MEDICAL RECORDS-International Medical Journal

Research Article

A Comparative Study on the Question-Answering Proficiency of Artificial Intelligence Models in Bladder-Related Conditions: An Evaluation of Gemini and ChatGPT 4.o

Mustafa Azizoglu[1,2], Sergey Klyuev[3]

[1]İstanbul Esenyurt Necmi Kadıoğlu State Hospital, Department of Pediatric Surgery, İstanbul, Türkiye
[2]İstinye University, Health Science Faculty, Department of Stem Cell and Tissue Engineering, İstanbul, Türkiye
[3]AO GK MEDSI, Department of Pediatric Surgery, Moscow, Russia

Abstract

Aim: The rapid evolution of artificial intelligence (AI) has revolutionized medicine, with tools like ChatGPT and Google Gemini enhancing clinical decision-making. ChatGPT's advancements, particularly with GPT-4, show promise in diagnostics and education. However, variability in accuracy and limitations in complex scenarios emphasize the need for further evaluation of these models in medical applications. This study aimed to assess the accuracy and agreement between ChatGPT 4.o and Gemini AI in identifying bladder-related conditions, including neurogenic bladder, vesicoureteral reflux (VUR), and posterior urethral valve (PUV).
Material and Method: This study, conducted in October 2024, compared ChatGPT 4.o and Gemini AI's accuracy on 51 questions about neurogenic bladder, VUR, and PUV. Questions, randomly selected from pediatric surgery and urology materials, were evaluated using accuracy metrics and statistical analysis, highlighting AI models' performance and agreement.
Results: ChatGPT 4.o and Gemini AI demonstrated similar accuracy across neurogenic bladder, VUR, and PUV questions, with true response rates of 66.7% and 68.6%, respectively, and no statistically significant differences ($p>0.05$). Combined accuracy across all topics was 67.6%. Strong inter-rater reliability ($\kappa=0.87$) highlights their agreement.
Conclusion: This study highlights the comparable accuracy of ChatGPT-4.o and Gemini AI across key bladder-related conditions, with no significant differences in performance.

Keywords: ChatGPT, Gemini, articifial intelligence, bladder

INTRODUCTION

The rapid evolution of artificial intelligence (AI) has significantly reshaped multiple disciplines, including medicine and healthcare (1-5). Among the most prominent advancements in AI are large language models (LLMs) like ChatGPT 4.o and Google Gemini (2-4). ChatGPT, developed by OpenAI, employs state-of-the-art natural language processing to generate human-like responses (2-4). Initially launched with GPT-3.5 in November 2022 and succeeded by GPT-4 in March 2023, ChatGPT has demonstrated considerable promise in various fields, including medical diagnostics and education. Similarly, Google Gemini, a competitor, integrates advanced neural networks to interpret and respond to complex medical inquiries. These advancements position both platforms as potential tools for supporting clinical decision-making (5,6).

Prior studies have explored ChatGPT's accuracy in standardized medical exams, revealing its potential to complement traditional methods. For example, GPT-4 has outperformed its predecessor GPT-3.5 in answering medical licensing questions, emphasizing its growing accuracy and reliability (7).

Despite the advancements, concerns persist regarding the reliability and consistency of these tools (8,9). Studies comparing ChatGPT, Gemini, and other AI models in specific medical contexts highlight variability in diagnostic accuracy and comprehensiveness. ChatGPT has been noted for its higher accuracy in certain scenarios but has also demonstrated limitations in nuanced or ambiguous

CITATION
Azizoglu M, Klyuev S. A Comparative Study on the Question-Answering Proficiency of Artificial Intelligence Models in Bladder-Related Conditions: An Evaluation of Gemini and ChatGPT 4.o. Med Records. 2025;7(1):201-5. DOI:1037990/medr.1601528

Received: 14.12.2024 **Accepted:** 10.01.2025 **Published:** 15.01.2025
Corresponding Author: Mustafa Azizoglu, İstanbul Esenyurt Necmi Kadıoğlu State Hospital, Department of Pediatric Surgery, İstanbul, Türkiye İstinye University, Health Science Faculty, Department of Stem Cell and Tissue Engineering, İstanbul, Türkiye
E-mail: mdmazizoglu@gmail.com

cases. Similarly, Gemini, while robust in specific areas, occasionally falters in providing context-sensitive responses, necessitating further evaluation (1).

This study aimed to assess the accuracy and agreement between ChatGPT 4.o and Gemini AI in identifying bladder-related conditions, including neurogenic bladder, vesicoureteral reflux (VUR), and posterior urethral valve (PUV).

MATERIAL AND METHOD

This study was conducted in October 2024. Since it involved an AI-based analysis, ethical approval was not required. The study focused on three primary bladder-related topics: VUR, PUV, and neurogenic bladder. A total of 51 questions were prepared, with 17 questions from each topic to ensure balanced representation across all areas. The questions were randomly selected from pediatric surgery and pediatric urology board examination study materials to reflect clinically relevant scenarios.

The prepared questions were simultaneously presented to both Gemini AI and ChatGPT 4.o models. Each model's responses were independently recorded and evaluated against a pre-determined answer key. The number of correct and incorrect answers provided by each model was calculated separately for each topic to assess their accuracy.

A text was prepared to train both artificial intelligence models in the question-and-answer format. This text was taught to both AI models and was as follows. "I am conducting a study to assess the diagnostic accuracy and agreement between ChatGPT and Gemini AI in identifying bladder-related conditions, including neurogenic bladder, VUR, and PUV. For this purpose, I will ask each of you a total of 51 questions, with 17 questions per topic. The questions will be the same for both models and are selected from the European board examination study materials in pediatric surgery and pediatric urology. Please answer these questions to the best of your ability. At the end, your correct and incorrect answers will be compared to the official answer key. And then, the results will be compared for a study." The responses provided by the models are presented in Figures 1 and 2. The questions asked to the AI were multiple choice questions (5 options) in a test format.

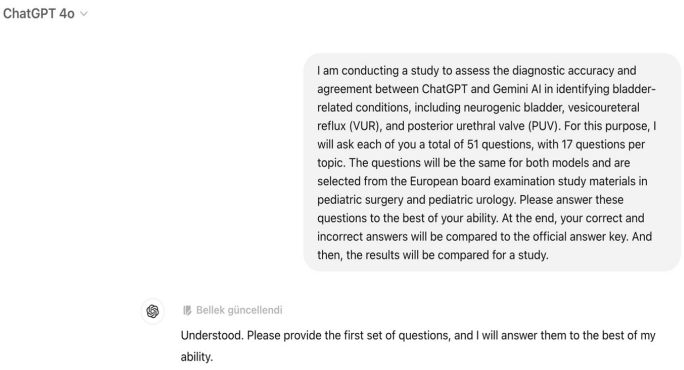

Figure 1. Instructions for AI model evaluation (for ChatGPT 4.o)

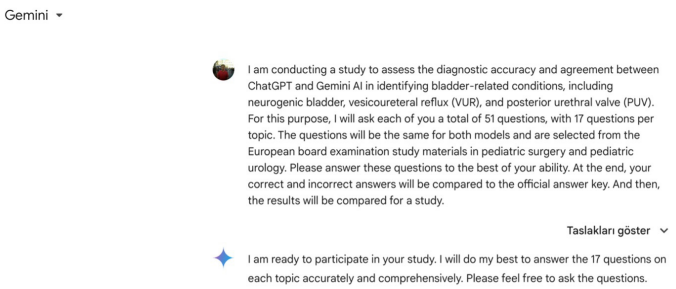

Figure 2. Instructions for AI model evaluation (for Gemini AI)

Statistical Analysis

For statistical analysis, the number of true and false responses for both AI models was compared for each topic using accuracy metrics and Cohen's Kappa to evaluate inter-model agreement. The categorical variables were compared with chi-square test. Univariate logistic regression analysis was performed to examine the factors associated with correct responses, using neurogenic bladder as the reference topic and ChatGPT 4.o as the reference AI model. Odds ratios (OR) and p-values were calculated to assess the statistical significance of the differences between the models and topics. Data visualization was achieved through a Sankey diagram to demonstrate the flow and distribution of correct and incorrect responses across topics and models. All statistical analyses were conducted using Jamovi software 2.4, with a significance threshold set at $p<0.05$.

RESULTS

For neurogenic bladder, ChatGPT 4.o correctly answered 11 questions (64.7%), while Gemini AI correctly answered 13 questions (76.5%), resulting in a combined true response rate of 70.6% (n=24). The false response rates were 35.3% for ChatGPT 4.o and 23.5% for Gemini AI, with a total false response rate of 29.4% (n=10). The difference in performance between the two AI models on neurogenic bladder-related questions was not statistically significant (p=0.452). Regarding VUR, ChatGPT 4.o and Gemini AI achieved true response rates of 64.7% (n=11) and 58.8% (n=10), respectively, with a total true response rate of 61.8% (n=21). The false response rates were 35.3% and 41.2% for ChatGPT 4.o and Gemini AI, respectively, contributing to a combined false response rate of 38.2% (n=13). The statistical comparison indicated no significant difference between the models on VUR-related questions (p=0.724). For PUV, both ChatGPT 4.o and Gemini AI performed identically, each with a true response rate of 70.6% (n=12) and a false response rate of 29.4% (n=5). The total response rates were identical across the models on PUV-related questions (p=1.000). Across all conditions, the combined true response rate was 66.7% for ChatGPT 4.o and 68.6% for Gemini AI, with a total true response rate of 67.6% (n=69). The combined false response rates were 33.3% for ChatGPT 4.o and 31.4% for Gemini AI, with a total false response rate of 32.4% (n=33). The overall performance comparison between the AI models showed no statistically significant difference (p=0.832) (Table 1).

Table 1. Accuracy between ChatGPT 4.o and Gemini AI

		ChatGPT 4.o	Gemini AI	Total	P-value
Neurogenic bladder	**True**	11 (64.7%)	13 (76.5%)	24 (70.6%)	0.452
	False	6 (35.3%)	4 (23.5%)	10 (29.4%)	
Vesicoureteral reflux	**True**	11 (64.7%)	10 (58.8%)	21 (61.8%)	0.724
	False	6 (35.3%)	7 (41.2%)	13 (38.2%)	
Posterior urethral valve	**True**	12 (70.6%)	12 (70.6%)	24 (70.6%)	1.000
	False	5 (29.4%)	5 (29.4%)	10 (29.4%)	
Total	**True**	34 (66.7%)	35 (68.6%)	69 (67.6%)	0.832
	False	17 (33.3%)	16 (31.4%)	33 (32.4%)	

A univariate logistic regression analysis was conducted to evaluate the factors associated with correct responses provided by the AI models. The analysis considered the topics (neurogenic bladder, VUR, and PUV) and the AI model used (ChatGPT 4.o vs. Gemini AI). For the topic, using neurogenic bladder as the reference category, the OR for VUR was 0.673 (p=0.443), indicating no statistically significant difference in the likelihood of correct responses between neurogenic bladder and VUR. Similarly, the OR for PUV was 1.000 (p=1.000), showing no difference in the probability of correct responses between neurogenic bladder and PUV. When comparing the AI models, using ChatGPT 4.o as the reference, the OR for Gemini AI was 1.09 (p=0.832) (Table 2).

Table 2. Univariate logistic regression analysis of AI model performance

	Univariate analysis	
	OR	**p-value**
Topic (ref: neurogenic bladder)		
VUR	0.673	0.443
PUV	1.000	1.000
AI-model (ref: ChatGPT 4.o)	1.09	0.832

The Cohen's kappa value for inter-rater reliability between ChatGPT 4.o and Gemini AI is 0.87, indicating a strong level of agreement (κ=0.87). The Sankey diagram demonstrates how each AI model's predictions distribute across topics and their respective accuracies (Figure 3).

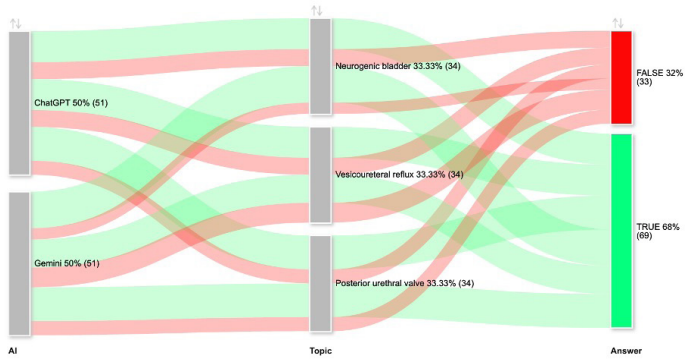

Figure 3. Sankey diagram

DISCUSSION

In our study, ChatGPT 4.o and Gemini AI showed comparable accuracy across neurogenic bladder, VUR, and PUV, with total true response rates of 66.7% and 68.6%, respectively, and no significant performance differences (p>0.05). Logistic regression confirmed no significant variability between topics or models. Strong inter-rater agreement (κ=0.87) highlights their potential as reliable tools.

The widespread adoption of artificial intelligence has significantly increased its use in various fields, including medicine (1-5,10). Numerous studies have been conducted in disciplines such as orthopedics, dermatology, and neurology to evaluate the accuracy of AI models in answering exam questions (11-14). While these studies have shown promising results, many also highlight the frequent occurrence of incorrect answers alongside correct ones. Furthermore, it has been emphasized that models like ChatGPT can occasionally produce misleading or incorrect responses, referred to as "hallucinations," underscoring the importance of being aware of such inaccuracies (15,16).

In this study, a total of 51 questions related to three key bladder conditions—neurogenic bladder, VUR, and PUV—were posed to both ChatGPT 4.o and Gemini AI, and their responses were compared. To minimize potential confounding factors, all questions were designed as multiple-choice. These questions were randomly selected from the European Board Examination study materials in pediatric surgery and pediatric urology. To the best of our knowledge, this is the first study to compare ChatGPT 4.o and Gemini AI models specifically for bladder-related conditions.

In a study on orthopedic and trauma surgery, the American and British equivalents of the French DES exam, the OITE (ABOS) and FRCS-Trauma and Orthopaedics, were tested. ChatGPT-4 generally performed well, approaching but not surpassing the residents' average scores (10). In another study, in the field of General Orthopedics, Ulus et al. reported GPT-4 achieving a higher success rate (75%) compared to GPT-3.5 (45%) (p=0.053) (17). In the

Traumatology domain, GPT-4 showed a notable success rate of 80%, significantly outperforming GPT-3.5 (p=0.010). Greif et al. found ChatGPT correctly identified the top diagnosis for 12 of 32 cases (37.5%), compared to 5 of 10 cases (50%) in their study (18). ChatGPT-3.5 included the correct diagnosis in 81% of cases, while ChatGPT-4 achieved 80%. These findings indicate ChatGPT-4 consistently outperforms its predecessor, GPT-3.5, as also demonstrated in the study by Azizoğlu et al (19). Furthermore, Demir et al. found ChatGPT 4.o provided more detailed and accurate answers to patient questions about keratoconus than Google Gemini and Microsoft Copilot, with 92% of responses rated as "agree" or "strongly agree" (1). Ronbinson et al highlight the potential of AI chatbots in addressing common urologic queries, particularly for benign prostatic hyperplasia (BPH) (20). Chatbot-generated responses demonstrated comparable accuracy, greater completeness, and higher perceived empathy compared to urologists' responses. However, patient trust and perception of AI remain challenges. They concluded that future studies should explore AI integration in broader urologic domains, address ethical concerns, and focus on improving patient confidence in AI-driven healthcare communications. In a meta-analysis of 193 studies, Zong et al. concluded that MedExamLLM is an open-source, freely accessible platform offering comprehensive performance evaluations and evidence on LLM capabilities in medical exams globally (21). It serves as a vital resource for educators, researchers, and developers in clinical medicine and AI. Despite its value, limitations include potential data source biases and exclusion of non-English studies, warranting future research to enhance LLM performance across diverse contexts. In our current study, we evaluated responses to bladder-related questions using ChatGPT 4.o and Gemini AI, assessing the accuracy and agreement of both AI tools. Our study found that ChatGPT 4.o and Gemini AI demonstrated comparable accuracy across conditions, with no significant performance differences, highlighting strong reliability.

Despite its strengths, this study has several limitations. First, the sample size of 51 questions may not comprehensively represent the complexity of bladder-related conditions. Additionally, the study relied on pre-determined answer keys, which may not account for nuanced clinical variations. Furthermore, the AI models were assessed using text-based questions, excluding imaging or laboratory data integration, which are critical in real-world diagnostics. The absence of clinical context in the questions may limit the applicability of findings to actual medical scenarios. Lastly, the models' outputs were evaluated in isolation, without considering collaborative human-AI interactions, potentially overlooking their full diagnostic potential.

CONCLUSION

In conclusion, this study highlights the comparable accuracy of ChatGPT 4.o and Gemini AI across key bladder-related conditions, with no significant differences in performance. These findings suggest the potential utility of AI tools in bladder-related conditions, emphasizing their reliability. However, further research is required to address limitations such as sample diversity and real-world applicability, ensuring these technologies can effectively complement clinical decision-making.

Financial disclosures: *The authors declared that this study has received no financial support.*

Conflict of interest: *The authors have no conflicts of interest to declare.*

Ethical approval: *Since the study does not involve human or animal subjects, obtaining ethical committee approval is not required.*

REFERENCES

1. Demir S. Evaluation of responses to questions about keratoconus using ChatGPT-4.0, Google Gemini and Microsoft Copilot: a comparative study of large language models on Keratoconus. Eye Contact Lens. 2024 Dec 4. doi: 10.1097/ICL.0000000000001158. [Epub ahead of print].
2. Sun SH, Chen K, Anavim S, et al. Large language models with vision on diagnostic radiology board exam style questions. Acad Radiol. 2024 Dec 3. doi: 10.1016/j.acra.2024.11.028. [Epub ahead of print].
3. Galvis-García E, Vega-González FJ, Emura F, et al. Inteligencia artificial en la colonoscopia de tamizaje y la disminución del error. Cir Cir. 2023;91:411-21.
4. De Busser B, Roth L, De Loof H. The role of large language models in self-care: a study and benchmark on medicines and supplement guidance accuracy. Int J Clin Pharm. 2024 Dec 7. doi: 10.1007/s11096-024-01839-2. [Epub ahead of print].
5. Ardila CM, Yadalam PK. ChatGPT's influence on dental education: methodological challenges and ethical considerations. Int Dent J. 2024 Dec 6. doi: 10.1016/j.identj.2024.11.014. [Epub ahead of print].
6. Meo AS, Shaikh N, Meo SA. Assessing the accuracy and efficiency of Chat GPT-4 Omni (GPT-4o) in biomedical statistics: Comparative study with traditional tools. Saudi Med J. 2024;45:1383-90.
7. Chen Y, Huang X, Yang F, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ. 2024;24:1372.
8. Bilgin IA, Percem AK, Aslan O. Artificial intelligence and robotic surgery in colorectal cancer surgery. J Clin Trials Exp Investig. 2024;3:83-4.
9. Yılmaz M. Revolutionizing laboratory medicine: the critical role of artificial intelligence and deep learning: Artificial intelligence and medical laboratory. The Injector. 2024;3:39-40.
10. Maraqa N, Samargandi R, Poichotte A, et al. Comparing performances of french orthopaedic surgery residents with the artificial intelligence ChatGPT-4/4o in the French diploma exams of orthopaedic and trauma surgery. Orthop Traumatol Surg Res. 2024 Dec 4. doi: 10.1016/j.otsr.2024.104080. [Epub ahead of print].

11. Giorgino R, Alessandri-Bonetti M, Luca A, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. Front Surg. 2023;10:1284015.

12. D'Agostino M, Feo F, Martora F, et al. ChatGPT and dermatology. Ital J Dermatol Venerol. 2024;159:566-71.

13. Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open. 2023;5:e000530.

14. Karakas C, Brock D, Lakhotia A. Leveraging ChatGPT in the pediatric neurology clinic: practical considerations for use to improve efficiency and outcomes. Pediatr Neurol. 2023;148:157-63.

15. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. 2023 Mar 15. doi: 10.48550/arXiv.2303.08774. [Preprint posted online].

16. Jin HK, Kim E. Performance of GPT-3.5 and GPT-4 on the Korean pharmacist licensing examination: comparison study. JMIR Med Educ. 2024;10:e57451.

17. Ulus SA. How does ChatGPT perform on the European board of orthopedics and traumatology examination? A comparative study. Academic Journal of Health Sciences. 2023;38:43-6.

18. Greif C, Mpunga N, Koopman IV, et al. Evaluating the effectiveness of ChatGPT4 in the diagnosis and workup of dermatologic conditions. Dermatol Online J. 2024;30. doi: 10.5070/D330464104.

19. Azizoglu M, Aydogdu B. How does ChatGPT perform on the European Board of Pediatric Surgery examination? A randomized comparative study. Academic Journal of Health Sciences. 2024;39:23-6.

20. Robinson EJ, Qiu C, Sands S, et al. Physician vs. AI-generated messages in urology: evaluation of accuracy, completeness, and preference by patients and physicians. World J Urol. 2024;43:48.

21. Zong H, Wu R, Cha J, et al. Large Language Models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res. 2024;26:e66114.