Eskisehir Med J. 2026; 7(3): 392-396
doi: 10.48176/esmj.2026.293

Original Article

COMPARATIVE EVALUATION OF CHATGPT AND GEMINI RESPONSES TO MUSCULOSKELETAL TRAUMA

FATMA HAKYEMEZ[1], EMINE CIHAN[2], CANSU SAHBAZ PIRINCCI[3]

[1] Selcuk University, Department of Medical Services and Techniques, First Aid Program, Vocational School of Health Sciences, Konya, Türkiye
[2] Selcuk University, Department of Therapy and Rehabilitation, Physiotherapy Program, Vocational School of Health Sciences, Konya, Türkiye
[3] University of Health Sciences, Gulhane Faculty of Physiotherapy and Rehabilitation, Ankara, Türkiye

ABSTRACT

Introduction: This research aimed to comparatively evaluate the performance of ChatGPT and Gemini in terms of medical accuracy and readability regarding first aid in musculoskeletal system traumas.

Methods: In the research, the responses given by both artificial intelligence systems to 10 first aid questions directed in the same format and in the Turkish language were analyzed. The responses were evaluated by 7 experts with clinical experience in the fields of first aid training and emergency health services. Flesch-Kincaid Grade Level scores were recorded for each response. Mann-Whitney U test was used in statistical analyses, and the significance level was accepted as $p<0.05$.

Results: In questions 3, 4, and 6, Gemini's response quality was found to be statistically significantly superior ($p=0.035$, $p=0.035$, and $p=0.037$). In the readability analysis, Flesch-Kincaid scores ranged between 9.85–13.46 for ChatGPT and 12.28–17.35 for Gemini. It was determined that ChatGPT responses were more readable, while Gemini responses required a higher reading level. The findings indicate an inverse relationship between the two models in terms of accuracy and readability.

Conclusion: Artificial intelligence-powered chatbots hold significant public health potential for rapid access to first aid information. However, it has been observed that more accurate responses are not always more readable. Therefore, it is thought that optimizing both the accuracy and readability of artificial intelligence-based health information will contribute to its safe and effective use by the public.

Keywords: Artificial Intelligence, Clinical Decision Support, Emergency Care, First Aid, Large Language Models, Musculoskeletal Trauma

INTRODUCTION

Musculoskeletal disorders are among the health problems that affect all age groups and can cause disability. These disorders, which can develop due to trauma or various other causes, may manifest as acute conditions such as low back pain, fractures, and sprains, as well as chronic diseases like osteoarthritis and rheumatoid arthritis. According to World Health Organization data, low back pain is reported as the leading cause of musculoskeletal disability, followed by fractures (1,2). Fractures, which have particularly high rates of emergency department visits and can be difficult to diagnose, account for approximately 80% of misdiagnosis rates due to being overlooked. Technological advancements, along with increased speed and mobility, as well as the presence of diseases affecting bone health, significantly increase the risk of trauma-related fractures (1,3). It is also emphasized that falls constitute a global public health problem and are the second leading cause of unintentional injury-related deaths. Furthermore, increased participation in sports activities is also reported to augment musculoskeletal injuries, with emergency room visits due to sports injuries being frequent in the 5-24 age range and most commonly involving injuries to the knee and ankle (4-6). Therefore, correct and timely intervention in musculoskeletal traumas will not only reduce the individual's vital risk but also contribute to the prevention of permanent functional losses. In this context, increasing first aid awareness and performing appropriate intervention by competent individuals are important as an effective strategy for preserving the quality of life of individuals.

It is reported that in recent years, individuals have been turning to AI-based applications for guidance and support in health-related situations, and this trend is expected to become widespread globally (7,8). AI, defined as computer software possessing human-like abilities such as learning, analysis, and classification, and enabling problem-solving by combining robust data sets, is emphasized to have the potential for use in many areas of healthcare, including diagnosis, treatment, care, research, and education (1,9,10). The literature reports that artificial intelligence is being used in different themes, including the diagnosis, prediction, and determination of the risk degree of diseases (11). It is stated that AI support can similarly provide positive contributions in diagnostic processes within the field of radiology and in the management of musculoskeletal traumas

Corresponding Author: Fatma Hakyemez, Department of Medical Services and Techniques, First Aid Program, Vocational School of Health Sciences, Selcuk University, Konya, Türkiye
E-mail: fatmahakyemez@selcuk.edu.tr
ORCID: 0000-0001-8390-087X

Submission Date: 06.04.2026 **Acception Date:** 23.07.2026
Cite as: Hakyemez F, Cihan E, Pirincci CS. Comparative Evaluation Of Chatgpt And Gemini Responses To Musculoskeletal Trauma. Eskisehir Med J. 2026; 7(3): 392-396. doi: 10.48176/esmj.2026.293.

(1,3). It is stated that artificial intelligence technology can offer advice on first aid and assist in guiding untrained individuals for intervention. Especially in situations where professional healthcare services may be delayed due to reasons such as geographical difficulties, it is foreseen that intervention initiated by lay rescuers, utilizing this technology, has the potential to improve outcomes in emergency situations (8,12). The literature also includes research evaluating the accuracy and reliability of artificial intelligence models in the field of first aid (12-14). However, despite the increasing number of studies evaluating artificial intelligence systems in the fields of healthcare and first aid, evidence regarding the quality and readability of responses generated by large language models, particularly for first aid interventions in musculoskeletal system traumas, remains limited. Considering that musculoskeletal injuries frequently require emergency first aid intervention and constitute a significant portion of trauma-related presentations to healthcare institutions, evaluating the reliability and readability of the information provided by artificial intelligence systems in this field is of critical importance. Especially in life-saving interventions such as first aid, the presented information must be evidence-based, and its accuracy must be evaluated by relevant experts. This study constitutes one of the few research efforts simultaneously evaluating both medical accuracy and readability; it questions not only whether the information is "accurate" but also whether it is "accessible." Therefore, this study aims to contribute to the existing literature by evaluating the responses provided by artificial intelligence models to frequently asked questions regarding first aid interventions in musculoskeletal system traumas in terms of medical accuracy and readability.

METHODS

Integration of AI-Powered Chatbots into the Research

In the research, OpenAI's ChatGPT-4 Turbo and Google's Gemini 2.0 Flash platforms were used. New user accounts were created for each platform to ensure impartiality during the data collection process and to reflect the public's natural way of posing questions. The operations were performed on a personal computer using a cleared-history browser. Both platforms were accessed via standard web interfaces; API access was not utilized, reflecting the typical usage pattern of lay users. "Temporary Chat" mode was activated in the ChatGPT system, and session and activity history were disabled on the other platforms. All questions were manually typed using the keyboard in a new conversation session each time, to prevent the chat history from influencing the responses and to avoid triggering the systems' self-learning.

Question Generation and Response Compilation Process

For the research conducted in a descriptive design, the following prompt was transmitted to both chatbots in the same format and in Turkish on May 12, 2025: “Can you list the 30 most frequently asked questions regarding first aid intervention in musculoskeletal system traumas/injuries since 2004?” The initial responses from each artificial intelligence system were recorded, and a comprehensive question pool was created. Additionally, search data for "first aid for musculoskeletal system traumas/injuries" since 2004 was analyzed via Google Trends to evaluate public interest in the subject. These data were used solely to contextualize public interest and guide the question selection process; they were not included in the statistical analysis. From the question pool consisting of 30 questions presented by both artificial intelligence systems, 10 questions were selected based on the consensus of two researchers who are experts in the field of first aid. In the selection of the questions, criteria of clinical significance, prevalence in the community, diversity of content (different scenarios such as acute trauma, chronic symptom management, transport/immobilization, open injury), and accessibility from the user's perspective were simultaneously observed. Through these criteria, questions that are not only frequently asked but also carry a critical risk of error in terms of first aid practice were prioritized (See Table 1). The identified questions were re-entered as separate prompts into the respective AI systems, and the initial responses received from each system were recorded directly.

Table 1. Research Questions

1	How should first aid be administered in case of a dislocation? Should I attempt to put it back in place?
2	What should I do to reduce swelling in joint sprains?
3	How is bleeding stopped in the area of a fracture?
4	How should someone suspected of having a neck or back trauma be handled? Why is immobilization important?
5	What are the roles of cold and hot applications in musculoskeletal system injuries, and when should each be applied?
6	How is first aid administered for muscle cramps? What might be the causes?
7	What are the potential harms of incorrect first aid applications in musculoskeletal system traumas/injuries?
8	How should first aid be administered if a fracture is suspected?
9	How is fracture stabilization done without a cast or splint?
10	What should be done in case of an open fracture?

Qualitative Evaluation of Responses

The responses were evaluated by 7 professionals with clinical experience in the field of first aid training and emergency medical services. The evaluators consist of first aid instructors, paramedics, and experts with academic and field experience in emergency medical services. All evaluators independently scored the responses using the predefined scoring criteria. To evaluate the quality of the responses, a 4-point Likert-type scoring system developed by Mika et al. (2023) was used. In the scale, 1 point represents the highest quality, and 4 points represent the lowest quality (See Table 2) (15).

Data Analysis

The readability level of the responses of the artificial intelligence (AI) applications was assessed through WordCalc software. Each response was analyzed individually through the readability assessment tool and the resulting Flesch-Kincaid Grade Level scores were recorded. To minimize potential evaluator bias during the evaluation process, the responses obtained from ChatGPT and Gemini were converted into a standardized text format by removing system names, logos, and interface information, and were presented to the evaluators in an anonymized manner. Statistical significance was determined by Mann-Whitney U test using SPSS software with a threshold of $p<0.05$.

Table 2. Method Used for Evaluating Responses

Score	Category	Description
1	Excellent response, no additional explanation needed	The response is highly accurate and comprehensive, requiring no further clarification.
2	Satisfactory, requires minimal explanation	The response is correct, but minimal additional explanation is needed to fully address the user's question.
3	Satisfactory, requires moderate explanation	The response is accurate, but a moderate level of additional explanation is needed to fully meet the user's needs.
4	Unsatisfactory, requires substantial explanation	The response contains significant inaccuracies or is overly generalized, potentially misleading the user.

RESULTS

In the 4-point Likert-type scoring system used for evaluating response quality, lower scores represent higher response quality (1 = highest quality, 4 = lowest quality). The median value for response quality was recorded as 1.00 for Gemini and 2.00 for ChatGPT in the majority of the ten questions examined for the quality of the responses provided by ChatGPT and Gemini (except for questions 1 and 9) (See Table 3). This indicates that Gemini's responses were generally assessed as being of higher quality. For questions 1 and 9, the two models were observed to be of equal quality.

Table 3. Responses of Artificial Intelligence Applications to The Questions

No	Quality of answer					
	ChatGPT		Gemini			
	Median	IQR	Median	IQR	z	p
1	2.00	1.00	2.00	2.00	-0.068	0.945
2	2.00	1.00	1.00	1.00	-1.041	0.298
3	2.00	0.00	1.00	0.00	-2.103	**0.035**
4	2.00	1.00	1.00	0.00	-2.109	**0.035**
5	2.00	1.00	1.00	1.00	-1.545	0.122
6	2.00	1.00	1.00	0.00	-2.082	**0.037**
7	2.00	1.00	1.00	1.00	-1.214	0.225
8	2.00	1.00	1.00	0.00	-1.663	0.096
9	2.00	1.00	2.00	2.00	-1.051	0.293
10	2.00	1.00	1.00	2.00	-0.431	0.667

Abb. IQR: Interquartile Range, Z: Mann Whitney U Test, p<0.05

In questions 3, 4, and 6, Gemini's median response quality (1.00) was found to be statistically significantly superior to ChatGPT's (2.00) (p<0.05= 0.035; 0.035; 0.037). In questions 1, 2, 5, 7, 8, 9, and 10, no statistically significant difference was found between the response qualities (p>0.05). However, in questions 2, 5, 7, 8, and 10, Gemini's median score (1.00) was lower (i.e., better) than ChatGPT's (2.00), but these differences did not exceed the statistical threshold. The Interquartile Range (IQR) values, especially in the questions where significant differences were observed (Questions 3, 4, 6), show that in some cases, the data dispersion around the median was quite low (IQR = 0.00), which indicates a high level of consistency among the evaluators. The findings reveal that Gemini provided statistically higher quality responses than ChatGPT in certain questions, while there was no significant difference in quality in the other questions.

The results of the Flesch-Kincaid Grade Level analysis, conducted to evaluate the readability of the responses given by both models to ten different questions, are presented in Figure 1. The findings indicate that the readability of the responses generated by Gemini is lower compared to ChatGPT. The Flesch-Kincaid scores range between 9.85 and 13.46 for ChatGPT, and between 12.28 and 17.35 for Gemini. The highest difference was recorded in Question 2, while the closest scores were recorded in Question 8.

Gemini's responses frequently reached Flesch-Kincaid Grade Levels of 15 and above, while ChatGPT's responses remained below 14. This clearly demonstrates that Gemini's responses require a post-high school or university-level reading ability, whereas ChatGPT's responses offer readability closer to the high school level.

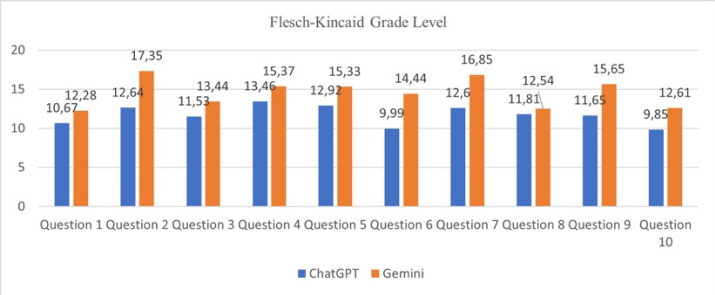


Figure 1. Flesch-Kincaid Grade Level

DISCUSSION

With advancements in digital technology, there has been an increased inclination towards online search engines and artificial intelligence-based applications for obtaining information, guidance, or support in the healthcare field, as in all other domains (7,8,16). The comprehension and effective application of information presented, especially in health content, depends on the readability and comprehensibility of the texts (17). The American Medical Association (AMA), the National Institutes of Health (NIH), and the U.S. Department of Health and Human Services have recommended that the readability level, a measure that indicates the comprehensibility of a text, should be at grade 6 or lower. Therefore, improving the readability level is seen as an important requirement for the effectiveness of AI-based health communication and patient education (18-20).

It is reported in the literature that most AI-based health materials are above the suggested grade reading levels, which may limit comprehensibility, and similar findings are observed with our results (16,17,19,20). Research reports that the combined use of chatbots such as ChatGPT, Perplexity, and Gemini is often preferred, and the easiest readability is found in Gemini, while the most difficult readability is found in Perplexity (16,19,20). In two separate studies conducted with four major language models, the responses regarding emergency care were also found to not meet the suggested readability level standard, and the readability levels of the responses generated by ChatGPT were observed to be higher (18,21). In this research, unlike the literature, the readability levels of the responses generated by

Gemini were found to be higher for all the questions examined. This divergence from the literature may be attributed to several factors. Model version updates may have shifted Gemini toward more detail-oriented outputs. Additionally, the clinical complexity of musculoskeletal trauma — involving anatomical terminology and multistep reasoning — likely elicited more terminology-dense responses compared to the general health topics examined in prior studies. Prompt structure may also have differentially influenced output complexity across platforms. These findings suggest that readability patterns observed in general health communication research may not be directly applicable to specialized clinical domains.

In addition to information being readable, its accuracy and reliability are also important. Information that is not evidence-based, incomplete, or incorrect can endanger individuals' lives and negatively affect public health (18-20). Studies evaluating the recommendations of chatbots regarding emergency care or first aid have reported that the information is sometimes misleading, may neglect critical actions, and therefore carries a risk of harming the victim (21-23). However, in studies where educational content for patient education and responses to frequently asked questions about lung cancer surgery and hip fractures, generated by ChatGPT, were evaluated by field experts, the responses were reported to be satisfactory and evidence-based, rarely requiring minimal or moderate clarification (24,25). In this study, expert evaluations demonstrated that none of the responses contained explicitly incorrect information that could potentially lead to major misunderstandings for the user. However, variations were observed among the responses in terms of completeness, clarity, and overall quality. Statistically significant differences were observed in questions 3, 4, and 6 ($p<0.05$); although no significant difference was found in questions 2, 5, 7, 8, and 10 ($p>0.05$), Gemini's median score (1.00) was observed to be superior. Overall, the findings indicate that the performance of the two models varied across individual questions, and statistically significant differences were observed only in specific scenarios. These findings contribute to the limited literature evaluating large language models within the context of first aid for musculoskeletal system traumas.

Limitations

In this study, the two artificial intelligence chatbots, ChatGPT and Gemini, were selected due to their accessibility, popularity, and ease of use; however, since snapshot responses dated May 12, 2025, were evaluated, it should be considered that different results might be obtained when the same questions are asked on different dates, and that model behaviors are continuously updated. While the limited number of analyzed questions (n = 10) and the selection of all experts from the field of first aid and emergency health services enriched the evaluator perspective, the exclusion of evaluators from different medical specialties (orthopedics, physical therapy, etc.) can be noted as a limitation. It is anticipated that future research conducted with larger question pools may increase generalizability. In the research, which has a descriptive design, multiple comparison corrections were not applied in order to preserve the power of statistical tests; analyses were conducted using raw scores and percentage changes to directly reflect clinical trends. The potential for variations in scoring due to the individual clinical experiences of the experts evaluating the artificial intelligence responses, and the fact that inter-rater reliability analysis (ICC/Kappa) was not the primary focus of our study, represent a limitation of this research. In future studies, the utilization of more standardized evaluation criteria with larger expert panels is recommended. The Flesch-Kincaid index was preferred to compare the readability of the artificial intelligence outputs with the global literature. However, since this formula was developed based on English language structure, the rich agglutinative system and high syllable rates of Turkish can result in scores that reflect a more difficult/advanced level than is actually the case. Although Turkish-specific readability formulas exist in the literature, the international index was utilized as a baseline in this study to compare the performance of global models on a standardized scale. In future studies, incorporating local readability metrics into the analysis is recommended. Furthermore, the fact that the chatbot responses were not evaluated on real patient data limits the clinical applicability of the findings. It is anticipated that future studies examining the impact of AI-generated responses on patient outcomes will provide more comprehensive insights into the functionality and reliability of these models. Additionally, generating the initial question pool using artificial intelligence systems may have introduced a potential risk of selection bias, even though the final questions were selected through expert evaluation.

CONCLUSION

This study reveals that artificial intelligence-powered chatbots generally produce acceptable responses to first aid questions concerning musculoskeletal system traumas. Gemini generated higher quality responses compared to ChatGPT in some of the evaluated scenarios; however, this difference does not apply to all questions, and the current findings are limited to the ten evaluated scenarios and cannot be generalized to all musculoskeletal trauma conditions. It is recommended that such technological tools should not replace professional healthcare services, but should only be utilized as a supportive source of information.

Ethics Committee Approval: This study does not involve human participants or biological data. Therefore, ethical approval was not required. The study was conducted in accordance with general principles of scientific research and publication ethics.

Financial Disclosure: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest: The authors declare no conflict of interest.

Informed Consent: Informed consent was not required, as no human participants or identifiable personal data were included in this study.

Authorship Contributions: FH: The authors confirm contribution to the manuscript as follows: conceptualization, methodology, supervision, writing – reviewing and editing; EC: Visualization, formal analysis; CSP: Data curation, investigation, writing – original draft preparation. All authors contributed to the analysis of results, reviewed this manuscript and revised it critically for important intellectual content, approve of version as submitted, and agree to be accountable for all aspects of the work.

REFERENCES

1. Ajmera P, Kharat A, Botchu R, Gupta H, Kulkarni V. Real-world analysis of artificial intelligence in musculoskeletal trauma. J Clin Orthop Trauma. 2021;22:101573. PMID: 34527511. DOI: https://doi.org/10.1016/j.jcot.2021.101573.
2. World Health Organization. Falls; Disability and health [Internet]. Geneva: World Health Organization; 2021 Apr 26 [cited 2025 Oct 22]. Available at: https://www.who.int/news-room/fact-sheets/detail/falls; https://www.who.int/news-room/fact-sheets/detail/disability-and-health.
3. Zech JR, Santomartino SM, Yi PH. Artificial intelligence (AI) for fracture diagnosis: An overview of current products and considerations for clinical adoption, from the AJR special series on AI applications. AJR Am J Roentgenol. 2022;219(6):869-78. PMID: 35731103. DOI: https://doi.org/10.2214/AJR.22.27873.
4. Örsçelik A, Yıldız Y. Kas-iskelet sistemi yaralanmaları: aşırı kullanım yaralanmaları. Turkiye Klinikleri J Sports Med-Special Topics. 2015;1(3):62-69.
5. Ercan S, Çetin C. Kas-iskelet sistemi yaralanmaları: alt ekstremite akut yaralanmaları. Turkiye Klinikleri J Sports Med-Special Topics. 2015;1(3):82-88.
6. Uysal, M. (ed.) (2023). Spora Çok Yönlü Yaklaşım. Özgür Yayınları. DOI: https://doi.org/10.58830/ozgur.pub329.
7. Baldwin AJ. Readability, accountability, and quality of burns first aid information available online. Burns. 2023;49(8):1823-1832. PMID: 37821277. DOI: https://doi.org/10.1016/j.burns.2023.03.002.
8. Bushuven S, Bentele M, Bentele S, et al. ChatGPT, Can You Help Me Save My Child's Life?" Diagnostic Accuracy and Supportive Capabilities to Lay Rescuers by ChatGPT in Prehospital Basic Life Support and Paediatric Advanced Life Support Cases – An In-silico Analysis. J Med Syst. 2023;21;47(1):123. PMID: 37987870. DOI: https://doi.org/10.1007/s10916-023-02019-x.
9. Akalın B, Veranyurt Ü. Sağlık 4.0 ve sağlıkta yapay zekâ. J Health Prof Res. 2022;4(1):57-64.
10. Yılmaz Y, Yılmaz DU, Yıldırım D, Akın E, Kaya DÖ. Artificial Intelligence and The Use of Artificial Intelligence in Health: Opinions of Health Sciences Students Health Sciences students. Suleyman Demirel Univ J Health Sci. 2021;12(3):297-308. DOI: https://doi.org/10.22312/sdusbed.950372.
11. Hoşgör H, Güngördü H. A qualitative research on the artificial intelligence in health. EJOSAT. 2022;(35):395-407. DOI: https://doi.org/10.31590/ejosat.1052614.
12. Birkun AA, Gautam A. Instructional support on first aid in choking by an artificial intelligence-powered chatbot. Am J Emerg Med. 2023;70:200-202. PMID: 37330383. DOI: https://doi.org/10.1016/j.ajem.2023.06.010.
13. Birkun, AA, Gautam A. Large language model–based chatbot as a source of advice on first aid in heart attack. Curr Probl Cardiol. 2024;49(1):102048. PMID: 37640177. DOI: https://doi.org/10.1016/j.cpcardiol.2023.102048.
14. Baldwin AJ. An artificial intelligence language model improves readability of burns first aid information. Burns. 2024;50(5):1122-1127. PMID: 38492982. DOI: https://doi.org/10.1016/j.burns.2024.03.005.
15. Mika AP, Martin JR, Engstrom SM, et al. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. JBone Joint Surg. 2023;105(19):1519–1526. DOI: 10.2106/JBJS.23.00209.
16. Özduran E, Hancı V, Erkin Y, Özbek İC, Abdulkerimov V. Assessing the readability, quality and reliability of responses produced by ChatGPT, Gemini, and Perplexity regarding most frequently asked keywords about low back pain. PeerJ. 2025;13:e18847. DOI: https://doi.org/10.7717/peerj.18847.
17. Sürme Y, Topan H, Maraş Baydoğan G. Howw succesful is AI in developing postsurgical wound care education matrerial? Wound Repair Regen. 2025;33(3):e70041. DOI: https://doi.org/10.1111/wrr.70041.
18. Ömür Arça D, Erdemir İ, Kara F, 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. DOI: http://dx.doi.org/10.1097/MD.0000000000038352.
19. Kara M, Özduran E, Kara MM, Özbek İC, Hancı V. Evaluating the readability, quality, and reliability of responses generated by ChatGPT, Gemini, and Perplexity on the most commonly asked questions about Ankylosing spondylitis. PLoS One. 2025;20(6):e0326351. DOI: https://doi.org/10.1371/journal.pone.0326351.
20. Özduran E, Akkoc I, Büyükçoban S, Erkin Y, Hanci V. Readability, reliability and quality of responses generated by ChatGPT, gemini, and perplexity for the most frequently asked questions about pain. Medicine (Baltimore). 2025;104(11):e41780. DOI: 10.1097/MD.0000000000041780.
21. Yau JY, Saadat S, Hsu E, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: Experimental comparative study. J Med Internet Res. 2024;4;26:e60291. PMID: 39496149. DOI:10.2196/60291.
22. 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):757-763. PMID: 37927093. DOI: 10.1017/S1049023X23006568.
23. Birkun AA, Gautam A. Dr. google's advice on first aid: Evaluation of the search engine's question-answering system responses to queries seeking help in health emergencies. Prehosp Disaster Med. 2023;38(3):345-351. PMID: 37165838. DOI: 10.1017/S1049023X23000511.
24. Troian M, Lovadina S, Ravasin A, et al. An assessment of ChatGPT's responses to common patient questions about lung cancer surgery: A preliminary clinical evaluation of accuracy and relevance. J Clin Med. 2025;14(5):1676. PMID: 40095693. DOI: https://doi.org/10.3390/jcm14051676.
25. Wrenn SP, Mika AP, Ponce RB, Mitchell PM. Evaluating ChatGPT's ability to answer common patient questions regarding hip fracture. J Am Acad Orthop Surg. 2024;32(14):656-659. PMID: 38743959. DOI: 10.5435/JAAOS-D-23-00877.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.