 4.0, IQR: 4.00-4.00) and Microsoft Copilot (median: 4.0, IQR: 3.00-4.00). The Friedman test demonstrated a statistically significant difference among the models [$\chi^2(2) = 21.58$, $p<0.001$], with post-hoc Wilcoxon comparisons confirming Gemini's superiority over both ChatGPT (p=0.008) and Copilot (p=0.001).

Inter-rater reliability between the two independent reviewers was excellent, with an ICC of 0.87 (95% CI: 0.82-0.91, p<0.001).

Comparison of Accuracy Scores

When evaluated for medical accuracy, Google Gemini achieved the highest median score of 6.0 (IQR: 5.00-6.00), corresponding to responses closest to being completely correct. ChatGPT ranked second (median score 5.0, IQR: 5.00-5.25), whereas Microsoft Copilot had the lowest accuracy (median score 4.5, IQR: 4.00-5.00).

The Friedman test revealed a statistically significant difference in accuracy scores among the models [$\chi^2(2) = 22.73$, $p<0.001$]. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction indicated that all inter-model differences were statistically significant.

Gemini > ChatGPT (p=0.012)

ChatGPT > Copilot (p=0.015)

Gemini > Copilot (p=0.001).

These findings indicate that in the context of scabies-related information, Google Gemini outperformed the other models in both content quality and medical accuracy (Figure 1).

Readability and Response Length Analysis

The comprehensibility of texts generated by AI models was assessed using the Ateşman readability formula, and the level of detail was evaluated by word count.

Table 2. Comparison of reliability (DISCERN), quality (GQS), accuracy, and readability (Ateşman) scores of AI models

Variables	Model	Mean ± SD	Median (IQR)	Min-max	p-value†
DISCERN score	ChatGPT	4.10±0.55	4.00 (4.00-4.00)	3-5	**<0.001**
	Google Gemini	**4.70±0.57**	**5.00 (4.75-5.00)**	**4-5**	
	Microsoft Copilot	3.70±0.47	4.00 (3.00-4.00)	3-4	
GQS score	ChatGPT	4.10±0.55	4.00 (4.00-4.00)	3-5	**<0.001**
	Google Gemini	**4.70±0.57**	**5.00 (4.75-5.00)**	**4-5**	
	Microsoft Copilot	3.70±0.47	4.00 (3.00-4.00)	3-4	
Accuracy score	ChatGPT	5.10±0.64	5.00 (5.00-5.25)	4-6	**<0.001**
	Google Gemini	**5.60±0.68**	**6.00 (5.00-6.00)**	**5-6**	
	Microsoft Copilot	4.45±0.60	4.50 (4.00-5.00)	4-5	
Word count	ChatGPT	172.6±37.6	171.0 (146.0-190.5)	118-283	**<0.001**
	Google Gemini	**297.4±28.3**	**293.5 (279.0-308.0)**	**254-352**	
	Microsoft Copilot	195.1±38.5	192.0 (159.8-220.3)	140-262	
Readability (Ateşman score)	ChatGPT	42.0±17.6	43.2 (40.0-49.7)	9.2-69.6	**<0.001**
	Google Gemini	**58.8±4.4**	**58.0 (56.1-59.9)**	**51.5-69.5**	
	Microsoft Copilot	52.2±13.5	56.3 (48.5-60.5)	2.8-63.3	

Ateşman readability score: Scores range from 0 to 100; higher scores indicate increased readability and clarity (0-29: Very difficult; 30-49: Difficult; 50-69: Moderately difficult; 70-89: Easy; 90-100: Very easy). †: Statistical analysis: The Friedman test was used for comparisons among related samples, followed by post-hoc Wilcoxon signed-rank tests with Bonferroni correction. A p-value of <0.05 was considered statistically significant. SD: Standard deviation, IQR: Interquartile range (Q1-Q3). Boldface: Indicates the values representing the statistically best performance in each respective category.

Response Length (Word Count)

Google Gemini produced the most detailed and comprehensive responses, with a mean length of 297.4±28.3 words. This was followed by Microsoft Copilot with 195.1±38.5 words and ChatGPT with 172.6±37.6 words. The differences in response length among the models were statistically significant (p<0.001). On average, Gemini's responses were approximately 72% longer than ChatGPT's and 52% longer than Copilot's (Figure 2).

Readability

According to the Ateşman readability formula, Google Gemini (mean score: 58.8) and Microsoft Copilot (mean score: 52.2) generated texts classified as "moderately difficult", whereas ChatGPT (mean score: 42.0) produced more complex sentence structures and fell within the "difficult category" (Figure 2). Despite generating the longest responses, Google Gemini maintained the highest level of readability among the models (p<0.001).

Correlation Analysis

Correlation analysis demonstrated a positive association between response length and quality scores. As word count increased, both DISCERN scores (r_S=0.40, p<0.01) and accuracy scores (r_S=0.32, p<0.05) increased. These findings suggest that more detailed explanations are generally associated with higher medical quality and accuracy.

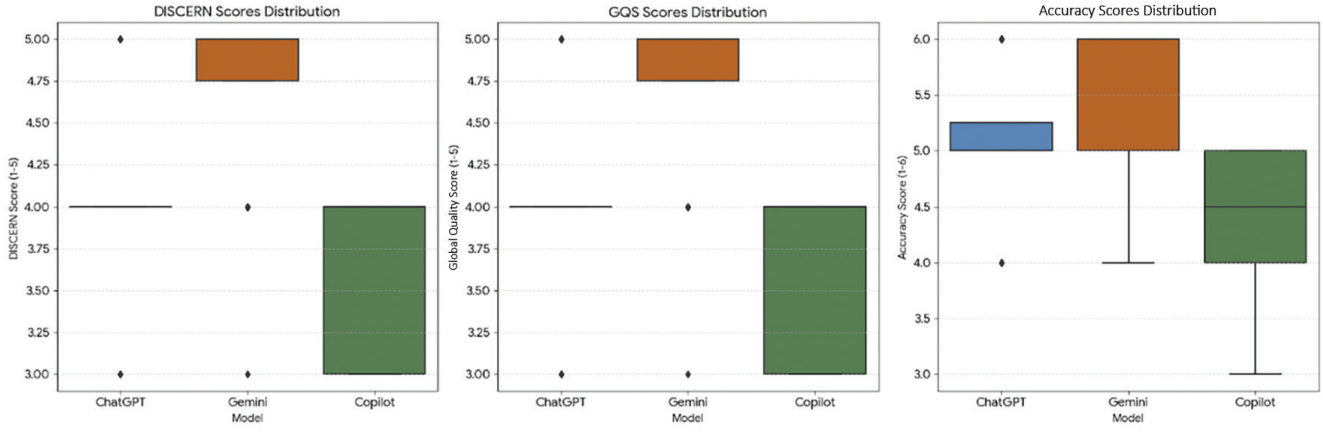

Figure 1. Distribution of DISCERN, global quality score (GQS), and accuracy scores of the artificial intelligence models. Comparison of DISCERN, GQS and accuracy scores among ChatGPT, Google Gemini, and Microsoft Copilot. Boxes represent the median and interquartile range, while whiskers indicate the minimum and maximum values. Statistical comparisons were performed using the Friedman test for related samples, with post-hoc Wilcoxon signed-rank tests and Bonferroni correction

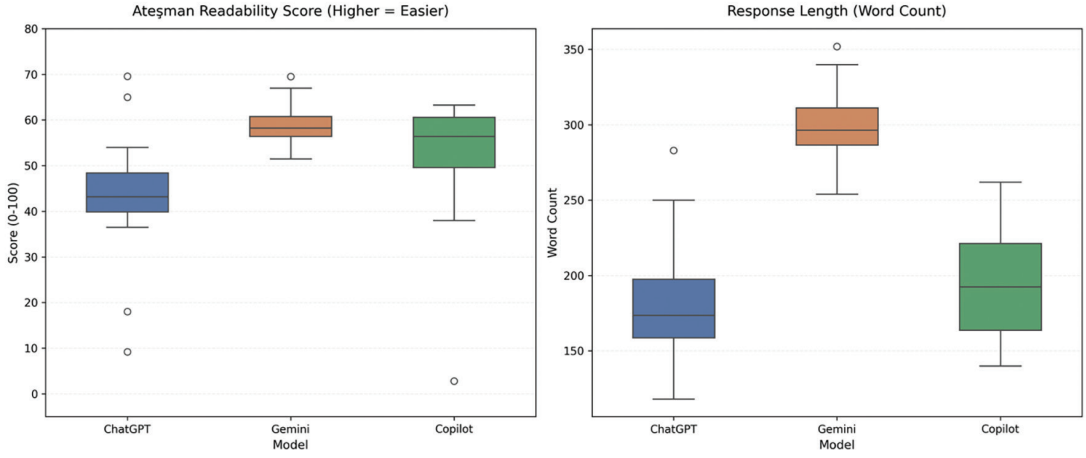

Figure 2. Comparison of Ateşman readability scores and response lengths (word count) of the artificial intelligence models. Boxes represent the median and interquartile range, while whiskers indicate the minimum and maximum values. Statistical comparisons were performed using the Friedman test for related samples, with post-hoc Wilcoxon signed-rank tests and Bonferroni correction

Discussion

To the best of our knowledge, this study is among the first to comparatively evaluate the performance of contemporary AI models in the context of scabies, a disease of critical public health importance in which patients frequently face barriers to accessing reliable information due to fear of stigma. The most striking finding of our study is that Google Gemini was statistically significantly superior to ChatGPT and Microsoft Copilot across all evaluated parameters, including DISCERN (quality), (GQS; content flow), accuracy, and readability.

Our findings indicate substantial performance differences among AI models, which may have meaningful clinical and social implications.

Comparison of DISCERN and GQS Scores

In our study, the Google Gemini model achieved significantly higher scores on both the DISCERN instrument and the GQS than ChatGPT and Microsoft Copilot. The DISCERN instrument is an internationally validated tool designed to assess the reliability of patient information materials and the balanced presentation of treatment options. Therefore, higher DISCERN scores reflect not only the quantity of information provided but also its structural integrity, neutrality, and trustworthiness (12).

Previous studies evaluating the quality of information generated by LLMs in dermatology have generally highlighted ChatGPT as a leading model. For example, Shapiro et al. (18) reported that ChatGPT achieved "moderate to good" DISCERN and GQS scores for general medical topics. Similarly, dermatology-focused studies have shown that ChatGPT provides fluent and well-structured responses, although it does not always present treatment alternatives in a balanced manner (19). In contrast, the higher DISCERN and GQS scores observed for Google Gemini in our study are noteworthy.

Several factors may explain this difference. First, Gemini's stronger integration with up-to-date and comprehensive academic data sources within the Google ecosystem may have enabled more guideline-based and systematic responses. Second, Gemini's performance in Turkish natural language processing appeared superior, particularly patient-friendly phrasing and logical flow. This advantage was reflected in higher scores on scales assessing content flow and user utility, such as the GQS and the Ateşman readability score (20-22).

These findings contribute to the literature by demonstrating that LLM performance may vary depending on both disease context and language setting.

Comparison of Accuracy Scores

Accuracy analysis is one of the most clinically critical aspects of our study. When accuracy scores were examined, Google Gemini (median: 6.0) produced responses that were nearly error-free, whereas Microsoft Copilot (median: 4.5) provided information that was occasionally incomplete or potentially misleading. Our results indicate that Google Gemini achieved the highest accuracy in responses related to scabies, whereas Microsoft Copilot demonstrated lower accuracy.

In particular, misinformation regarding critical topics such as transmission routes and treatment protocols may contribute to increased disease spread, suggesting that Copilot should be used with greater caution in this context. ChatGPT, by contrast, demonstrated a "safe middle-ground" profile—less optimal than Gemini but more reliable than Copilot. Previous studies have reported variability in the accuracy performance of LLMs (23). For instance, in the study by Zorlu et al. (24) on hidradenitis suppurativa, Microsoft Copilot achieved relatively high accuracy scores, whereas in our study it exhibited the lowest accuracy. In another investigation by Valentini et al. (25) involving sarcoma patients, ChatGPT accuracy rates ranged between 70% and 85%, although some responses were found to be clinically incomplete or misleading.

This discrepancy may be attributed to Copilot's reliance on internet-based information retrieval (e.g., Bing Search), which may direct the model toward non-scientific forum or blog sources in conditions such as scabies, where misinformation and disinformation are prevalent. In contrast, Gemini appears to produce more guideline-oriented and reliable responses, likely because of its integration with Google's extensive academic data infrastructure.

In a comprehensive review on the use of AI in medicine, Topol emphasized that misinformation in infectious diseases can have serious public health consequences (26). Within this framework, the lower accuracy scores observed for Copilot in our study warrant particular caution, especially for diseases like scabies in which incorrect information is widespread in the community.

The relatively low accuracy of Copilot observed in our study represents a significant risk for common diseases that are prone to misinformation on forums, blogs, and social media. Inaccurate or incomplete information may contribute to increased transmission and delays in appropriate treatment (27).

Taken together, our findings suggest that Gemini's more guideline-consistent and coherent responses, in contrast to Copilot's occasional provision of incomplete or misleading information, underscore that the clinical and public health impact of LLM use—particularly in infectious diseases—may be highly dependent on model selection.

Readability and Response Length Analysis

In patient education, comprehensibility is as critical as the accuracy of the information provided (28). In our study, readability was assessed using the Ateşman readability formula, and Google Gemini was found to generate responses that were both the longest and the most readable. In contrast, ChatGPT produced shorter responses with more complex sentence structures, resulting in lower readability scores.

In a study evaluating online health information, McInnes and Haglund (29) reported that most patient education materials are written above the average literacy level of the general population. In our study, Gemini's use of clearer and more structured language suggests a more patient-friendly approach that contrasts with this general trend.

Campos et al. (30), in their analysis of AI-generated responses, noted that while models may produce grammatically fluent outputs, the use of complex sentence structures can adversely affect readability. This observation aligns with our findings, in which ChatGPT generated shorter yet less readable responses.

In the study by Zorlu et al. (24) on hidradenitis suppurativa, a positive relationship was identified between response length and quality scores; however, the authors emphasized that excessively technical language may hinder patient comprehension. This finding provides an important point of comparison for our results, as Google Gemini was able to maintain higher readability despite producing longer responses.

The positive correlation observed in our study between word count and both the DISCERN and accuracy scores suggests that more detailed explanations are generally associated with higher-quality, more accurate medical information. However, this finding should be interpreted with caution and not as a simple "longer is better" rule. Response length is best understood as a proxy for explanatory depth and topical coverage rather than as a direct determinant of content quality. Increased response length has several potential drawbacks that, beyond a certain threshold, may compromise rather than enhance the value of patient-oriented information. (i) Excessive verbosity may introduce redundancy, diluting the most clinically relevant messages. (ii) Longer texts may reduce overall clarity by burying key recommendations within lengthy explanations. (iii) Higher word counts typically correlate with more complex sentence structures, which can reduce readability and impede comprehension in patients with limited health literacy. (iv) Extended responses may include peripheral or tangential content that, while not factually incorrect, distracts from the core informational need. Therefore, response length should be considered alongside, rather than as a substitute for, validated indicators of structural quality, accuracy, and readability.

In our study, the apparent advantage of longer responses depended critically on whether the additional length translated into substantively informative content delivered in patient-accessible language. Gemini's ability to preserve high readability despite producing the longest responses indicates that this model achieves a favorable balance between informational depth and comprehensibility—a balance not observed in models that produce shorter but linguistically more complex outputs (e.g., ChatGPT) or longer outputs without commensurate gains in clarity. This pattern reinforces the view that, in patient education contexts, the quality of information delivery is determined not by volume alone but by the integrated combination of accuracy, structural completeness, and readability (31).

Collectively, these findings underscore that when LLMs are used for patient education, evaluation should consider not only the amount of information provided but also how that information is presented.

Study Limitations

This study has several limitations that should be acknowledged when interpreting the findings. First, the evaluation was limited to a single dermatological condition (scabies). Although scabies is clinically and epidemiologically important and has substantial public health implications, the present findings cannot be directly extrapolated to other medical fields, dermatological disorders, or healthcare domains in which the performance of AI chatbots may differ depending on disease complexity,

the volume and quality of available training data, and the degree to which authoritative guidelines are available. Disease-specific replication studies across diverse clinical areas are therefore warranted before drawing broader conclusions about the comparative utility of AI chatbots in patient education.

Second, the numbers of analyzed questions (n=20) and AI models (n=3) were inherently limited. Although these numbers were selected to balance content diversity with the feasibility of structured manual scoring, larger question sets and the inclusion of additional AI platforms in future studies would enhance external validity.

Third, AI chatbots are subject to continuous, rapid updates—including changes to underlying model architecture, training data, fine-tuning procedures, and safety filters—that may meaningfully alter response quality, accuracy, and readability over time. Consequently, the findings reported here represent a cross-sectional snapshot of model performance as of January 8, 2026, and should not be interpreted as reflecting the long-term or future capabilities of any of the evaluated platforms. Periodic re-evaluation will be necessary to track the evolution of AI model performance in healthcare contexts.

Fourth, AI-generated outputs are inherently stochastic, and the same prompt may yield slightly different responses on repeated queries. In the present study, only the first generated response to each query was analyzed to reflect the real-world experience of a patient consulting these chatbots. While this design enhances ecological validity, it does not capture potential intra-model output variability. Future studies employing repeated queries with statistical pooling, or sensitivity analyses across paraphrased versions of the same prompt, could provide additional insight into the reproducibility of AI-generated medical information and the extent of prompt-sensitivity effects.

Fifth, the study was conducted exclusively in Turkish. Given that LLM performance—including factual accuracy, fluency, and readability—may vary substantially across languages and cultural contexts, the generalizability of the present results to other linguistic settings is limited. Cross-linguistic studies would help clarify the extent to which the observed performance differences are language-dependent.

Finally, although the accuracy assessment was anchored to predefined international and national reference standards (the IACS Consensus Criteria, the EADV/IUSTI European guideline, and the Turkish Society of Dermatology guideline), expert evaluation inevitably involves a degree of subjectivity. The high inter-rater reliability observed in this study (ICC =0.87) supports the robustness of the assessment, but fully objective, automated evaluation frameworks may further enhance reproducibility in future research.

Conclusion

This study is among the first to perform a comparative evaluation of the performance of LLMs in delivering patient-oriented information about scabies in terms of quality, accuracy, and readability. Our findings indicate that although all evaluated models generally provided acceptable levels of information, significant differences were observed in content quality, medical accuracy, and information presentation.

Google Gemini achieved the highest performance in the DISCERN and GQS scores, as well as in accuracy assessments. ChatGPT yielded moderately consistent results, whereas Microsoft Copilot exhibited lower performance, particularly in critical areas such as treatment guidance and contact management. These findings indicate that LLMs do not provide consistent quality in patient education and that model selection plays a decisive role in the reliability of the information delivered.

Analyses of readability and response length revealed that more detailed and structured responses are generally associated with higher quality and accuracy scores. However, presenting information in a clear, simplified, and patient-appropriate manner was found to be as important as the quantity of information provided.

Large language models may serve as supportive tools for patient education in common and highly contagious dermatological conditions such as scabies. Nevertheless, patients should be clearly informed that AI-generated outputs do not replace professional medical advice. The use of AI-based information resources, within the framework of physician guidance and up-to-date clinical guidelines, is essential to ensure both patient safety and public health.

Ethics

Ethics Committee Approval: Because the study is based on the analysis of publicly available data and AI outputs, and does not involve human or animal subjects, it does not require ethical committee approval.

Informed Consent: Not necessary.

Footnotes

Acknowledgements

The author would like to thank Ilgaz Genç, MD, and Merve Cansu Kaya, MD, for their contribution as independent evaluators in the blinded assessment of the AI-generated responses.

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

References

1. Engelman D, Yoshizumi J, Hay R, Osti M, Micali G, Norton S, et al. The 2020 international alliance for the control of scabies consensus criteria for the diagnosis of scabies. Br J Dermatol. 2020;183(5):808-820.
2. Karimkhani C, Colombara DV, Drucker AM, Norton SA, Hay R, Engelman D, et al. The global burden of scabies: a cross-sectional analysis from the global burden of disease study 2015. Lancet Infect Dis. 2017;17(12):1247-1254.
3. Romani L, Steer AC, Whitfeld MJ, Kaldor JM. Prevalence of scabies and impetigo worldwide: a systematic review. Lancet Infect Dis. 2015;15(8):960-967.
4. Worth C, Heukelbach J, Fengler G, Walter B, Liesenfeld O, Feldmeier H. Impaired quality of life in adults and children with scabies from an impoverished community in Brazil. Int J Dermatol. 2012;51(3):275-282.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.
6. Recker F, Neubauer R, Wittek A, Scholten N. Large language models and women's health: a digital companion for informed decision-making. Arch Gynecol Obstet. 2025;312(3):663-670.
7. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. J Med Internet Res. 2024;26:e22769.
8. Chow JC, Li K. Large language models in medical chatbots: opportunities, challenges, and the need to address AI risks. Information. 2025;16(7):549.
9. Yun HS, Bickmore T. Online health information–seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res. 2025;27:e68560.
10. Index.dev. ChatGPT Stats 2025: 800M users, traffic data & usage breakdown [Internet]. Index.dev; 2025. Available from: https://www.index.dev/blog/chatgpt-statistics
11. Google LLC. Google Trends. Mountain View, CA: Google LLC; [accessed: 08 January 2026]. Available from: https://trends.google.com
12. 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.
13. Yeung AWK. Evaluation of content quality of online health information by global quality score: a case study of researchers misnaming it and citing secondary sources. Publications. 2025;13(2):23.
14. Goodman R, Patrinely J, Stone Jr C, Zimmerman E, Donald R, Chang S, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483.
15. Salavastru CM, Chosidow O, Boffa MJ, Janier M, Tiplica GS. European guideline for the management of scabies. J Eur Acad Dermatol Venereol. 2017;31(8):1248-1253.
16. Tu̇rk Dermatoloji Derneği. Uyuz tanı ve tedavi rehberi [İnternet]. Ankara: Tu̇rk Dermatoloji Derneği. Erişim adresi: https://turkdermatoloji.org.tr/icerik/detay/401
17. Ateşman E. Tu̇rkçede okunabilirliğin ölçu̇lmesi. Dil Dergisi.1997;58:71-74.
18. Shapiro J, Avitan-Hersh E, Greenfield B, Khamaysi Z, Dodiuk-Gad RP, Valdman-Grinshpoun Y, et al. The use of a ChatGPT-4-based chatbot in teledermatology: a retrospective exploratory study. JDDG J Dtsch Dermatol Ges. 2025;23(3):311-319.
19. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the specialty certificate examination in dermatology. Clin Exp Dermatol. 2024;49(7):686-691.
20. Fattah FH, Salih AM, Salih AM, Asaad SK, Ghafour AK, Bapir R, et al. Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: a scoping review. Front Digit Health. 2025;7:1482712.
21. Arbel Y, Gimmon Y, Shmueli L. Evaluating the potential of large language models for vestibular rehabilitation education: comparison of ChatGPT, Google Gemini, and clinicians. Phys Ther. 2025;105(4):pzaf010.
22. Sami MA, Samad MA, Parekh K, Suthar PP. Comparative accuracy of ChatGPT 4.0 and Google Gemini in answering pediatric radiology text-based questions. Cureus. 2024;16(10):e70897.
23. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198.
24. Zorlu Ö, Gu̇nal U, Yazici S. The quality, accuracy, and readability of information about hidradenitis suppurativa provided by artificial intelligence: comparative analysis of ChatGPT, Copilot, and Perplexity. Medicine (Baltimore). 2025;104(39):e44728.
25. Valentini M, Szkandera J, Smolle MA, Scheipl S, Leithner A, Andreou D. Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients? Front Public Health. 2024;12:1303319.
26. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
27. Engelman D, Steer AC. Control strategies for scabies. Trop Med Infect Dis. 2018;3(3):98.
28. Weiss BD. Health literacy and patient safety: help patients understand. Manual for clinicians. American Medical Association Foundation; 2007.
29. McInnes N, Haglund BJ. Readability of online health information: implications for health literacy. Inform Health Soc Care. 2011;36(4):173-189.
30. Campos VM, Heringer DL, Costa GP, Oliva HN, Anand A. Readability, linguistic complexity, and stigma in ChatGPT responses to opioid use disorder FAQs: a comparative analysis. Am J Addict. 2026;35(3):362-369.
31. Li Y, Ouyang H, Lin G, Peng Y, Yao J, Chen Y. Evaluation of the measurement properties of online health information quality assessment tools: a systematic review. Int J Nurs Sci. 2025;12(2):130-136.