Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Artificial Intelligence Evaluate Online Health Information? A Comparative Assessment of Scabies-Related YouTube Videos Using Human Experts and ChatGPT-5

Bengisu MERAL KETENCI
meralbengisu@gmail.com

konya city hospital
ozge sevil KARSTARLI BAKAY
Pamukkale University

Article

Keywords: Scabies, Social Media, Artificial Intelligence, Natural Language Processing, Health Communication

Posted Date: March 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8924512/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Scabies is a highly contagious parasitic skin disease for which patients frequently seek information on YouTube, although the reliability of available content is uncertain. Artificial intelligence (AI), particularly large language models, has emerged as a potential tool for assessing online health information; however, its concordance with expert evaluation remains unclear. This cross-sectional study analyzed the first 50 English-language YouTube videos retrieved using the term “scabies disease.” Two dermatology specialists independently evaluated videos using the DISCERN instrument, JAMA benchmark criteria, and the Global Quality Scale (GQS). Video characteristics were recorded, and sources were classified as professional or non-professional. Corrected transcripts were analyzed with ChatGPT-5 to generate Accuracy and Completeness scores. Readability was assessed using Flesch Reading Ease and Flesch–Kincaid Grade Level. Professionally produced videos scored significantly higher than non-professional videos across all human-based quality measures ($p < 0.01$). AI-generated scores were also higher for professional content but showed only moderate correlation with expert assessments. ChatGPT Completeness demonstrated moderate discrimination in identifying higher-quality videos (AUC = 0.668). Overall, AI reflected general quality trends but did not replicate expert judgment, suggesting a complementary rather than substitutive role.

1.Introduction

Scabies is a contagious parasitic skin disease caused by *Sarcoptes scabiei* var. *hominis*, affecting hundreds of millions worldwide[1]. Despite being recognized for centuries, it remains a major public health concern, particularly in low-resource and densely populated settings[2]. The infestation typically presents with intense nocturnal pruritus and erythematous papules or burrows, often complicated by secondary bacterial infections that may lead to significant morbidity, especially in children and immunocompromised individuals[3].

Although scabies is treatable with topical or oral scabicides, difficulties such as persistent pruritus, fear of reinfestation, and treatment misuse are common. In pediatric cases, parental anxiety and misinformation further complicate management[4]. Consequently, patients and caregivers increasingly seek reliable online information to enhance disease understanding and treatment adherence.

Digital platforms have transformed public access to health information, with YouTube emerging as a widely used source for dermatologic content, including scabies[5]. Its visual format and accessibility make it particularly appealing for non-professional audiences[6]. However, the absence of content regulation allows dissemination of inaccurate or incomplete medical information, potentially leading to inappropriate treatment practices or delayed medical consultation[7]. To our knowledge, scabies-related YouTube content has not previously been evaluated using an integrated methodology combining human expert assessment and large language model–based analysis.

Given these concerns, the present study evaluates the reliability and educational quality of scabies-related YouTube videos using established human-based assessment tools—DISCERN, JAMA, and GQS—together with AI-assisted evaluation using the ChatGPT-5 model and readability analyses. Beyond content characterization, this approach specifically addresses whether artificial intelligence can meaningfully align with human expert judgment in evaluating online health information, or whether it functions primarily as a complementary screening tool rather than a human-level evaluator. By examining scabies-related videos—a common yet underexplored topic in the digital health literature—this study aims to contribute to the growing discussion on the role of large language models in assessing the quality of online medical information.

2. Materials and Methods

In this descriptive study, a YouTube search (https://www.youtube.com) was conducted on September 10, 2025, using the term “scabies disease.” To minimize algorithmic bias, the search was performed in incognito mode on a cleared browser without a logged-in account, and the term was selected based on commonly used keywords from highly viewed scabies-related videos. Results were sorted by view count to reflect typical user behavior. Because most viewers primarily engage with the highest-ranked search outputs and videos beyond the 50th position receive minimal traffic, the first 50 videos meeting the inclusion criteria were selected, replicating real-world viewing patterns[8]. To ensure reproducibility, the search was repeated on a second computer and a different internet network on the same day. Video titles and descriptions were screened independently by two dermatology specialists. Inclusion criteria required videos to be in English, at least 60 seconds in duration, publicly accessible, and to provide audio-visual medical or educational information directly related to scabies. Exclusion criteria included poor or absent audio, irrelevant content (general health topics, unrelated dermatologic diseases, or personal stories), promotional material without educational value, duplicate uploads (retaining the earliest or most-viewed version), videos under 60 seconds, and clips consisting solely of on-screen text without narration. All videos were screened using the same criteria, and categorization by source and presenter type (physician, academic institution, patient-generated, health website, news media) was conducted only after the inclusion phase during subgroup analysis. These procedures were adapted from previously published methodologies[9].

As this study involved only publicly available online content and did not include human participants, identifiable personal data, or animal subjects, approval from an ethics committee was not required.

2.1 Evaluation of Video Content

The included videos were independently evaluated by two dermatology specialists who are co-authors of this manuscript. To establish a standardized framework, prior YouTube-based medical content analyses were reviewed. Each video was classified into one of seven thematic domains: (1) definition and transmission; (2) clinical features; (3) diagnostic procedures; (4) treatment and management; (5) prevention and hygiene; (6) prognosis and follow-up; and (7) other related topics. The “other” category encompassed content on special populations, public health messages, complications, and common myths.

The interobserver reliability between the two dermatologists was evaluated using the kappa statistic. In cases of disagreement, a final decision was made through joint re-evaluation and discussion to reach a consensus.

2.2 Video Characteristics

For every video included in the analysis, the duration (in seconds), total number of views, and time elapsed since upload (in months) were recorded. The latter parameter was also converted into days on air to ensure greater temporal precision in descriptive statistics. Viewer interaction was quantified through an Engagement Index (EI), calculated using the following formula:

EI = ((likes + comments) / views) × 100.

This index provided a standardized indicator of audience engagement relative to the visibility of each video.

Videos were classified into five distinct categories based on their source of upload:

Physicians

Content created and shared by dermatologists or other medical doctors explaining diagnostic or therapeutic aspects of scabies.

Patients or Individual Users

Personal experiences or self-reported accounts uploaded by individuals affected by scabies.

Health-Related Websites or Online Platforms

Non-academic and non-governmental sources delivering general health or dermatology-related information.

Academic Institutions or Professional Organizations

Universities, teaching hospitals, and dermatology societies producing scientifically grounded educational material.

News Agencies

Informative or public health–oriented content released by national or international media outlets.

When duplicate or nearly identical videos were detected, only the version with the earliest publication date and highest view count was retained, while the others were excluded to avoid redundancy.

The presenter's gender was identified as female, male, or both, according to observable visual or auditory characteristics. If both male and female speakers appeared, the video was classified as "both genders." When gender determination was uncertain, it was omitted from the dataset.

Ultimately, the videos were grouped into two overarching categories according to their source of production, rather than their intended audience:

(1) Professional videos, produced by physicians, universities, or professional organizations; and

(2) Non-professional videos, created by patients, individual users, health-related websites, or news agencies.

2.3 Evaluation of Informational Quality and Educational Value of Videos

2.3.1 Human Reviewer Evaluation

The informational quality of the videos was assessed by two independent dermatology specialists, each with more than five years of clinical experience in dermatological practice. Prior to data collection, both reviewers underwent standardized instruction on applying the DISCERN, JAMA, and Global Quality Scale (GQS) instruments to ensure interrater consistency.

The modified DISCERN tool, derived from the original instrument developed by Charnock et al., was used to evaluate the reliability and quality of treatment-related health information[10]. It consists of five yes/no items scored as "yes" (1 point) or "no" (0 points), yielding a total score from 0 (lowest quality) to 5 (highest quality).

The JAMA benchmark criteria, originally introduced by Silberg et al., were also applied to assess the authorship, attribution, disclosure, and currency of each video[11]. Each fulfilled criterion was assigned one point, producing a total JAMA score between 0 and 4.

In addition, the 5-point Global Quality Scale (GQS) developed by Bernard et al., was used to rate the overall educational value and coherence of the videos (1–2 = low, 3 = moderate, 4–5 = high quality)[12]. Each video was reviewed independently by both raters, and discrepancies were resolved by consensus after discussion. Inter-rater reliability for DISCERN, JAMA, and GQS evaluations was measured using Cohen's κ (κ = 0.87) and intraclass correlation coefficient (ICC(2,k) = 0.90), indicating excellent agreement.

2.3.2 AI-Based Evaluation

To complement human assessment, the corrected transcripts of all included videos were analyzed using the ChatGPT-5 model (OpenAI, 2025 version). A standardized prompt was used to evaluate each video for medical accuracy and informational completeness on a 5-point Likert scale (1 = poor, 5 = excellent). The model outputs were automatically parsed and statistically compared with human reviewer scores to explore agreement levels. Transcripts were automatically generated using YouTube's closed captions feature and manually proofread by the researchers to ensure accuracy.

2.3.3 Readability Analyse

Readability metrics were calculated for each video transcript using the online tool Read-Able (https://www.webfx.com/tools/read-able/). The Flesch Reading Ease Score (FRES) and Flesch–Kincaid Grade Level (FKGL) were computed to determine linguistic accessibility and complexity of the content[13]. Higher FRES values indicate easier readability, whereas higher FKGL values correspond to more complex, higher-grade-level language.

2.4 Statistical Analysis

Analyses were performed using SPSS version 31 (IBM Corp.). Statistical significance was set at $p < 0.05$. The strength of the correlations was interpreted as strong (0.70–0.99), moderate (0.50–0.69), weak (0.01–0.49), or none (0), according to the correlation coefficient.

3.Results

A total of 92 YouTube videos related to scabies were screened, and 50 met the inclusion criteria. The selection process is shown in Fig. 1. Table 1 summarizes video characteristics, which showed wide variation in duration, time online, and engagement metrics. The mean DISCERN, GQS, and JAMA scores were 3.46 ± 1.1, 3.42 ± 1.0, and 2.23 ± 1.1. AI-based evaluations yielded mean ChatGPT Accuracy and Completeness scores of 4.62 ± 0.92 and 4.51 ± 0.85. Readability analysis showed mean FRES and FKGL values of 64.8 ± 6.3 and 9.2 ± 1.0. Treatment and management was the most frequently covered domain (86%), followed by diagnostics (46%), clinical features (42%), and definitions/transmission (38%). As shown in Fig. 2, preventive hygiene (28%), prognosis/follow-up (18%), and other topics (12%) were the least represented categories.

Table 1

Descriptive statistics of video characteristics, engagement metrics, and quality scores This table summarizes the descriptive statistics for all variables included in the study, including video characteristics (length, days on air, time since upload), viewer engagement metrics (total views, likes, comments, engagement index), human-rated quality scores (DISCERN, JAMA, GQS), and AI-based evaluation metrics (Accuracy and Completeness). Readability indices (FRES and FKGL) are also presented. For each variable, the mean, standard deviation (SD), median, minimum (Min), and maximum (Max) values are reported. These descriptive metrics provide an overview of the distribution and variability of the dataset used for subsequent correlation and multivariate regression analyses.

Video length, sec	**Mean**	**SD**	**Median**	**Min**	**Max**
	347.44	**229.04**	**302.0**	**63.0**	**896.0**
Days on air	1539.1	1340.92	955.5	69.0	4594.0
Time since upload, months	51.3	44.7	31.85	2.3	153.13
Total views, n	251428.06	359526.13	93500.0	1500.0	1674457.0
Likes, n	2247.98	4631.84	529.5	0.0	24000.0
Comments, n	278.4	485.02	68.5	0.0	2405.0
Engagement index	1.11	1.07	0.79	0.01	6.61
DISCERN	3.32	1.17	3.0	1.0	5.0
JAMA	2.14	1.11	2.0	0.0	4.0
GQS	3.28	1.26	3.0	1.0	5.0
AI ,ACCURACY	4.54	1.03	5.0	1.0	5.0
AI ,COMPLETENESS	4.53	0.8	5.0	2.0	5.0
FRES	64.45	4.46	64.6	54.8	74.1
FKGL	8.41	0.85	8.5	5.24	9.8

Video source analysis

Quality indicators differed significantly across the five video source categories (Table S1). For JAMA scores, one-way ANOVA demonstrated a significant group effect ($F = 7.93$, $p < 0.001$). Tukey's HSD indicated that patient-generated videos scored significantly lower than those from physicians ($p = 0.002$) and academic institutions ($p < 0.001$). For DISCERN, group differences were more pronounced ($F = 10.42$, $p < 0.001$), with patient-generated videos scoring significantly lower than physicians ($p < 0.001$), health-related websites ($p = 0.011$), academic institutions ($p < 0.001$), and news agencies ($p = 0.035$). GQS also differed significantly by source ($F = 10.09$, $p < 0.001$); patient-generated videos consistently scored the lowest, performing significantly worse than physicians ($p < 0.001$), health websites ($p = 0.006$), academic institutions ($p < 0.001$), and news media ($p = 0.043$).

Correlation analysis

Engagement metrics demonstrated strong intercorrelations (Table 2), including views–likes (r = 0.85, p < 0.001), views–comments (r = 0.83, p < 0.001), and likes–comments (r = 0.86, p < 0.001). Readability indices showed a strong inverse correlation (FRES–FKGL: r = − 0.79, p < 0.001). Human-based quality scores were also strongly interrelated: DISCERN–GQS (r = 0.95, p < 0.001), DISCERN–JAMA (r = 0.83, p < 0.001), and JAMA–GQS (r = 0.82, p < 0.001). AI-based metrics showed moderate correlations with human ratings, such as DISCERN–AI Accuracy (r = 0.45, p = 0.001) and GQS–AI Accuracy (r = 0.46, p = 0.001).

Table 2

Spearman correlation matrix of video characteristics, engagement metrics, quality scores, AI-based evaluation metrics, and readability indices. Spearman correl are shown. EI represents a composite index of views, likes, and comments. DISCERN, JAMA, and GQS are human-rated quality scores, while AI Accuracy and AI-based metrics. FRES and FKGL denote readability indices. **Statistical significance was defined as p < 0.05** (*p < 0.05, **p < 0.01, ***p < 0.0

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Video length (sec)	–											
2. Days on air	−0.174	–										
3. Time since upload (months)	−0.174	1.000***	–									
4. Total views	−0.188	0.738***	0.738***	–								
5. Likes	0.013	0.705***	0.705***	0.848***	–							
6. Comments	−0.026	0.655***	0.655***	0.829***	0.856***	–						
7. Engagement Index (EI)	0.471***	−0.209	−0.209	−0.306*	0.114	0.071	–					
8. DISCERN	0.198	−0.059	−0.059	0.205	0.157	0.116	−0.137	–				
9. JAMA	0.148	−0.118	−0.118	0.071	−0.008	−0.045	−0.260	0.830***	–			
10. GQS	0.215	−0.038	−0.038	0.241	0.186	0.146	−0.126	0.946***	0.818***	–		
11. AI Accuracy	−0.011	−0.386**	−0.386**	−0.277	−0.255	−0.255	−0.052	0.453***	0.414**	0.457***	–	
12. AI Completeness	0.325*	−0.424**	−0.424**	−0.362*	−0.250	−0.263	0.042	0.445***	0.464***	0.469***	0.712***	–
13. FRES	−0.428**	0.254	0.254	0.096	0.045	0.101	−0.042	−0.186	−0.230	−0.216	−0.272	−0.408*
14. FKGL	0.458***	−0.359*	−0.359*	−0.166	−0.040	−0.066	0.252	0.172	0.156	0.200	0.255	0.462**

Professional vs. non-professional comparison

When sources were categorized as professional (physicians + academic institutions) or non-professional (patients, health websites, news agencies), significant differences emerged (Table 3). Professional videos scored higher in DISCERN (4.10 ± 0.9 vs. 2.80 ± 1.0, p = 0.0001), JAMA (2.80 ± 0.95 vs. 1.70 ± 0.99, p = 0.0007), GQS (4.05 ± 1.1 vs. 2.77 ± 1.1, p = 0.0006), AI Accuracy (4.99 ± 0.04 vs. 4.23 ± 1.24, p = 0.0045), and AI Completeness (4.84 ± 0.49 vs. 4.33 ± 0.91, p = 0.012). Readability scores did not differ significantly (p > 0.05).

Table 3

Comparison of video characteristics, engagement metrics, quality scores, AI-based evaluations, and readability indices between professional and non-professional video sources

Values are presented as mean (*) values for professional (n = 20) and non-professional (n = 30) video sources. Group comparisons were performed using the Mann–Whitney U test, and corresponding p-values are indicated by (**). P-values shown in bold indicate statistically significant differences between groups ($p < 0.05$). Video characteristics include video length, days on air, and time since upload. Engagement metrics include total views, likes, comments, and the Engagement Index (EI). Human-rated quality scores include DISCERN, JAMA, and the Global Quality Scale (GQS). Artificial intelligence–based evaluation metrics include AI Accuracy and AI Completeness. Readability indices include the Flesch Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL).

Variable	Professional (Mean) (n = 20)	Non-Professional (Mean) (n = 30)	p **
Video length, sec *	393.35	316.83	0.26317
Days on air*	1483.25	1576.33	0.80449
Time since upload, months*	49.44	52.54	0.80449
Total views, n*	241372.9	258131.5	0.74383
Likes, n*	2083.2	2357.83	0.85077
Comments,n*	297.2	265.87	0.3781
Engagement index *	1.28	1.0	1.0
DISCERN *	4.1	2.8	**0.00011**
JAMA *	2.8	1.7	**0.00068**
GQS *	4.05	2.77	**0.00056**
AI ACCURACY *	4.99	4.23	**0.00453**
AI COMPLETENESS*	4.84	4.33	**0.01211**
FRES *	64.3	64.55	0.92898
FKGL *	8.41	8.42	0.67706

Agreement Between Human and AI-Based Evaluations

Agreement between human assessment scales and AI-based quality scores was evaluated using the weighted Cohen's kappa method (Table 4). In the overall sample, no statistically significant agreement was observed between ChatGPT Accuracy scores and either the GQS or DISCERN scales ($\kappa = 0.000$ for both comparisons; $p = 1.000$). In contrast, ChatGPT Completeness scores demonstrated a low-to-moderate but statistically significant agreement with human assessment scales (GQS: $\kappa = 0.281$, $p = 0.0036$; DISCERN: $\kappa = 0.257$, $p = 0.0064$). In subgroup analyses stratified by source type (physician vs non-physician), the agreement between ChatGPT Completeness scores and human assessment scales did not reach statistical significance (all $p > 0.05$) (Fig. 3A and 3B)

Table 4

Agreement Between Human and AI-Based Quality Assessments

Comparison	Overall		Professional		Non-profesional	
	κ	p	κ	p	κ	p
ChatGPT Accuracy – GQS	0,000	1,000	0,000	1,000	0,000	1,000
ChatGPT Accuracy – DISCERN	0,000	1,000	0,000	1,000	0,000	1,000
ChatGPT Completeness – GQS	**0,281**	**0,0036**	0,115	0,249	0,239	0,094
ChatGPT Completeness – DISCERN	**0,257**	**0,0064**	0,034	0,583	0,203	0,144
Agreement between human evaluation scales (DISCERN and Global Quality Scale [GQS]) and AI-based quality scores (ChatGPT Accuracy and ChatGPT Completeness) assessed using the weighted Cohen's kappa method. Results are presented for the overall sample and stratified by source type (physician vs non-physician).						
Abbreviations:						
κ, Cohen's kappa						

ROC Analysis

Receiver operating characteristic (ROC) analysis was performed to evaluate the discriminative performance of ChatGPT Completeness in identifying higher-quality videos (Table S2). ChatGPT Completeness demonstrated a statistically significant diagnostic performance with an area under the curve (AUC) of 0.668 (95% CI: 0.543–0.781; $p = 0.004$). Using a cut-off value of 5.0, the sensitivity and specificity were 84.2% and 48.4%, respectively (Fig. 4).

4.Discussion

This study provides a comprehensive evaluation of the educational quality and medical accuracy of YouTube videos on scabies, integrating human expert assessments, AI-based large language model (LLM) analyses, and readability metrics. Of 92 screened videos, 50 met the inclusion criteria. Overall, DISCERN, GQS, and JAMA scores indicated moderate-to-high informational quality, and both human and AI-based measures were comparable to previous video-based evaluations[5, 9]. Scores were significantly higher among professionally sourced videos, highlighting marked heterogeneity and the greater reliability of professional content.

The internet is now a major medium for public health education, particularly for contagious and stigmatizing dermatoses[14, 15]. Patients frequently seek diagnostic and therapeutic information online before consulting a physician, a pattern also reported for acne, psoriasis, atopic dermatitis, and skin care[16, 17]. However, this behavior exposes users to inaccurate or incomplete information[18]. This pattern has been observed not only in dermatology-related videos but also across content analyses from other medical specialties, where user-generated materials have consistently been reported to demonstrate lower accuracy and completeness compared with professionally produced content[16–21].

No significant differences were observed between professional and non-professional sources in visibility or popularity metrics, contrasting with some earlier reports. This aligns with evidence that views, likes, and comments do not reliably reflect medical content quality and with cross-platform observations showing that highly viewed dermatology videos—particularly on TikTok—are frequently produced by non-professional users and do not correspond to higher reliability[10, 22, 23]. Thus, the superiority of professional videos appears independent of simple popularity measures, underscoring the limitations of relying on engagement counts as proxies for educational value.

Professionally produced videos achieved significantly higher DISCERN, JAMA, and Global Quality Scale (GQS) scores, reflecting stronger adherence to evidence-based information and ethical transparency[21]. AI-based evaluations demonstrated a parallel trend, with professionally sourced videos showing significantly higher ChatGPT Accuracy and Completeness scores. Moderate and statistically significant positive correlations were observed between human assessment scales and AI-based quality measures, indicating that AI models are able to capture general trends in content quality. However, despite these correlations, analyses of inter-rater agreement using Cohen's kappa revealed no significant agreement for ChatGPT Accuracy scores and only low-to-moderate agreement for ChatGPT Completeness scores. These findings suggest that, while AI-based metrics may rank content quality in a direction similar to that of human experts, discrepancies—particularly in accuracy assessments—may be related to the inability of AI systems to fully represent information derived from access-restricted scientific literature and expert experience. This distinction is further supported by the ROC analysis, in which ChatGPT Completeness demonstrated a statistically significant but moderate discriminative performance for identifying higher-quality videos, characterized by high sensitivity but limited specificity. This pattern suggests that the Completeness metric may be useful for preliminary screening and sensitivity-oriented evaluations, but should not be interpreted as a standalone decision-making tool with high discriminative precision.

These findings are consistent with a recent systematic review and meta-analysis reporting substantial heterogeneity and only moderate overall accuracy of large language models in medical question-answering tasks, despite their frequent generation of guideline-aligned or partially accurate responses[24]. In this context, the "Humans in Charge" framework proposed for the integration of artificial intelligence in healthcare is further supported by our results[25, 26]. Collectively, our findings suggest that AI-derived quality metrics are better positioned as complementary and scalable tools, particularly for large-scale content screening and preliminary assessment, rather than as replacements for expert human evaluation.

Thematic analysis showed that videos predominantly addressed definitions, symptoms, and treatment, whereas preventive hygiene, contact management, and post-scabetic care were infrequently discussed. Similar underrepresentation of prevention and follow-up has been reported in online medical education and dermatology-related social media content[27–31]. In scabies, these omissions are clinically important, as guidelines stress simultaneous treatment of contacts, environmental decontamination, management of post-scabetic pruritus, and clear indications for re-evaluation[32]. Non-professional materials rarely included such structured recommendations, underscoring the need for expert-driven, guideline-aligned, and clinically comprehensive content[33].

This study has limitations. YouTube content evolves rapidly, and our cross-sectional design captures only one time point. The modest sample size (n = 50) and restriction to English-language videos limit generalizability to other settings and to populations with lower health literacy. Although DISCERN, JAMA, and GQS are frequently used in dermatology and general medical video evaluations, their dermatology-specific validity is incomplete, and they may miss domains such as cultural sensitivity or shared decision-making[19, 20, 32]. AI-based metrics were derived from a single LLM configuration and text transcripts only; non-verbal cues like demonstrations, production quality, and presenter credibility were not directly assessed. As LLM architectures and training data change, performance and alignment with experts may shift, necessitating ongoing recalibration across conditions, platforms, and languages[34]. Finally, we did not examine whether higher-quality videos translate into improved patient knowledge, behavior, or clinical outcomes.

5. Conclusions

YouTube videos on scabies exhibit substantial variability in reliability, completeness, and educational quality. Professional sources consistently outperformed non-professional ones in both human and AI-based evaluations, yet readability and coverage of prevention, contact management, and post-treatment care remain suboptimal. The lack of association between popularity and quality highlights that online visibility is a poor proxy for educational accuracy. Within this context, our findings suggest that artificial intelligence can evaluate online health information at the level of general quality trends and content coverage, but does not replicate expert judgment in assessing medical accuracy or clinical nuance. Increasing the prominence of evidence-based dermatologic content, aligning video narratives with scabies guidelines, and integrating AI-assisted—but human-supervised—quality monitoring may therefore represent a pragmatic approach to promoting more reliable, scalable, and equitable digital health communication.

Declarations

Conflicting Interest :

The authors declare no conflicts of interest.

7.Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

8.Acknowledgements

The authors thank the independent reviewers for their methodological contributions during the evaluation process.

9.Author Contributions

Bengisu Meral Ketenci conceived and designed the study, conducted data collection and statistical analyses, interpreted the results, and wrote the original manuscript draft. Özge Sevil Karstarlı Bakay contributed to data evaluation, provided methodological support, critically revised the manuscript for important intellectual content, and supervised the study. Both authors reviewed and approved the final version of the manuscript.

10.Data Availability Statement

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Publicly available video data were obtained from YouTube. Derived evaluation scores and analytical datasets are not publicly available due to analytical processing but may be shared upon request for academic purposes.

11.Additional Information

The authors declare no competing interests.

Author Contribution

Bengisu Meral Ketenci conceived and designed the study, conducted data collection and statistical analyses, interpreted the results, and wrote the original manuscript draft. Özge Sevil Karstarlı Bakay contributed to data evaluation, provided methodological support, critically revised the manuscript for important intellectual content, and supervised the study. Both authors reviewed and approved the final version of the manuscript.

Acknowledgement

The authors thank the independent reviewers for their methodological contributions during the evaluation process.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Publicly available video data were obtained from YouTube. Derived evaluation scores and analytical datasets are not publicly available due to analytical processing but may be shared upon request for academic purposes.

References

1. Sunderkötter, C., Wohlrab, J. & Hamm, H. Scabies: epidemiology, diagnosis, and treatment. *Dtsch. Arztebl Int.* **118**, 695–704 (2021).
2. Anderson, K. L. & Strowd, L. C. Epidemiology, diagnosis, and treatment of scabies in a dermatology office. *J. Am. Board. Fam Med.* **30**, 78–84 (2017).
3. Hicks, M. I., Elston, D. M. & Scabies *Dermatol. Ther.* **22**, 279–292 (2009).
4. Romani, L. et al. Prevalence of scabies and impetigo worldwide: a systematic review. *Lancet Infect. Dis.* **15**, 960–967 (2015).
5. Ozdemir Kacer, E. & Kacer, I. Evaluating the quality and reliability of YouTube videos on scabies in children: a cross-sectional study. *PLoS One*. **19**, e0310508 (2024).
6. Kang, E. et al. The quality of evidence of and engagement with video medical claims. *JAMA Netw. Open.* **9**, e2552106 (2026).
7. Alice, U. et al. Assessing the reliability of YouTube content for plastic surgery patient information in Africa with the modified DISCERN and JAMA scores. *Ann. Plast. Surg.* **94**, 403–408 (2025).
8. Advanced Web Ranking. Google organic CTR history. (2018). https://www.advancedwebranking.com/ctrstudy/
9. Gurok, N. G. et al. Scabies on YouTube: the quality, accuracy, and reliability of the videos. *Turk. J. Dermatol.* **19**, 80–86 (2025).
10. 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*. **53**, 105–111 (1999).
11. Serifler, S. & Gul, F. Evaluating tonsillectomy-related YouTube videos via a human expert review and ChatGPT-4: a multi-method quality analysis. *BMC Med. Educ.* **25**, 1157 (2025).
12. Altun, A. et al. Evaluation of YouTube videos as sources of information about complex regional pain syndrome. *Korean J. Pain*. **35**, 319–326 (2022).
13. Meyer, M. K. R. et al. Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. *Aesthetic Plast. Surg.* **49**, 1868–1873 (2025).

14. Schick, T. S. et al. Impact of digital media on the patient journey and patient–physician relationship among dermatologists and adult patients with skin diseases: qualitative interview study. *J. Med. Internet Res.* **25**, e44129 (2023).
15. Wojtara, M. Use of social media for patient education in dermatology: narrative review. *JMIR Dermatol.* **6**, e42609 (2023).
16. Reinhardt, L. et al. Quality, understandability and reliability of YouTube videos on skin cancer screening. *J. Cancer Educ.* **38**, 1667–1674 (2023).
17. Guzman, A. K. et al. Evaluation of YouTube as an educational resource for treatment options of common dermatologic conditions. *Int. J. Dermatol.* **59**, e65–e67 (2020).
18. Szeto, M. D. et al. Social media in dermatology and an overview of popular social media platforms. *Curr. Dermatol. Rep.* **10**, 97–104 (2021).
19. Karadag, A. S. et al. Social media use in dermatology in Turkey: challenges and tips for patient health. *JMIR Dermatol.* **7**, e51267 (2024).
20. Meral, H. B. et al. Evaluating the educational value and content quality of YouTube videos on myasthenia gravis. *Muscle Nerve*. **72**, 1067–1073 (2025).
21. Oydanich, M. et al. An analysis of the quality, reliability, and popularity of YouTube videos on glaucoma. *Ophthalmol. Glaucoma*. **5**, 306–312 (2022).
22. Tackett, S. et al. Medical education videos for the world: an analysis of viewing patterns for a YouTube channel. *Acad. Med.* **93**, 1150–1156 (2018).
23. Zheng, D. X., Mulligan, K. M. & Scott, J. F. TikTok and dermatology: an opportunity for public health engagement. *J. Am. Acad. Dermatol.* **85**, e25–e26 (2021).
24. Liu, M. et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J. Med. Internet Res.* **26**, e60807 (2024).
25. Jaleel, A. et al. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations: systematic review and meta-analysis. *JMIR Med. Educ.* **11**, e68070 (2025).
26. Herbert, A. S. et al. An evaluation of the readability and content-quality of pelvic organ prolapse YouTube transcripts. *Urology* **154**, 120–126 (2021).
27. Nigro, A. R. et al. Quality and readability of online educational Mohs micrographic surgery resources. *Dermatol. Surg.* **50**, 904–907 (2024).
28. White-Williams, C. et al. Addressing social determinants of health in the care of patients with heart failure: a scientific statement from the American Heart Association. *Circulation* **141**, e841–e863 (2020).
29. Aghajani, M. et al. Evaluating the quality and readability of online information about hidradenitis suppurativa: a systematic review. *Clin. Exp. Dermatol.* **50**, 1937–1944 (2025).
30. Unal, S. Demirel Ogut, N. Dermatologists' way of informative content about dermatology and cosmetology on social media. *J. Cosmet. Dermatol.* **24**, e70148 (2025).
31. Barrutia, L. et al. Benefits, drawbacks, and challenges of social media use in dermatology: a systematic review. *J. Dermatolog Treat.* **33**, 2738–2757 (2022).
32. Uzun, S. et al. Clinical practice guidelines for the diagnosis and treatment of scabies. *Int. J. Dermatol.* **63**, 1642–1656 (2024).
33. Cooper, B. R. et al. Social media as a medium for dermatologic education. *Curr. Dermatol. Rep.* **11**, 103–109 (2022).
34. Vrdoljak, J. et al. Evaluating large language and large reasoning models as decision support tools in emergency internal medicine. *Comput. Biol. Med.* **192**, 110351 (2025).

Figures

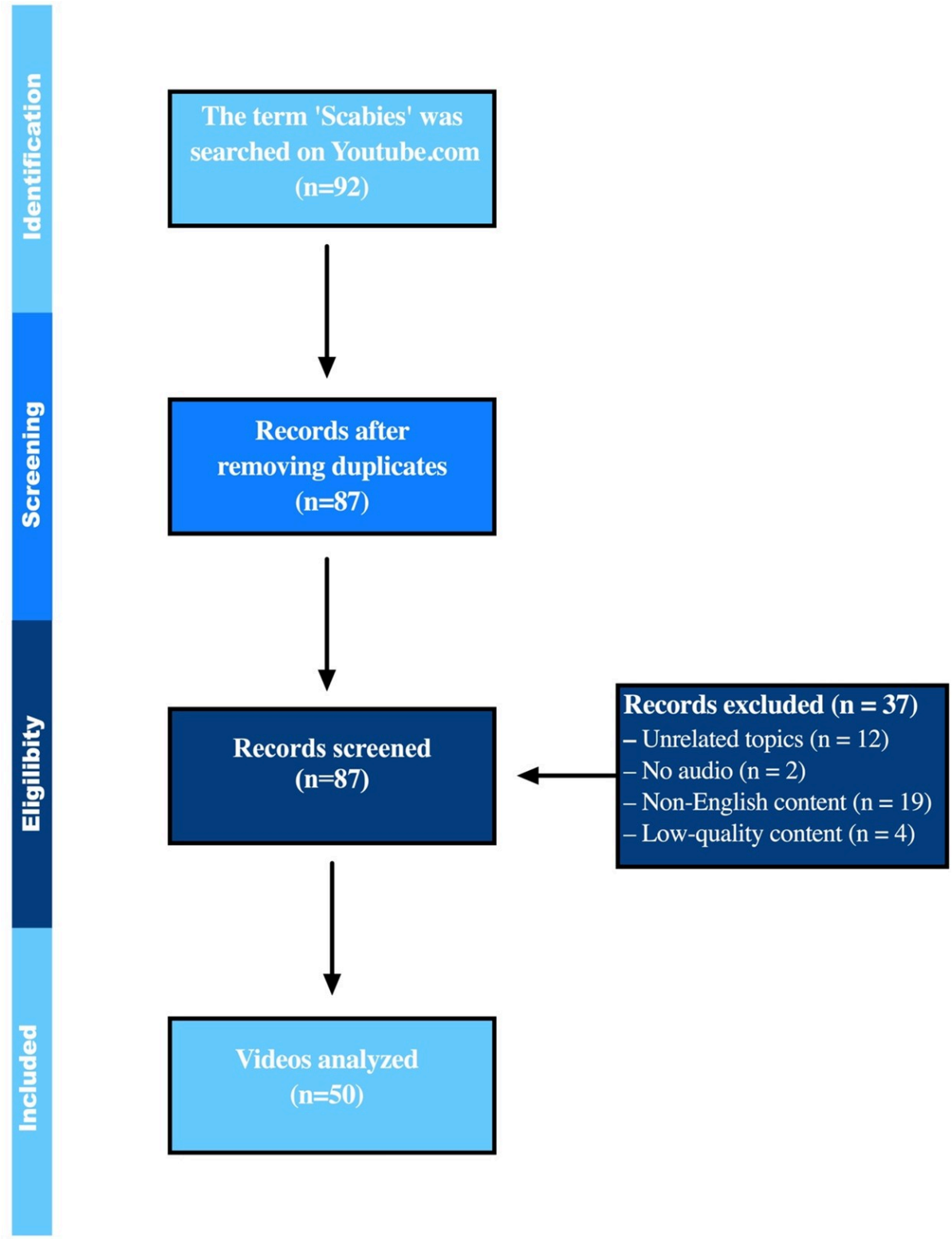


Figure 1

PRISMA-based flow diagram of video selection.

Flow diagram illustrating the identification, screening, eligibility assessment, and inclusion process of scabies-related YouTube videos. A search using the term “scabies” yielded 92 records. After removal of duplicate entries, 87 videos were screened. Thirty-seven videos were excluded due to unrelated content (n = 12), absence of audio (n = 2), non-English language (n = 19), or low informational quality (n = 4). Fifty videos met the predefined inclusion criteria and were included in the final analysis.

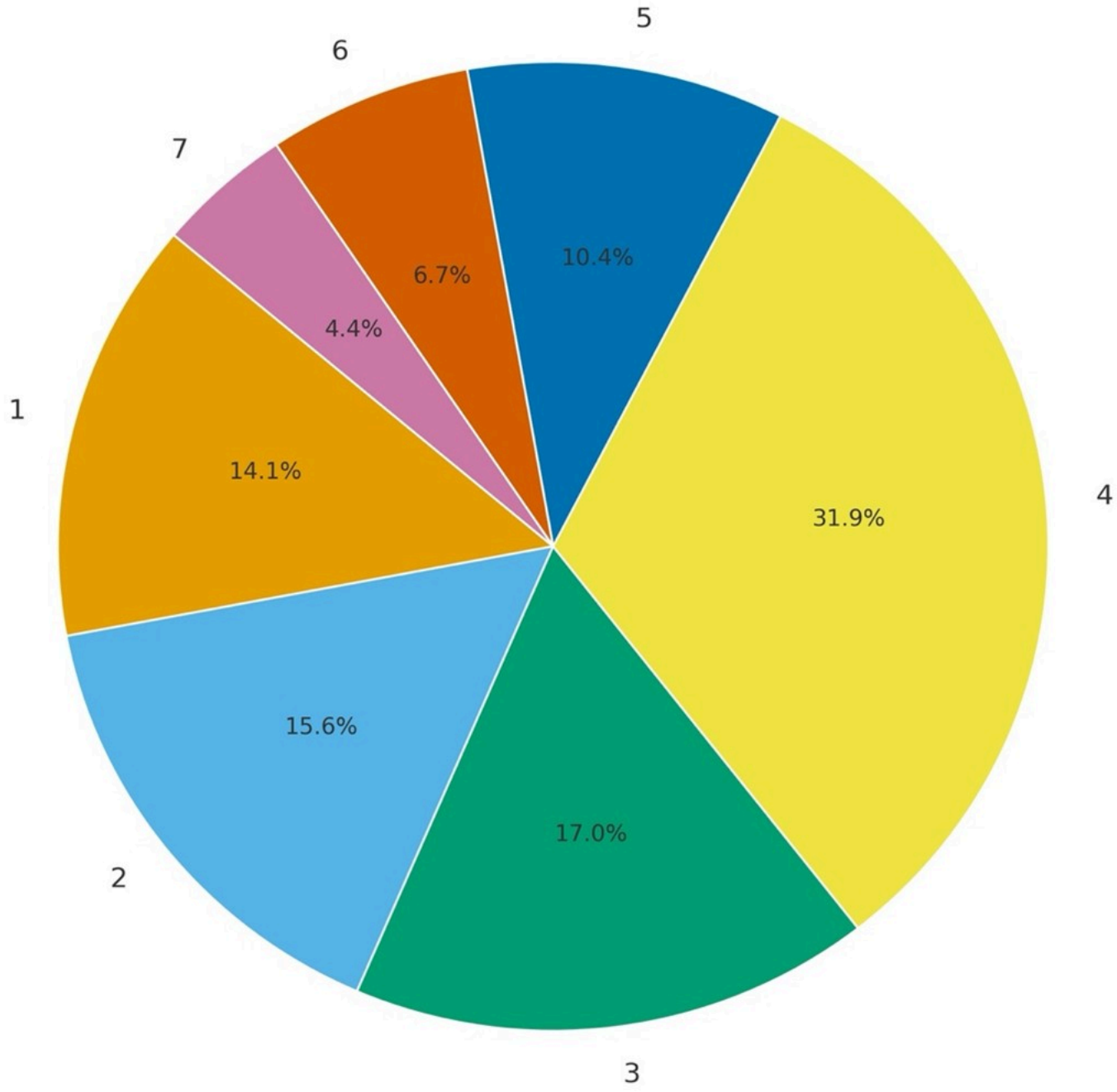


Figure 2

Distribution of scabies-related YouTube video content categories.
Pie chart showing the proportional distribution of included videos across predefined content categories. Each segment represents one of seven numerically coded content categories (1–7), and percentage values indicate the relative frequency of videos within each category among the 50 included videos.

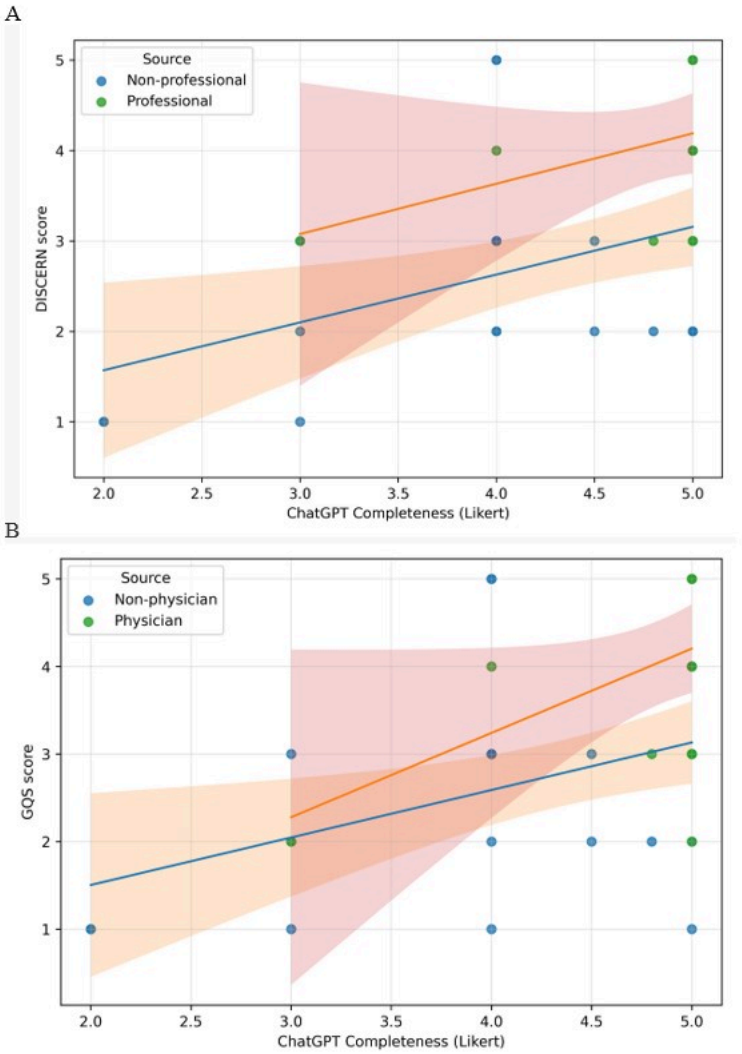


Figure 3

Association between ChatGPT Completeness and human-based quality scores by source type.
(A) Scatter plot showing the relationship between ChatGPT Completeness scores (Likert scale) and DISCERN scores, stratified by source type (physician vs non-physician). Solid lines represent linear regression models, and shaded areas indicate 95% confidence intervals.
(B) Scatter plot showing the relationship between ChatGPT Completeness scores (Likert scale) and Global Quality Scale (GQS) scores, stratified by source type (physician vs non-physician). Solid lines represent linear regression models, and shaded areas indicate 95% confidence intervals.

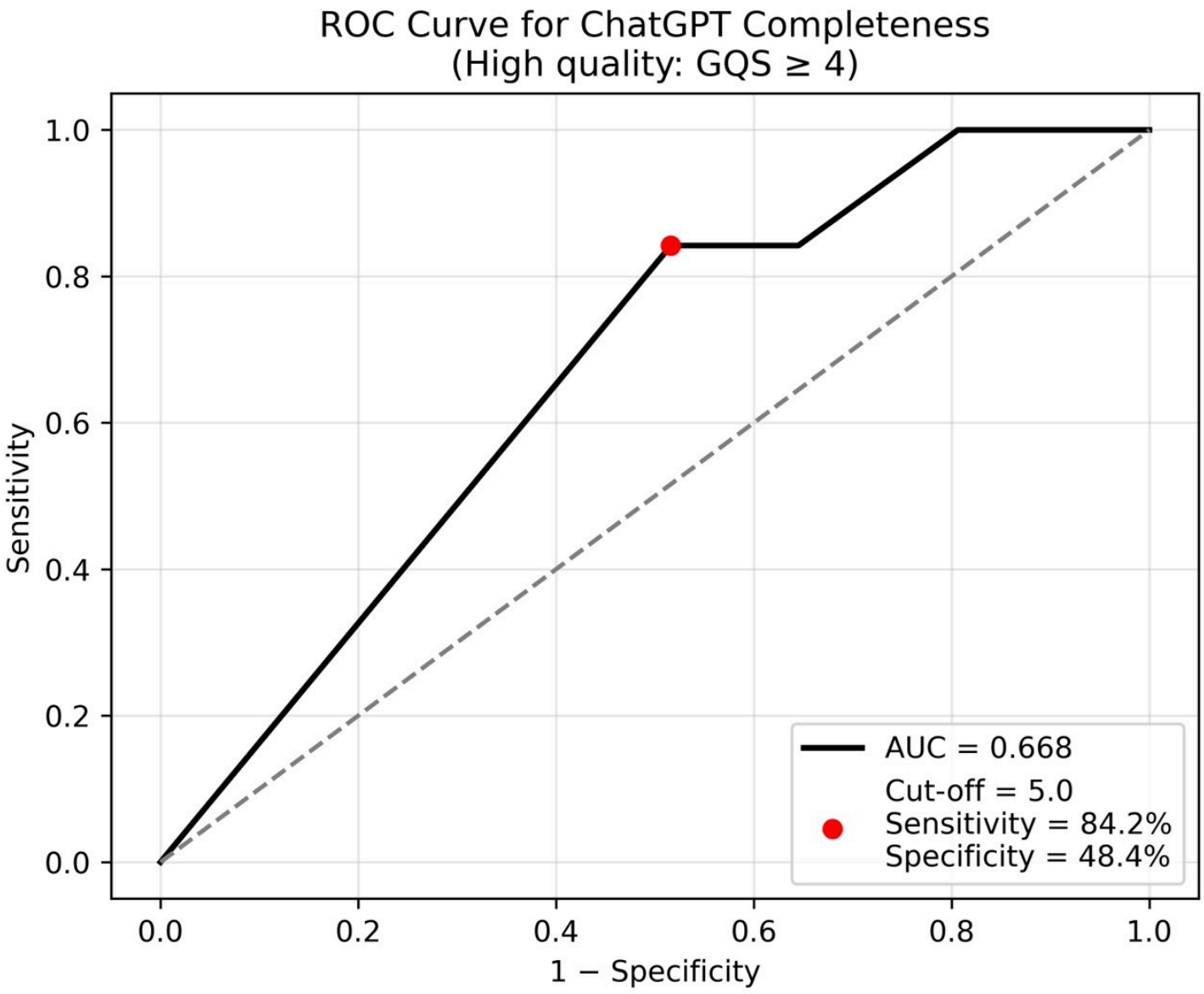


Figure 4

Receiver operating characteristic (ROC) curve of ChatGPT Completeness for identifying high-quality videos.
Receiver operating characteristic (ROC) curve evaluating the discriminatory performance of ChatGPT Completeness in identifying higher-quality scabies-related YouTube videos, defined as Global Quality Scale (GQS) ≥ 4. The area under the curve (AUC), optimal cut-off value, sensitivity, and specificity are shown.