

The Responses of Artificial Intelligence to Questions About Urological Emergencies: A Comparison of 3 Different Large Language Models

Ubeyd Sungur¹, Yusuf Arıkan², Ahmet Tuğrul Türkay¹, Hakan Polat¹

¹ Department of Urology, University of Health Sciences, Bakirkoy Dr. Sadi Konuk Training and Research Hospital, Istanbul, Türkiye

² Department of Urology, University of Health Sciences, Tepecik Training and Research Hospital, Izmir, Türkiye

Submitted: 2025-02-22

Accepted: 2025-05-29

Corresponding Author;

Ubeyd Sungur, MD

Address: University of Health Sciences Bakirkoy Dr. Sadi Konuk Training and Research Hospital, Department of Urology, Zuhuratbaba Mah, Dr. Tevfik Sağlam Cd No:11, 34147 Bakırköy/İstanbul, Türkiye

E-mail: ubeydsungur@gmail.com

ORCID

U.S. [0000-0002-8910-9859](https://orcid.org/0000-0002-8910-9859)

Y.A. [0000-0003-0823-7400](https://orcid.org/0000-0003-0823-7400)

A.T.T. [0009-0005-4210-8954](https://orcid.org/0009-0005-4210-8954)

H.P. [0000-0003-1525-1243](https://orcid.org/0000-0003-1525-1243)

Abstract

Objective: This study aimed to compare the accuracy and adequacy of responses provided by three different large language models (LLMs) utilizing artificial intelligence technology to fundamental questions related to urological emergencies.

Material and Methods: Nine distinct urological emergency topics were identified, and a total of 63 fundamental questions were formulated for each topic, including two related to diagnosis, three related to disease management, and two related to complications. The questions were posed in English on three different free AI platforms (ChatGPT-4, Google Gemini 2.0 Flash, and Meta Llama 3.2), each utilizing different infrastructures, and responses were documented. The answers were scored by the authors on a scale of 1 to 4 based on accuracy and adequacy, and the results were compared using statistical analysis.

Results: When all question-answer pairs were evaluated overall, ChatGPT exhibited slightly higher accuracy rates compared to Gemini and Meta Llama; however, no statistically significant differences were detected among the groups (3.8 ± 0.5 , 3.7 ± 0.6 , and 3.7 ± 0.5 , respectively; $p=0.146$). When questions related to diagnosis, treatment management, and complications were evaluated separately, no statistically significant differences were detected among the three LLMs ($p=0.338$, $p=0.289$, and $p=0.407$, respectively). Only one response provided by Gemini was found to be completely incorrect (1.6%). No misleading or wrong answers were observed in the diagnosis-related questions across all three platforms. In total, misleading answers were observed in 2 questions (3.2%) for ChatGPT, three questions (4.7%) for Gemini, and two questions (3.2%) for Meta Llama.

Conclusion: LLMs predominantly provide accurate results to basic and straightforward questions related to urological emergencies, where prompt treatment is critical. Although no significant differences were observed among the responses of the three LLMs compared in this study, the presence of misleading and incorrect answers should be carefully considered, given the evolving nature and limitations of this technology.

Keywords: urological emergencies, artificial intelligence, large language models

Cite; Sungur U, Arıkan Y, Türkay AT, Polat H. The Responses of Artificial Intelligence to Questions About Urological Emergencies: A Comparison of 3 Different Large Language Models. New J Urol. 2025;20(2):89-96. doi: <https://doi.org/10.33719/nju1645041>

INTRODUCTION

Urological emergencies are clinical conditions that require immediate initiation of treatment, as delays in patient admission can lead to irreversible consequences (1). While time is critical in testicular torsion, early medical and surgical intervention is essential in conditions such as urosepsis, which can lead to multiorgan dysfunction and the need for intensive care (2,3).

As technology has advanced, both patients and healthcare providers have increasingly turned to the Internet to research the conditions they encounter (4). Large language models (LLM) developed using artificial intelligence (AI) technology have demonstrated the ability to respond to queries and provide rapid data on even highly specialized topics. Over the years, this rapidly evolving technology has led to the development of AI assistants such as ChatGPT, Google Gemini, and Meta Llama AI, each utilizing distinct infrastructures. The use of AI assistants in medical contexts has gained increasing attention in recent years, paving the way for numerous studies (5). Many studies have been published on the efficacy of their application in various diseases (6). However, the adequacy and reliability of the responses provided by these assistants, which are easily accessible to patients in time-sensitive urological emergencies, remain questionable.

This study aimed to evaluate the accuracy and adequacy of responses provided by three different AI-powered platforms to fundamental questions regarding urological emergencies, focusing on diagnosis, treatment management, and complications.

MATERIAL AND METHODS

Nine different topics were identified as urological emergencies: Testicular Torsion, Hematuria, Obstructive Uropathies, Penile Fracture, Urosepsis, Paraphimosis, Fournier's Gangrene, Priapism, and Trauma. For each topic, seven questions were prepared: two related to diagnosis, three related to treatment management, and two related to complications. While preparing the questions, instead of expecting lengthy responses from each AI platform, the command "Can you answer in one paragraph?" was used to request concise answers. The questions were selected focusing on frequently asked fundamental questions. The

list of questions is provided in Table 1. Each question was asked separately on new pages to ChatGPT, Gemini, and Meta Llama to prevent any influence from usage history, and responses were documented. The answers were evaluated by the authors participating in the study and scored on a scale of 1 to 4 (1: Completely incorrect, 2: Correct but misleading, 3: Correct but insufficient, 4: Completely correct). The scores were recorded and grouped based on diagnostic questions, treatment management questions, and complication-related questions, followed by statistical analysis. No real patients or patient information were shared in this study. This study was conducted following the Helsinki Declaration.

Statistical Analysis

The mean and deviation, number, and percentage values for the answers given to the questions in three different subgroups and the total for each AI model were documented. Results were analyzed using a non-parametric test. The Friedman test was employed for the comparison of the three groups. A p-value of < 0.05 was considered statistically significant. For the analysis of the study, IBM Statistical Package for Social Sciences SPSS 26.0.1 (IBM, Corp., Armonk, NY, USA) was utilized.

Large Language Models in Artificial Intelligence

ChatGPT-4 is an AI model developed by OpenAI, offering a more comprehensive language understanding and generation capacity compared to its predecessors. It is developed in the USA, with headquarters in San Francisco, California. This model can be utilized across various domains, ranging from everyday conversational language to technical or scientific texts. While interacting with a user, it comprehends the context of the question and generates appropriate responses. Additionally, it seamlessly adapts to multilingual content, enabling smooth communication in different languages.

Gemini 2.0 Flash is an AI model developed by Google DeepMind, with primary facilities located in Mountain View, California. This model is capable of processing visual and textual data simultaneously. This feature allows the model to respond to text-based questions while also interpreting and analyzing visual content. Its most notable characteristic is its ability to integrate information learned from diverse data sources, enabling it to make sense of complex scenarios.

LLama 3.2 is an AI model developed by Meta Llama AI, with major operations based in Menlo Park, California. LLama 3.2 is an AI system that stands out for its efficiency among language models. Its ability to deliver high performance with lower computational power has made it a preferred tool for large-scale projects and diverse applications. The model learns from a vast number of textual sources and provides accurate responses even in complex textual contexts.

RESULTS

When the responses to the 18 diagnosis-related questions were compared, the mean scores for ChatGPT, Gemini, and Meta Llama were calculated as 3.8 ± 0.4 , 3.8 ± 0.4 , and 3.6 ± 0.5 , respectively ($p=0.338$). ChatGPT provided completely correct answers to 15 (83.3%) questions, while Gemini and Meta Llama provided completely correct answers to 14 (77.8%) and 11 (61.1%) questions, respectively. None of the three platforms provided completely incorrect or misleading answers to any of the diagnosis-related questions.

When the responses to the 27 treatment management-related questions were compared, the mean scores for ChatGPT, Gemini, and Meta Llama were calculated as 3.9 ± 0.5 , 3.6 ± 0.8 , and 3.8 ± 0.5 , respectively ($p=0.289$). ChatGPT provided completely correct answers to 24 (88.9%) questions, while Gemini and Meta Llama provided completely correct answers to 21 (77.8%) and 22 (81.5%) questions, respectively. Gemini provided a completely incorrect answer to 1 (3.7%) question, while the other platforms had no completely wrong answers. Insufficient and misleading answers were observed in 3 (11.1%), 5 (18.5%), and 5 (18.5%) questions for ChatGPT, Gemini, and Meta Llama, respectively.

When the responses to the 18 complication-related questions were compared, the mean scores for ChatGPT, Gemini, and Meta Llama were calculated as 3.8 ± 0.5 , 3.8 ± 0.5 , and 3.6 ± 0.6 , respectively ($p=0.407$).

Overall, when considering all topics, the mean scores for ChatGPT, Gemini, and Meta Llama were calculated as 3.8 ± 0.5 , 3.7 ± 0.6 , and 3.7 ± 0.5 , respectively ($p=0.146$). ChatGPT provided completely correct answers to 54 (85.7%) questions, while Gemini and Meta Llama provided completely correct answers to 50 (79.4%) and 45 (71.4%) questions, respectively. The mean scores and percentages of correct answers for the three platforms are presented in Table 2. The mean scores of three LLMs are shown in Figure 1.

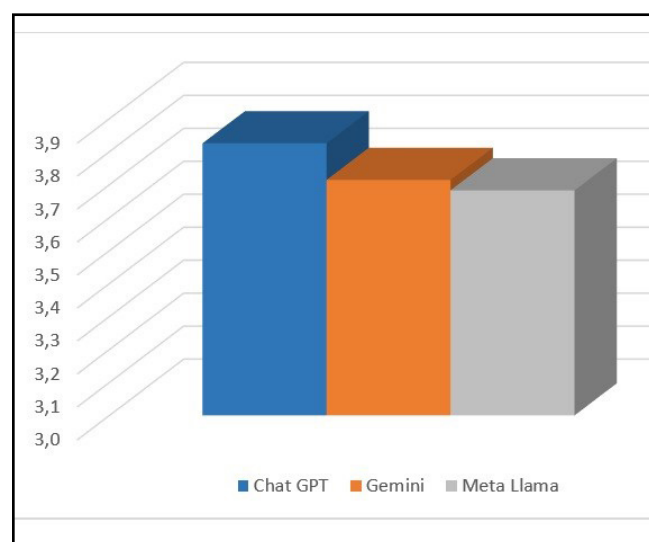


Figure 1. Mean values graph of three different AI models

Table 1. The Questions of Urological Emergencies

1.1.1 What could be the cause of sudden onset scrotal pain?
1.1.2 What is the reason for upward displacement and tenderness of the testicle in the scrotal region?
1.2.1 What is the treatment for testicular torsion, and when should it be performed?
1.2.2 Are there any non-surgical treatment options for testicular torsion?
1.2.3 From which site is surgery for testicular torsion performed?
1.3.1 What are the consequences of delayed treatment in testicular torsion?
1.3.2 Does testicular torsion lead to infertility?
2.1.1 What are the causes of blood in urine?

2.1.2 What could be the reason for black fragments in the urine?
2.2.1 What are the treatments for hematuria?
2.2.2 Is hematuria an emergency, and does it require surgery?
2.2.3 Are there non-surgical treatments for hematuria?
2.3.1 What are the outcomes if hematuria is left untreated?
2.3.2 What complications can occur in individuals with hematuria?
3.1.1 What causes decreased urine output and abdominal distension?
3.1.2 What are the reasons for nausea, vomiting, and severe flank pain?
3.2.1 What are the treatments for acute urinary retention?
3.2.2 What treatments are available for obstruction caused by kidney or ureteral stones?
3.2.3 Is urinary tract obstruction an emergency, and how soon should it be treated?
3.3.1 What complications arise from urinary retention?
3.3.2 What are the consequences of untreated obstruction due to kidney or ureteral stones?
4.1.1 What causes a sound and loss of erection during sexual activity?
4.1.2 What leads to bruising and loss of rigidity in the erect penis?
4.2.1 What is the treatment for penile fracture?
4.2.2 Are there non-surgical treatment options for penile fracture?
4.2.3 When should surgery for penile fracture be performed?
4.3.1 What are the potential complications of penile fracture?
4.3.2 What are the risks if penile fracture repair is not performed?
5.1.1 What causes fever and pain during urination, or perineal pain?
5.1.2 What causes fever in a person passing a kidney stone?
5.2.1 Is acute prostatitis an emergency, and how should it be managed?
5.2.2 How is obstructive pyelonephritis treated?
5.2.3 Are there non-invasive treatment options for fever caused by ureteral stones?
5.3.1 What are the long-term outcomes of urosepsis due to urological disease?
5.3.2 What happens if fever during stone passage is left untreated?
6.1.1 What causes bruising and swelling of the penis in children?
6.1.2 Why does the inability to retract the foreskin occur?
6.2.1 What treatments are available for phimosis-paraphimosis?
6.2.2 Are there non-surgical treatments for paraphimosis?
6.2.3 Does phimosis-paraphimosis require emergency intervention?
6.3.1 Does untreated paraphimosis lead to permanent consequences as the child grows?
6.3.2 What are the outcomes of untreated paraphimosis?
7.1.1 What causes bruising and discharge in the perineal area?
7.1.2 What is the reason for pain, hardness, and discoloration in the scrotum?
7.2.1 Is Fournier gangrene an emergency, and how is it treated?
7.2.2 Are there non-surgical treatments for Fournier gangrene?
7.2.3 How is the surgical area repaired after Fournier gangrene surgery?

7.3.1 What complications may develop in a person with Fournier gangrene?
7.3.2 What are the risks of delayed treatment for Fournier gangrene?
8.1.1 What causes prolonged and persistent penile erection?
8.1.2 How long is persistent erection considered abnormal?
8.2.1 Is priapism an emergency, and what treatments are available?
8.2.2 What are the types of priapism, and how are they diagnosed?
8.2.3 Can priapism be resolved with non-invasive methods or medications?
8.3.1 What are the long-term complications of priapism?
8.3.2 What are the risks of untreated priapism?
9.1.1 How is renal injury diagnosed after a traffic accident?
9.1.2 What causes abdominal distension and flank pain following gynecological surgery?
9.2.1 How is the treatment decision made for blunt or penetrating renal injuries?
9.2.2 What are the treatment types for bladder perforation, and how is the decision made?
9.2.3 How is treatment managed in cases of ureteral injury during surgery?
9.3.1 What complications can arise in conservatively managed renal trauma?
9.3.2 What are the risks of delayed treatment for ureteral injury during surgery?

Table 2. Comparative analysis between the different large language models about urological emergency questions

	Chat-GPT	Gemini	Meta Llama	
Question Topics, mean \pm SD	(n = 63)	(n = 63)	(n = 63)	p
Answer Proficiency, (n, %)				
Diagnostic questions	3.8 \pm 0.4	3.8 \pm 0.4	3.6 \pm 0.5	0.338
Completely Correct	15 (83.3)	14 (77.8)	11 (61.1)	
Correct but insufficient	3 (16.7)	4 (22.2)	7 (38.9)	
Correct but misleading	0 (0)	0 (0)	0 (0)	
Completely incorrect	0 (0)	0 (0)	0 (0)	
Questions regarding treatment management	3.9 \pm 0.5	3.6 \pm 0.8	3.8 \pm 0.5	0.289
Completely Correct	24 (88.9)	21 (77.8)	22 (81.5)	
Correct but insufficient	2 (7.4)	3 (11.1)	4 (14.8)	
Correct but misleading	1 (3.7)	2 (7.4)	1 (3.7)	
Completely incorrect	0 (0)	1 (3.7)	0 (0)	
Questions regarding complications	3.8 \pm 0.5	3.8 \pm 0.5	3.6 \pm 0.6	0.407
Completely Correct	15 (83.3)	15 (83.3)	12 (66.7)	
Correct but insufficient	2 (11.1)	2 (11.1)	5 (27.7)	
Correct but misleading	1 (5.6)	1 (5.6)	1 (5.6)	
Completely incorrect	0 (0)	0 (0)	0 (0)	
Total	3.8 \pm 0.5	3.7 \pm 0.6	3.7 \pm 0.5	0.146
Completely Correct	54 (85.7)	50 (79.4)	45 (71.4)	
Correct but insufficient	7 (11.1)	9 (14.3)	16 (25.4)	
Correct but misleading	2 (3.2)	3 (4.7)	2 (3.2)	
Completely incorrect	0 (0)	1 (1.6)	0 (0)	

DISCUSSION

AI applications have advanced rapidly in recent years, becoming an integral part of daily life. In the field of healthcare, they have been the subject of studies in a wide range of areas, including disease diagnosis, treatment management, prediction of complications, and interpretation of imaging and pathology examinations (6). Large language models (LLMs) powered by AI provide rapid responses by interpreting written text, scanning open sources, and summarizing information (7). This capability raises the possibility of their use by both patients and healthcare providers. Although algorithms developed for use by healthcare providers have not yet entered routine practice, their widespread adoption is anticipated in the near future. Meanwhile, the accuracy and adequacy of these platforms, which are used by patients to obtain information, have become a topic of interest. The correctness and adequacy of responses provided by LLMs in patient education have been examined across various subtopics (8). This study aimed to investigate whether the responses generated by LLMs to basic questions in urological emergencies, which may require time-sensitive decision-making, are consistent with the literature, accurate, and reliable.

Urological emergencies encompass a variety of conditions, ranging from testicular torsion, which requires immediate intervention, to hematuria, which may allow for a relatively longer diagnostic window but can still lead to urgent outcomes. The lack of awareness of testicular torsion among patients and their families, delayed hospital presentation, and the potential for organ loss or future infertility can result in devastating consequences. A study investigating the causes of delayed testicular torsion found that only 23.8% of cases underwent timely surgery. Misdiagnosis and the initial consultation with a non-urologist were identified as risk factors for orchiectomy, emphasizing the importance of proper technical training and referral to prevent delays in the diagnosis and treatment of testicular torsion (9). In cases of testicular torsion presenting with scrotal pain, consulting a large language model (LLM) in remote areas with limited healthcare access could potentially reduce the time to initial presentation, thereby preventing orchiectomy.

Another example is urolithiasis, a highly prevalent condition in the general population. Although hospital visits and the

need for analgesic treatment due to renal colic are common, patients may prefer to manage the condition without seeking medical attention based on prior experiences or anecdotal information. However, the development of fever and infection during this process may result in complicated urinary tract infections, such as pyelonephritis with obstruction, which, if left untreated, may progress to sepsis and multiorgan failure (10). Therefore, the lack of awareness among patients about the risk of sepsis in cases of renal colic complicated by infection may result in adverse outcomes in individuals who do not seek medical care. A comprehensive study examining factors related to mortality in obstructive pyelonephritis concluded that delayed decompression was associated with increased mortality, with higher rates observed in weekend admissions (11).

Another condition, penile fracture, occurs as an unexpected medical event in men. The dramatic presentation, including an audible snap during sexual intercourse and the appearance of hematoma, often signals the urgency of the situation even to untrained individuals. However, in such acute medical scenarios, the accuracy and reliability of responses provided by a free AI platform, which patients might consult to determine the urgency and potential complications, are of critical importance. In this study, we prioritized evaluating the responses of LLMs for patient education and guidance in these contexts.

Our results demonstrated that AI platforms generally provide accurate and adequate responses to basic questions regarding urological emergencies. While similar responses were predominantly observed across the three different AI assistants, no statistically significant differences were found among the results. The recent study examining the use of ChatGPT for self-diagnosis in orthopedic conditions suggested that, although it could serve as a potential initial step in accessing healthcare, it contained inconsistent results and emphasized the necessity of including clear language encouraging users to seek expert medical opinions (12). Another study investigating the use of AI platforms for emergency medical conditions highlighted that, even if the results are consistent, the ambiguity of sources and the presence of misleading information regarding the timing of medical interventions should be carefully considered due to potential risks (13). Scott et al., in their study evaluating AI-

generated responses to urology patient messages, noted that ChatGPT performed better on simple questions compared to complex ones, suggesting its potential to assist care teams (14). A recent systematic review examining the use of LLMs in patient care underscored the need for caution due to the uncertainties inherent in this technology (15).

Furthermore, ethical considerations must be addressed, particularly concerning the reliance on AI tools without professional Supervision. As AI systems evolve, ensuring transparency in source attribution and decision-making logic becomes essential. Healthcare professionals must be aware of the limitations of these tools and use them as supplementary rather than primary decision-making instruments.

Studies involving LLMs must take into account several limitations. First, the instability of the platforms used, their ongoing development, and their potential for rapid evolution over time highlight the necessity of interpreting findings based on the specific conditions of the platforms at the time of the study. We emphasize that our study focused on basic and straightforward questions, with responses summarized in paragraph form for evaluation. The likelihood of inaccuracies or misleading information may increase with more complex and lengthy responses. Since we aimed to investigate basic questions in emergency scenarios, we believe it would be inappropriate to conclude complex urological emergency conditions based on these questions and answers. Given the continuous advancement and widespread adoption of these platforms, we consider it crucial to assess and research their accuracy and reliability consistently.

CONCLUSION

Three different AI-based LLM models (ChatGPT-4, Google Gemini 2.0 Flash, and Meta Llama 3.2), which are freely accessible to patients, predominantly provided accurate responses to basic and simple questions related to urological emergency conditions. There was no significant difference in the summarized responses among the three platforms. While the use of AI assistants in patient education and guidance is becoming increasingly widespread, it is important to recognize that the current technology is not yet capable of delivering fully qualified and adequate healthcare services, given the potential for misleading or incorrect responses.

Funding: The authors declare that this study received no financial support.

Data Sharing Statement: The results of the study were not published in full or in part in the form of abstracts.

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical Approval: Chat-GPT, Google Gemini, and Meta Llama are publicly available artificial intelligence models, and there are no animal or human research participants in our study. For these reasons, our study did not require ethics committee approval.

Informed Consent: There is no need for informed consent in this study.

Authorship Contribution: Ubeyd Sungur: Conception and design, Data analysis and interpretation, Drafting the manuscript, Critical revision of the manuscript for the content, Statistical analysis. Yusuf Arıkan: Critical revision of the manuscript for the content and supervision. Ahmet Tuğrul Türkay: Data acquisition, Drafting the manuscript. Hakan Polat: Data acquisition and supervision.

REFERENCES

1. Rosenstein D, McAninch JW. Urologic emergencies. *Med Clin North Am.* 2004;88:495-518. [https://doi.org/10.1016/S0025-7125\(03\)00190-1](https://doi.org/10.1016/S0025-7125(03)00190-1)
2. Sharp VJ, Kieran K, Arlen AM. Testicular torsion: Diagnosis, evaluation, and management. *Am Fam Physician.* 2013;88:835-840. <https://pubmed.ncbi.nlm.nih.gov/24364548/>
3. Wagenlehner FM, Lichtenstern C, Rolfes C, et al. Diagnosis and management for urosepsis. *Int J Urol.* 2013;20:963-970. <https://doi.org/10.1111/iju.12200>
4. Stoumpos AI, Kitsios F, Talias MA. Digital Transformation in Healthcare: Technology Acceptance and Its Applications. *Int J Environ Res Public Health.* 2023;20:3407. <https://doi.org/10.3390/ijerph20043407>
5. Topol EJ. High-performance medicine: the convergence

- of human and artificial intelligence. *Nat Med.* 2019;25:44-56. <https://doi.org/10.1038/s41591-018-0300-7>
6. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc Neurol.* 2017;2:230-243. <https://doi.org/10.1136/svn-2017-000101>
 7. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ.* 2023;23:689. <https://doi.org/10.1186/s12909-023-04698-z>
 8. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artif Intell Rev.* 2024;57:1-48. <https://doi.org/10.1007/s10462-024-10921-0>
 9. Yi H, Wang D, Wu X, et al. Analysis of factors associated with delayed diagnosis and treatment of testicular torsion in 1005 cases from Chongqing city, China: a cross-sectional study. *Sci Rep.* 2023;13:1-10. <https://doi.org/10.1038/s41598-023-49820-9>
 10. Hsiao CY, Chen TH, Lee YC, et al. Urolithiasis Is a Risk Factor for Uroseptic Shock and Acute Kidney Injury in Patients With Urinary Tract Infection. *Front Med.* 2019;6:288. <https://doi.org/10.3389/fmed.2019.00288>
 11. Haas CR, Li G, Hyams ES, Shah O. Delayed Decompression of Obstructing Stones with Urinary Tract Infection is Associated with Increased Odds of Death. *J Urol.* 2020;204:1256-1262. <https://doi.org/10.1097/JU.0000000000001182>
 12. Kuroiwa T, Sarcon A, Ibara T, et al. The Potential of ChatGPT as a Self-Diagnostic Tool in Common Orthopedic Diseases: Exploratory Study. *J Med Internet Res.* 2023;25:e47621. <https://doi.org/10.2196/47621>
 13. Yau JYS, 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;26:e60291. <https://doi.org/10.2196/60291>
 14. Scott M, Muncey W, Seranio N, et al. Assessing Artificial Intelligence-Generated Responses to Urology Patient In-Basket Messages. *Urol Pract.* 2024;11:793-798. <https://doi.org/10.1097/UPJ.0000000000000637>
 15. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med.* 2025;5:26. <https://doi.org/10.1038/s43856-024-00717-2>