Original Article
Özgün Araştırma

Endourol Bull. 2026;18(3):126–133. doi: 10.54233/endourolbull-1848567

Evaluating Kidney Cyst Information from Chatbots: A Comparative Study

Chatbotlardan Elde Edilen Böbrek Kisti Bilgilerinin Değerlendirilmesi: Karşılaştırmalı Bir Çalışma

Burak Koseoğlu [1*], **Eser Ördek** [2], **Kerem Gencer Kutman** [3]

[1] *Department of Urology, Polatli Duatepe State Hospital, Ankara, Türkiye*
[2] *Department of Urology, Hatay Mustafa Kemal University, Faculty of Medicine, Hatay, Türkiye*
[3] *Department of Urology, Sincan Training and Research Hospital, Ankara, Türkiye*

ABSTRACT

Objective: Kidney cysts are the most common acquired lesions of the kidney. This study aims to assess the reliability and quality of responses provided by three different artificial intelligence chatbots to questions related to kidney cysts.

Materials and Methods: Using the Semrush platform, we analyzed kidney cyst-related search trends over five years. We examined 25 questions, categorized into four subgroups: "general information", "symptoms and diagnostic methods", "treatment and follow-up procedures", and "side effects and complications". These questions were posed to three chatbots (ChatGPT-4, Gemini Pro, Llama 3.1 Large). Response readability was assessed using Flesch-Kincaid Grade Level (FKGL) test and Flesch Reading Ease Score (FRES) test. Reliability and quality were evaluated using the Ensuring Quality Information for Patients (EQIP) score and modified DISCERN score.

Results: Llama had a significantly higher FKGL score than ChatGPT ($p = 0.022$), with no significant difference between Llama and Gemini ($p = 0.500$), or Gemini and ChatGPT ($p = 0.525$). On the FRES scale, ChatGPT scored higher than Gemini, and Gemini outperformed Llama ($p = 0.006$ and $p = 0.012$, respectively). The median modified DISCERN score was lowest for ChatGPT (2), compared to Gemini (3) and Llama (3) ($p < 0.001$). The mean EQIP score was 71.22 ± 3.91 for ChatGPT, 77.72 ± 1.16 for Gemini and 73.26 ± 2.70 for Llama ($p < 0.001$). ChatGPT had significantly lower quality scores than the other two chatbots.

Conclusion: Although ChatGPT generated the most readable responses according to the FRES score, it demonstrated lower quality and reliability than Gemini and Llama. Overall, chatbot responses remained challenging in terms of readability and comprehension, highlighting the need for further improvements before they can be considered a reliable standalone source of patient education.

Keywords: artificial intelligence, chatbot, chatgpt, gemini, kidney cysts, llama

Cite As: *Koseoglu B, Ordek E, Kutman KG. Evaluating Kidney Cyst Information From Chatbots: A Comparative Study. Endourol Bull. 2026;18(3):126–133.* https://doi.org/10.54233/endourolbull-1848567

Corresponding Author : *Burak Köseoğlu, MD. Polatli Duatepe State Hospital Department of Urology 06900, Polatli, Ankara-Türkiye*
e-mail: burakkoseoglu1993@gmail.com
Received : December 24, 2025 ***Accepted*** : August 10, 2026

Copyright © 2026 Endourology Bulletin
This is an Open Access article distributed under the terms of the Attribution Non-Commercial ShareAlike 4.0 International License.

ÖZET

Amaç: Böbrek kistleri, böbreğin en sık görülen edinsel lezyonlarıdır. Çalışmanın amacı, böbrek kistleriyle ilgili sorulara üç yapay zekâ sohbet botu tarafından verilen yanıtların güvenilirliğini ve kalitesini değerlendirmektir.

Gereç ve Yöntemler: Semrush platformu kullanılarak böbrek kistleriyle ilişkili arama eğilimleri beş yıllık dönem boyunca analiz edildi. Yirmi beş soru; "genel bilgiler", "semptomlar ve tanı yöntemleri", "tedavi ve takip yaklaşımları" ile "yan etkiler ve komplikasyonlar" olarak gruplandı. Sorular üç sohbet botuna (ChatGPT-4, Gemini Pro, Llama 3.1 Large) yöneltildi. Yanıtların okunabilirliği Flesch-Kincaid Sınıf Düzeyi (FKGL) testi ve Flesch Okunabilirlik Kolaylığı Skoru (FRES) testiyle değerlendirildi. Güvenilirlik ve kalite ise Hastalar İçin Bilgi Kalitesinin Sağlanması (Ensuring Quality Information for Patients – EQIP) skoru ve modifiye DISCERN skoru kullanılarak analiz edildi.

Bulgular: Llama'nın FKGL skoru ChatGPT'ye kıyasla anlamlı derecede yüksekti (p = 0,022); Llama ile Gemini (p = 0,500) ve Gemini ile ChatGPT (p = 0,525) arasında fark saptanmadı. FRES ölçeğinde ChatGPT, Gemini'den; Gemini ise Llama'dan yüksek skor elde etti (sırasıyla p = 0,006 ve p = 0,012). Medyan modifiye DISCERN skoru ChatGPT için en düşük değerdi (2); bu değer Gemini (3) ve Llama (3) ile karşılaştırıldığında anlamlı fark gösterdi (p < 0,001). Ortalama EQIP skoru ChatGPT için 71,22 ± 3,91, Gemini için 77,72 ± 1,16 ve Llama için 73,26 ± 2,70 bulundu (p < 0,001). ChatGPT, diğer sohbet botlarına kıyasla daha düşük kalite skorlarına sahipti.

Sonuçlar: ChatGPT, FRES skoruna göre en okunabilir yanıtları oluşturmasına rağmen, kalite ve güvenilirlik açısından Gemini ve Llama'ya kıyasla daha düşük performans göstermiştir. Genel olarak sohbet botlarının yanıtları okunabilirlik ve anlaşılabilirlik açısından zorlayıcı bulunmuş olup, hasta eğitimi için güvenilir bağımsız bir kaynak olarak değerlendirilebilmeleri için daha fazla geliştirmeye ihtiyaç vardır.

Anahtar Kelimeler: böbrek kistleri, ChatGPT, Gemini, Llama, sohbet botu, yapay zekâ

INTRODUCTION

Renal cysts are the most frequently observed acquired lesions of the kidney (1). Often detected incidentally during imaging studies conducted for other indications, these lesions represent a highly prevalent urological condition and constitute a common finding in routine clinical practice (2). Although most renal cysts are benign and asymptomatic, concerns regarding their malignant potential, follow-up strategies, and treatment options frequently prompt patients to seek additional health information. Consequently, the demand for reliable patient-oriented health information remains considerable.

The internet, which became widespread in the last quarter of the 20th century, has emerged as the primary method for information retrieval today. The rapid development of information and communication technologies has substantially increased public access to health information and continuously reshaped health information-seeking behaviors (3). In recent years, artificial intelligence (AI) technologies have gained prominence as an interactive tool for accessing such information (4). AI encompasses technologies that enable computers to mimic human learning, decision-making, and creativity, and these systems are now integrated into various platforms, allowing patients immediate access to medical information and treatment options (5).

Several studies have examined the quality and reliability of AI chatbots' responses to medical inquiries (5–8). However, to the best of our knowledge, no published study has specifically evaluated chatbots in the context of renal cysts. Therefore, the aim of this study is to evaluate the readability, quality, and reliability of chatbot-generated responses pertaining to renal cysts.

MATERIALS AND METHODS

This cross-sectional observational study was initiated on December 7, 2024. The Semrush platform (Semrush Inc., Boston, MA, USA) was used to identify search trends for "kidney cyst" and related keywords. The 50 most relevant kidney cyst-related questions identified by the platform were screened according to predefined eligibility criteria,

including relevance to kidney cysts, clarity of expression, and suitability for patient-oriented questioning. Duplicate, repetitive, poorly formulated, and unrelated questions were excluded, resulting in a final set of 25 eligible questions (Appendix 1). To avoid bias prior to the study, all personal web browser accounts were logged out, and browsing histories were cleared. A 5-year timeframe was selected from the Semrush data source, and the search volumes and regional distributions were examined. Based on their content, the questions were categorized under four main headings: "general information", "symptoms and diagnostic methods", "treatment and follow-up procedures", and "side effects and complications".

In the second phase of the study, the 25 identified questions were presented once to three publicly accessible chatbot models (ChatGPT-4, Gemini Pro, and Llama 3.1 Large) on December 7, 2024, and all responses were generated and recorded on the same day. Each question was submitted only once to each chatbot without regeneration or repeated prompting, and the first response generated by the chatbot was used for analysis. The number of words, sentences, and syllables in each response was documented. Subsequently, the responses were subjected to statistical analyses focusing on readability, quality, and reliability.

The readability of the responses obtained was evaluated using the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRES) scores. The FKGL score identifies the level of education at which the text can be comfortably read. A lower FKGL score indicates a more readable and easily understandable text (9). Lower scores (0–6) correspond to an elementary school level, while higher scores (>12) indicate a university-level text. The formula for calculating the FKGL score is presented below.

FRES is another metric used to assess the readability of responses. The target score in this system is 60 or above. As the score increases, the readability of the responses also improves (10). A score between 80 and 100 indicates that the text is easily readable, while a score ranging from 0 to 30 suggests that the text is hard to understand. The calculation is shown below.

The quality and reliability of the responses were assessed using the modified DISCERN score and the Ensuring Quality Information for Patients (EQIP) tool. The modified DISCERN score was calculated by asking five questions with yes/no answers. Each "yes" or "no" response was scored as 1 point or 0 points, respectively, by two urologists. In cases of disagreement, a third reviewer was consulted to finalize the score. The questions were as follows: (a) Is the text clear and understandable? (b) Are reliable sources of information used? (c) Is the information presented both balanced and unbiased? (d) Are additional sources of information listed for patient reference? (e) Are areas of uncertainty mentioned? The EQIP tool, on the other hand, is an assessment instrument comprising 20 questions, each of which can be answered with "yes," "partly," "no," or "does not apply". As with the modified DISCERN score, the EQIP score was assigned by two urologists, with "yes" scoring 1 point, "partly" 0.5 points, and "no" 0 points. The calculation of the EQIP score is shown below. According to the EQIP score, 76%–100% indicates adequate quality, 51%–75% indicates good quality with minor issues, 26%–50% indicates serious quality issues, and 0%–25% indicates severe quality issues (11).

Statistical Analysis

Statistical analyses were performed using IBM SPSS (v 27.0.1.0). Descriptive statistics for normally distributed variables were reported as mean ± standard deviation, whereas nonparametric or ordinal data were reported as median (IQR). Normality was assessed using the Shapiro-Wilk test, and homogeneity of variances was evaluated with Levene's test. Comparisons among the three chatbot groups were performed using one-way analysis of variance (ANOVA). When significant differences were detected, Tukey's HSD (Honestly Significant Difference) test was used for post hoc comparisons if the assumption of equal variances was met. In cases of unequal variances, the Games-Howell test was applied. For ordinal variables, including Modified DISCERN scores, the Kruskal–Wallis test was used, followed by Bonferroni-adjusted pairwise comparisons when appropriate. A p value < 0.05 was considered statistically significant.

Because this study did not involve any personal data or patient information, no ethical committee approval was required. The data used in this study were obtained from publicly accessible versions of large language models (LLMs) and open-source databases.

RESULTS

The 25 questions posed to the chatbots were evaluated using readability, quality, and reliability metrics. The average number of words per response for ChatGPT, Gemini, and Llama was 130.16 ± 39.82, 228.80 ± 67.33, and 191.20 ± 71.72, respectively ($p < 0.001$). The average number of sentences per response was 12.92 ± 5.58 for ChatGPT, 17.76 ± 6.75 for Gemini, and 12.40 ± 6.04 for Llama ($p = 0.005$). Post-hoc analyses showed that ChatGPT's responses contained significantly fewer words than those of Gemini ($p < 0.001$) and Llama ($p = 0.002$). However, no statistically significant difference was observed between Gemini and Llama ($p = 0.100$). Additionally, post-hoc analyses indicated that Gemini's responses had significantly more sentences per response compared to ChatGPT and Llama ($p = 0.021$ and $p = 0.009$, respectively), whereas no significant difference was noted between ChatGPT and Llama ($p = 1.00$).

Readability was assessed using FKGL and FRES scores (Table 1). There was a statistically significant difference in FKGL scores among the three chatbots ($p = 0.026$). Pairwise comparisons revealed that Llama had a significantly higher FKGL score than ChatGPT ($p = 0.022$); however, there was no significant difference between Llama and Gemini ($p = 0.500$) or between Gemini and ChatGPT ($p = 0.525$). For the FRES scale, post-hoc analyses showed that ChatGPT had a significantly higher FRES score than Gemini, and Gemini had a significantly higher score than Llama ($p = 0.006$ and $p = 0.012$, respectively).

The reliability and quality of the responses were evaluated using the modified DISCERN score and EQIP tool (Table 1). ChatGPT had lower modified DISCERN scores compared to both Gemini and Llama ($p < 0.001$ and $p = 0.014$, respectively). No statistically significant difference was observed between Gemini and Llama ($p = 0.056$). For the EQIP score, Gemini scored significantly higher than both ChatGPT and Llama ($p < 0.001$ and $p < 0.001$, respectively), and Llama had a significantly higher score than ChatGPT ($p = 0.039$).

Analyses of the responses by question subgroup revealed no significant differences in readability measures. For the questions related to "general information", Gemini's responses had significantly higher modified DISCERN and EQIP scores. In the subgroups of "symptoms and diagnostic methods", and "treatment and follow-up procedures", Gemini had higher EQIP scores than the other chatbots, whereas the modified DISCERN scores were similar across all three. Within the "treatment and follow-up" subgroup, the modified DISCERN scores were also similar for all three chatbots; however, Gemini had a significantly higher EQIP score than the other two. Regarding "side effects and complications", all three chatbots had similar EQIP scores; however, ChatGPT's modified DISCERN score was significantly lower than that of Llama, with no statistically significant difference observed between ChatGPT and Gemini.

Table 1. The results of the readability, quality, reliability evaluation of chatbots

Variables	ChatGPT	Gemini	Llama	p value
FRES, mean ± s.d	51.69±5.45[a]	45.74±5.80[b]	40.25±8.09[c]	<0.001
FKGL, mean ± s.d	20.52±2.53[a]	21.38±1.89[ab]	22.26±2.21[b]	0.026
Modified DISCERN, median (min-max)	2 (1-3)[c]	3 (1-4)[a]	3 (1-4)[b]	<0.001
EQIP, mean ± s.d.	71.22±3.91[c]	77.72±1.16[a]	73.26±2.70[b]	<0.001

Different superscript letters indicate statistically significant differences between groups ($p < 0.05$). EQIP: The Ensuring Quality Information for Patients, FKGL: Flesch–Kincaid Grade Level, FRES: Flesch Reading Ease score, s.d: Standard Deviation

DISCUSSION

The role of AI in healthcare is expanding every day. AI-based LLMs are frequently used by patients and caregivers to obtain information on various diseases (12). Owing to their high prevalence, kidney cysts represent one of the topics most commonly queried in AI chatbots. In this study, we aimed to evaluate the readability, quality, and reliability of answers provided by three chatbots regarding kidney cysts.

To our knowledge, this is the first study in the literature to assess answers generated by AI chatbots for questions related to kidney cysts. According to our findings, ChatGPT provided more readable answers than Llama and Gemini, yet it scored lower than the other two chatbots in terms of quality and reliability. This apparent paradox may be explained by the shorter and more concise nature of ChatGPT's responses. Although brevity may improve readability, it may also reduce the comprehensiveness of the information provided, leading to lower EQIP and modified DISCERN scores. Although Gemini had higher EQIP scores than the other two chatbots, all three chatbots' EQIP scores indicated good quality with minor issues. Gemini's higher scores may be attributable to its provision of longer, more detailed answers. Subgroup analyses showed that Gemini had higher EQIP scores in all but the "side effects and complications" subgroup. These findings suggest that Gemini delivers more detailed, reliable, and higher-quality information, making it a better resource for addressing complex and challenging questions about kidney cysts.

In recent years, the use of chatbots in healthcare has attracted considerable attention, prompting numerous studies on this subject (5–7,13). Nearly all these studies have evaluated the readability and comprehensibility of chatbot responses. Texts that are difficult for patients and their caregivers to understand may lead to the spread of misinformation; therefore, the readability and comprehensibility of chatbot-generated content are of paramount importance (14). One study reported low readability in ChatGPT-generated text on spinal cord injuries (15). In that study, the readability scores, as assessed by FKGL and FRES, were 14.84 ± 1.79 and 26.24 ± 13.81, respectively. Behers and colleagues, in their assessment of responses from four different chatbots—including ChatGPT and Gemini—regarding cardiac catheterization, noted that the responses were difficult to read (10). However, the texts generated by Gemini had a lower FKGL and a higher FRES, indicating that it produced more readable content. Similarly, in an evaluation of educational brochures for urological patients, chatbot-generated answers—including those from Llama—were found to exceed the average adult reading level (16). Consistent with prior research, our findings indicate that chatbot-generated responses are often written at a level that exceeds the average reader's comprehension capacity. Among the evaluated models, ChatGPT produced the most readable content; however, its readability still corresponded to a university-level text according to the FKGL score. The relatively high FKGL values observed in the present study may be explained by the frequent use of multisyllabic medical terminology, disease classifications, and detailed explanatory language generated by the chatbots. Because the evaluated questions focused on specialized concepts such as Bosniak classification, polycystic kidney disease, and minimally invasive treatment approaches, the resulting responses naturally contained vocabulary that increased the calculated reading grade level. Post-hoc comparisons showed that ChatGPT's responses were significantly easier to read than those of Gemini, and Gemini's were in turn more readable than those of Llama. Notably, Llama's outputs aligned with the lowest FRES values and the highest FKGL scores, indicating text that would likely be challenging even for highly educated readers. These results reveal a clear gradient in readability among the chatbots and underscore the importance of optimizing AI-generated health content to meet established readability standards, especially when intended for patient education.

Various assessment tools such as the Patient Education Materials Assessment Tool (PEMAT), the Global Quality Score (GQS), the modified DISCERN score, and the EQIP score have been utilized in recent studies to evaluate the accuracy and overall quality of health-related content produced by AI-powered chatbots (5,17,18). One study that assessed content related to erectile dysfunction compared five different chatbots, including ChatGPT, using the EQIP score (5). In contrast to our findings, that study reported a significantly lower mean EQIP score for ChatGPT (40 ± 4.2), indicating serious quality issues, whereas our results showed acceptable quality with minor issues, reflected by a notably higher

mean score (71.22 ± 3.91). In an evaluation of patient education materials on female urinary incontinence produced by chatbots, the EQIP scores for the Gemini chatbot were higher than those of ChatGPT and Llama (19). These results align with the EQIP scores observed in our study. In our study, while ChatGPT received lower scores for quality and reliability than the other two chatbots, Gemini outperformed both on these metrics.

The modified DISCERN score is another commonly used tool in recent studies for evaluating the reliability of health-related information provided by AI chatbots (4,17). In their study on cardiopulmonary resuscitation content, Arca and colleagues found that Gemini provided more reliable responses than ChatGPT, as reflected by higher modified DISCERN scores, which is consistent with our results (17). In another study, health information provided by AI-powered chatbots regarding premature ejaculation was evaluated. The modified DISCERN score, used to assess reliability, was similar for Gemini and Llama, whereas ChatGPT received a lower score (4). This statistical outcome aligns with the findings of our study.

The reliability and quality of information provided by AI-supported chatbots can vary considerably between different models. Overall, the literature suggests that chatbot responses do not consistently meet adequate quality standards (5,7,20). Therefore, efforts to improve the quality of AI-generated content remain crucial. One strategy to achieve this involves expanding the training datasets used by AI chatbots (5). Incorporating specific healthcare parameters into AI machine-learning processes could facilitate the generation of more accurate responses. In addition, the evaluated chatbots differ in their accessibility and deployment characteristics. Unlike proprietary models such as ChatGPT and Gemini, Llama is an open-source large language model that can be deployed and customized locally, potentially improving accessibility in privacy-sensitive or resource-limited settings. It is also important to remember that no AI-based model can fully replace a comprehensive medical examination or the consultation of a qualified healthcare professional.

From a practical perspective, chatbot-generated information may serve as an introductory educational resource for patients and caregivers. Patients seeking brief and relatively easier-to-read information may benefit from ChatGPT's concise responses, whereas users requiring more comprehensive information may find Gemini more informative despite its lower readability. Therefore, chatbot-generated content should be interpreted as a complementary educational tool rather than a substitute for professional medical advice.

Limitations

This study has several limitations. First, the question set was restricted to the 25 most frequently searched queries identified through the Semrush platform, which may not fully represent the wide range of information needs of patients with kidney cysts. Second, the EQIP tool used to assess response quality was originally developed for evaluating written patient information rather than AI-generated chatbot content. Although it has been widely used in similar studies, dedicated assessment instruments specifically designed for LLM outputs are still needed. Third, all evaluations were performed using English-language responses, which may limit the generalizability of our findings to other languages and cultural contexts. Fourth, only specific versions of three publicly available chatbot models were evaluated at a single time point. Because LLMs are continuously updated, their responses and performance may change over time. Fifth, each question was submitted only once to each chatbot, and only the initial response was analyzed. Since chatbot outputs may vary across repeated interactions due to their probabilistic nature, future studies incorporating multiple response generations may provide a more comprehensive assessment of response quality and reliability. Finally, although the modified DISCERN and EQIP scores were independently evaluated by two urologists and disagreements were resolved through discussion with a third reviewer, formal inter-rater agreement statistics (e.g., Cohen's kappa or intraclass correlation coefficient) were not calculated, which may limit the assessment of scoring consistency.

CONCLUSION

All three chatbots evaluated in this study provided responses to kidney cyst-related questions at a readability level beyond the recommended threshold, rendering their content somewhat challenging to comprehend. Although ChatGPT produced the most readable responses, it demonstrated lower quality and reliability than Gemini and Llama. Since none of the three chatbots achieved optimal scores for quality and reliability, our findings underscore the importance of users exercising caution and seeking assistance from a healthcare professional when relying on chatbot-generated information for matters concerning health. Future research should focus on developing AI-specific scoring systems for quality assessment. Investigators should further examine advancements aimed at delivering comprehensible, high-quality information on kidney cysts to chatbot users. Meanwhile, chatbot developers should strive to enhance the clarity and reliability of the information provided.

Conflict of Interest: The authors have no relevant financial or non-financial interests to disclose.

Funding: None

Ethical Approval: No ethical approval was needed because this is not a human study, but only online information was used.

Informed Consent: No informed consent was needed because this is not a human study, but only online information was used.

Author Contributions: B Koseoglu: Project development, data collection, manuscript writing; KG Kutman: Manuscript editing, supervision; E Ordek: Data collection, data analysis

REFERENCES

1. Chang CC, Kuo JY, Chan WL, Chen KK, Chang LS. Prevalence and clinical characteristics of simple renal cyst. J Chin Med Assoc JCMA. 2007;70(11):486–91. https://doi.org/10.1016/S1726-4901(08)70046-7
2. Mensel B, Kühn JP, Kracht F, Völzke H, Lieb W, et al. Prevalence of renal cysts and association with risk factors in a general population: an MRI-based study. Abdom Radiol N Y. 2018;43(11):3068–74. https://doi.org/10.1007/s00261-018-1565-5
3. Zhao YC, Zhao M, Song S. Online Health Information Seeking Behaviors Among Older Adults: Systematic Scoping Review. J Med Internet Res. 2022;24(2):e34790.
4. Anıl H, Kayra MV. The digital dialogue on premature ejaculation: evaluating the efficacy of artificial intelligence – driven responses. Int Urol Nephrol. 2025;57:2829–36. https://doi.org/10.1007/s11255-025-04461-x
5. Şahin MF, Ateş H, Keleş A, Özcan R, Doğan Ç, et al. Responses of Five Different Artificial Intelligence Chatbots to the Top Searched Queries About Erectile Dysfunction: A Comparative Analysis. J Med Syst. 2024;48(1):38. https://doi.org/10.1007/s10916-024-02056-0
6. McMahon AK, Terry RS, Ito WE, Molina WR, Whiles BB. Battle of the bots: a comparative analysis of ChatGPT and bing AI for kidney stone – related questions. World J Urol. 2024;42(1):600. https://doi.org/10.1007/s00345-024-05326-1
7. Gül Ş, Erdemir İ, Hanci V, Aydoğmuş E, Erkoç YS. How artificial intelligence can provide information about subdural hematoma: Assessment of readability, reliability, and quality of ChatGPT, BARD, and perplexity responses. Medicine (Baltimore). 2024;103(18):e38009. https://doi.org/10.1097/MD.0000000000038009

8. Lee Y, Shin T, Tessier L, Javidan A, Jung J, et al. Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in generating clinician-level bariatric surgery recommendations. Surg Obes Relat Dis Off J Am Soc Bariatr Surg. 2024;20(7):603–8. https://doi.org/10.1016/j.soard.2024.03.011

9. Kincaid J, Fishburne R, Rogers R, Chissom B. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted personnel. Inst Simul Train [Internet]. 1975 Jan 1. Available from: https://stars.library.ucf.edu/istlibrary/56

10. Behers BJ, Vargas IA, Behers BM, Rosario MA, Wojtas CN, et al. Assessing the Readability of Patient Education Materials on Cardiac Catheterization From Artificial Intelligence Chatbots: An Observational Cross-Sectional Study. Cureus. 2024;16(7):e63865. https://doi.org/10.7759/cureus.63865

11. Ghani S, Fan KS, Fan KH, Lenti L, Raptis D. Using the Ensuring Quality Information for Patients Tool to Assess Patient Information on Appendicitis Websites: Systematic Search and Evaluation. J Med Internet Res. 2021;23(3):e22618. https://doi.org/10.2196/22618

12. Tustumi F, Andreollo NA, Aguilar-Nascimento JE de. Future of the language models in health care: the role of ChatGPT. Arq Bras Cir Dig. 2023;36:e1727. https://doi.org/10.1590/0102-672020230002e1727

13. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The Advent of Generative Language Models in Medical Education. JMIR Med Educ. 2023;9:e48163. https://doi.org/10.2196/48163

14. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review. J Med Internet Res. 2024;26:e56930. https://doi.org/10.2196/56930

15. Temel MH, Erden Y, Bağcıer F. Information Quality and Readability: ChatGPT's Responses to the Most Common Questions About Spinal Cord Injury. World Neurosurg. 2024;181:e1138–44. https://doi.org/10.1016/j.wneu.2023.11.062

16. Pompili D, Richa Y, Collins P, Richards H, Hennessey DB. Using artificial intelligence to generate medical literature for urology patients: a comparison of three different large language models. World J Urol. 2024;42(1):455. https://doi.org/10.1007/s00345-024-05146-3

17. Ömür Arça D, Erdemir İ, Kara F, Shermatova N, Odacioğlu M, et al. Assessing the readability, reliability, and quality of artificial intelligence chatbot responses to the 100 most searched queries about cardiopulmonary resuscitation: An observational study. Medicine (Baltimore). 2024;103(22):e38352. https://doi.org/10.1097/MD.0000000000038352

18. Musheyev D, Pan A, Loeb S, Kabarriti AE. How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies? Eur Urol. 2024;85(1):13–6. https://doi.org/10.1016/j.eururo.2023.07.004

19. Malak A, Şahin MF. How Useful are Current Chatbots Regarding Urology Patient Information? Comparison of the Ten Most Popular Chatbots' Responses About Female Urinary Incontinence. J Med Syst. 2024;48(1):102. https://doi.org/10.1007/s10916-024-02125-4

20. Birkun AA, Gautam A. Large Language Model (LLM) - Powered Chatbots Fail to Generate Guideline - Consistent Content on Resuscitation and May Provide Potentially Harmful Advice. Prehosp Disaster Med. 2023;38(6):757–63. https://doi.org/10.1017/S1049023X23006568