Evaluating the Performance of different large language models on health consultation and patient education in urolithiasis

Haifeng Song (shf1990@hotmail.com)
Tsinghua University
Yi Xia
Southeast University
Zhichao Luo
Tsinghua University
Hui Liu
Tsinghua University
Yan Song
Sheng Jing Hospital of China Medical University
Xue Zeng
Tsinghua University
Tianjie Li
Tsinghua University
Guangxin Zhong
Tsinghua University
Jianxing Li
Tsinghua University
Ming Chen
Southeast University
Guangyuan Zhang
Southeast University
Bo Xiao
Tsinghua University

Research Article

Keywords: urolithiasis, health consultation, large language model, ChatGPT, artificial intelligence

Posted Date: August 29th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3293294/v1

Abstract

Objectives

To evaluate the effectiveness of four large language models (LLMs) (Claude, Bard, ChatGPT4, and New Bing) that have large user bases and significant social attention, in the context of medical consultation and patient education in urolithiasis.

Materials and methods

In this study, we developed a questionnaire consisting of twenty-one questions and two clinical scenarios related to urolithiasis. Subsequently, clinical consultations were simulated for each of the four models to assess their responses to the questions. Urolithiasis experts then evaluated the model responses in terms of accuracy, comprehensiveness, legibility, human care, and clinical case analysis ability based on a predesigned 5-point Likert scales. Visualization and statistical analyses were then employed to compare the four models and evaluate their performance.

Results

All models yielded relatively qualified results, except for Bard, which failed to provide a valid response to Question 13. Claude consistently scored the highest in all dimensions compared with the other three models. ChatGPT4 ranked second in accuracy, with a relatively stable output across multiple tests, but shortcomings were observed in empathy and care for counsellors. The Bard model exhibited the lowest accuracy and overall performance. Claude and ChatGPT4 both had a high capacity to analyze clinical cases of urolithiasis. Overall, the Claude model emerged as the best performer in urolithiasis consultations and education.

Conclusion

Claude demonstrated superior performance compared with the other three in urolithiasis consultation and education. This study highlights the remarkable potential of LLMs in medical health consultations and patient education, although professional review, further evaluation, and modifications are still required.

Introduction

Urolithiasis is a common disease of the urinary system, with an incidence rate ranging from 1–10% worldwide. The high recurrence rate and detrimental effects of urolithiasis impose significant health and economic burdens on both patients and society. Therefore, early diagnosis, intervention, and strict follow-up are crucial for effective urolithiasis management, complication reduction, and prevention of disease recurrence(1, 2). However, patients often lack reliable information regarding diagnosis, prognosis,

treatment options, side effects, and preventive measures at all stages of decision-making and treatment. Providing appropriate medical consultation services to patients with urolithiasis or suspected cases plays an important role in patient education, which can significantly improve prognosis and alleviate the burden of urolithiasis(3).

Large Language Models (LLMs) represent a type of artificial intelligence (AI) model that generates natural language text from copious amounts of data. Utilizing deep neural networks and machine learning algorithms, such as transformers, LLMs are trained on vast quantities of text data and are capable of various natural language tasks, including summarization, translation, question answering, conversation, and even poetry generation. LLMs, as represented by ChatGPT, have also shown unique innovation and remarkable efficiency in various medical scenarios, such as answering medical and public health inquiries(4, 5), facilitating computer-aided diagnosis(6), providing treatment advice(7), and providing healthcare education(8). These applications demonstrate the potential of LLMs in improving the quality and efficiency of medical consultations and patient education.

In urology, the application and evaluation of LLMs are limited. Görtz et al. established a chatbot named PROSCA and evaluated its performance in providing patient information regarding early detection of prostate cancer(9). The showed that PROSCA was well-received by patients and served as an additional informative tool to benefit them. Similarly, Zhu et al. reported that LLMs (ChatGPT, YouChat, and NeevaAI) can accurately address the fundamental inquiries from patients with prostate cancer, and analyze specific scenarios to a certain extent(10). However, the performance of LLMs for the consultation and education of patients with urolithiasis remains unexplored and requires evaluation. This study aimed to evaluate the effectiveness of four large language models (Bard, Claude, ChatGPT4, and New Bing) that have large user bases, robust qualifications, and significant social attention, in the context of medical consultations and patient education regarding urolithiasis.

Materials and Methods

We designed a set of urolithiasis-related questions and clinical scenarios ranging from basic to advanced. Subsequently, we simulated clinical consultations for each of the four models to assess their responses to these inquiries. The answers provided by the models were evaluated by urolithiasis experts based on objective and rigorous standards, and the results were collected. Finally, statistical analyses were performed to compare the scores of the four models and evaluate their performance.

Questions and clinical scenarios design:

A set of 21 questions that address the common concerns of patients with urolithiasis was collected. These questions were curated through an analysis of queries from online consultation platforms, surveys conducted among hospitalized urolithiasis patients, and incorporation of the researchers' clinical experience. The questions were categorized into simple and complex types with difficulty levels ranging

from general urolithiasis knowledge to cutting-edge diagnostic and therapeutic techniques. Two case scenarios with different complexities were created based on clinical experience.

Model selection and test

Four LLMs that possess substantial user bases, impressive backgrounds, and significant social attention were evaluated. The models included in this study were Bard, Claude, ChatGPT4, and New Bing. Each model underwent three rounds of testing for all questions and the resulting outcomes were recorded. The tests were performed in late April 2023.

Scoring Criteria and procedure

5-point Likert scale was used to evaluate the LLM outputs based on five dimensions: accuracy, comprehensiveness, legibility, human caring, and case analysis performance. Accuracy pertains to the correctness and adherence to scientific knowledge and clinical guidelines in the provided information. Comprehensiveness evaluates the extent to which the response addresses all relevant aspects of the question or case scenario. Legibility assesses the clarity of the response in terms of its logical structure, language usage, and ease of comprehension for the target audience. Human care measures the degree to which the response demonstrates empathy towards patients, addresses their concerns, and respects their values and preferences. Lastly, the case analysis performance measures the proficiency of the LLM in interpreting case scenarios, identifying key issues, and providing a coherent and effective approach or solution. The Likert scale was defined as follows:

1-Unacceptbale: The LLM's response significantly lacks in the particular criterion.

2-Poor: The LLM's response lacks in the criterion but not to a severe extent.

3-Fair: The LLM's response adequately but not exceptionally meets the criterion.

4-Good: The LLM's response aligns well with the criterion.

5-Excellent: The LLM's response excels in the criterion, exceeding the standard expectation.

Three associate chief physicians (Y. S., B. X., and G. Z.) with expertise in urolithiasis from different medical centers were recruited as reviewers to evaluate the results of the three tests across the five dimensions. To comprehensively evaluate the models, we synthesized the scores of the three reviewers as the final scores for each model. Discrepancies in scoring were resolved by taking the median value or following the majority view.

Visualization and Statistical analysis

The online tool HIPLOT (https://hiplot.org) was used to create dot plots, illustrating the individual model scores for each question, and radar plots, enabling a comparison of the overall scores of various models. All statistical analyses were performed using R version 4.3.0 (The R Foundation for Statistical Computing, Vienna, Austria). Non-parametric Wilcoxon tests were used to compare scores between different groups.

Results

Characteristics of questions and clinical scenarios

After collection and screening, we included 21 questions covering various aspects of urolithiasis, ranging from concepts to diagnosis, treatment, prevention, and follow-up. Additionally, we designed two clinical case that dealt with emergencies caused by ureteral stones. One case involved a straightforward case of renal colic caused by ureteral stone obstruction, while the other involved a complex situation leading to septic shock caused by ureteral stone obstruction. The questions and cases are shown in Table 1.

Table 1
Questions and clinical scenarios

Category	Number	Question
I. Overview	1	What is kidney stone?
	2	What is the specific mechanism of kidney stone formation?
	3	Are kidney stones hereditary?
	4	What harm can kidney stones cause?
	5	What are the characteristics of kidney stones with different compositions?
II. Diagnosis	6	How to determine if I have urolithiasis?
	7	What auxiliary examinations should kidney stone patients undergo for diagnosis and assessment?
	8	Why do kidney stones cause pain?
	9	What are the characteristics of pain caused by kidney stones and how to differentiate it from other diseases?
	10	Is it possible not to have a CT scan if I have kidney stones because of the radiation?
III. Treatment	11	Can kidney stones be eliminated naturally, and if so, how to increase the probability of natural elimination of stones
	12	How to perform conservative treatment for kidney stones
	13	How to relieve pain and other symptoms caused by kidney stones?
	14	What are the various treatment methods for urolithiasis and their respective characteristics?
	15	What type of stones are suitable for extracorporeal shock wave lithotripsy?
	16	Under what circumstances does kidney stone require surgical treatment?
	17	What is the difference between ureteroscopic lithotripsy and percutaneous nephrolithotomy for kidney stones?
	18	What are the potential risks and complications of surgical treatment for kidney stones?
IV. Follow-up and Prevention	19	What type of kidney stones are prone to recurrence?
	20	How to prevent kidney stone recurrence through diet and lifestyle adjustments?
	21	How to follow up after kidney stone surgery?

Category	Number	Question
Cases 1	22	Young male, 23 years old, sudden onset of left lumbar back pain for 2 hours, accompanied by nausea and vomiting, admitted to the emergency department, physical examination revealed left lumbar percussion pain, routine urine test showed occult blood, ultrasound indicated left ureteral dilation and mild hydronephrosis of the left kidney. According to the medical history, what disease should be considered, and what additional tests should be performed?
Case 2	23	Female patient, 62 years old, left lumbar pain accompanied by fever for 2 days, admitted to the emergency department. Body temperature 38 degrees Celsius, heart rate 110 beats/min, blood pressure 80/40mmHg, blood routine showed white blood cell count $23x10^9$/L, neutrophil percentage 95%, kidney function showed blood creatinine 300μmol/L, blood glucose 23mmol/L. Abdominal CT indicated left kidney hydronephrosis, perinephric infiltration, and a 0.8 cm high-density shadow at the distal end of the left ureter. What disease should this patient consider? What additional tests are needed, how to urgently deal with it? What is the cause of the patient's hypotensive shock?

Performance of the models in answering questions of urolithiasis

The detailed outputs of the different models for the questions and cases are presented in the Supplementary Material. We summarized the ratings given by the experts for each question.

The accuracy scores of the four models for all questions are shown in Fig. 1A. The Claude model performed the best in terms of accuracy, and most answers scored 4 points or higher, with only two questions scoring lower than 4 points. ChatGPT4 ranked second in terms of accuracy. Bard's responses had the lowest accuracy, with most questions scoring 3 points or less and question 13 scoring only 1–2 points. ChatGPT4 exhibited the most stable output, with the smallest fluctuation in scores across the three tests.

The scores for comprehensiveness of the responses are shown in Fig. 1B. The Claude model again had the highest median score of 5. ChatGPT4 obtained a median score of 4, while both the Bard and New Bing models had median scores of 3. Notably, the Bard model demonstrated the most stable output results.

Figure 1C shows the readability scores of the output answers; all four models had a median score of 4. Based on the performance in the three tests, ChatGPT4 yielded the most stable output.

Figure 1D shows the scores for human care. The Claude model exhibited the highest median score of 4, while the other three models achieved a median score of 3. Furthermore, Claude had the most stable output, with almost all questions scoring 4 points and exhibiting minimal variation in scores.

Case analysis performance of the models

We conducted a detailed evaluation and comparative analysis of the performance of the two case analyses. The ratings given by experts showed that both the Claude and ChatGPT4 models had ratings of 4 or above for both cases in terms of their capacity to analyze cases. However, Bard's ratings were significantly lower than those of the other three models. The performance of each model is shown in Fig. 2.

The comparison of overall scores for the models

We calculated the score rate of each model for each question in different dimensions by dividing the total score of each model for each question by the full score, The corresponding radar charts were then generated based on this data. Figure 3A, 3B, and 3C present the results of the three tests. Notably, Claude demonstrated remarkable stability across all three tests, consistently achieving scores of 80 or higher in all dimensions. ChatGPT4's performance was relatively balanced and excellent in all aspects, except for human care. Both Claude and ChatGPT4 showed significantly higher overall scores than the other two models, with Claude performing noticeably better than ChatGPT4 (Fig. 3D).

Discussion

The application of ChatGPT and other LLMs in the medical field has sparked a lively debate in the academic community since their emergence(11). Several studies have evaluated the performance of ChatGPT3.5 in various medical domains and tasks. For example, ChatGPT3.5 passed all three stages of the United States Medical Licensing Examination (USMLE), a comprehensive assessment for medical licensure in the US (12). In addition, when employed in the radiological diagnosis of breast cancer screening and the evaluation of breast pain severity, ChatGPT exhibited moderate accuracy(13). ChatGPT3.5 also demonstrated the potential to improve health education by providing consultation to healthcare providers and offering accessible and understandable medical knowledge to the general public(14, 15). Thus, LLMs, such as ChatGPT, have shown unique advancements and remarkable efficiencies in solving medical issues. However, different LLMs have different strengths and limitations in terms of functionality, performance, and reliability. Similarly, various AI LLMs exhibit diverse features and capabilities when applied to different medical scenarios.

In this study, we compared the performance of four state-of-the-art LLMs currently available, including Bard, Claude, ChatGPT4, and New Bing, in health consultation and patient education in urolithiasis. Claude was developed by Anthropic and founded by former employees of OpenAI. Bard, developed by Google, was built based on its own language model, LaMDA. ChatGPT4, developed by OpenAI, was built based on the latest GPT-4 language model. New Bing developed by Microsoft, is also based on GPT4; unlike the other three models, New Bing can output images and access the Internet for real-time data and information. It can also provide sources and references for its answers. Overall, the findings of the present study are promising. All models yielded relatively qualified results, except for Bard, which failed to provide a valid response to Question 13. The study indicated that the Claude model consistently outperformed the other three models, exhibiting the highest scores across all dimensions including

accuracy, comprehensiveness, readability, and empathy, regardless of question complexity. ChatGPT4 ranked second in accuracy, with a relatively stable output across multiple tests, but there were still shortcomings in empathy and care for counsellors. The Bard model had the lowest accuracy and overall performance, with lower scores in comprehensiveness, readability, and human caring. The case analysis evaluation also showed that both the Claude and ChatGPT4 models demonstrated strong capabilities in case analysis, while Bard's performance in this regard was significantly inferior. Overall, the Claude model emerged as the top performer in urolithiasis consultations and education.

Our study had several important limitations. First, the relatively small number of inputs, specifically the urolithiasis patients' questions and case scenarios, restricted the depth and scope of our analysis. Increasing the number of questions and cases and categorizing them by question type and complexity could reveal more specific and profound differences in the performance of LLMs in simulating different clinical tasks. Second, LLMs are being rapidly updated and our study only represents the performance of the four models in their respective versions until late April 2023. With ongoing model updates, their performances may improve over time. Third, we designed only one language pattern for questioning, but it is important to recognize that different questioning styles may yield varying results from the models. Therefore, it is necessary to design more standardized and rigorous prompts and conduct more tests to evaluate the output of these models in the future.

With the rapid development of natural language processing and artificial intelligence, LLMs can make full use of medical big data, and through cross-collaboration with researchers, clinical healthcare practitioners, patients, and health policymakers, they will have an unprecedented impact on all aspects of healthcare in the future and further promote a paradigm shift in healthcare(16, 17). Although current evaluations show promising prospects for LLMs' application of LLMs in healthcare consultations, certain concerns remain. These models are not specialized medical LLMs based on professional materials within the field. LLMs utilize a vast amounts of data from various internet sources for training and text generation, but these training data are not all peer reviewed and may introduce biases. The lack of transparency in the black-box nature of LLMs compromises objectivity and accuracy during the answering process(18–21). Additionally, the training data may have temporal limitations. Except for New Bing, the other three models cannot provide any sources or evidence for their claims, which may raise concerns and suspicions when the outputs deviate from current clinical practices and the latest medical advancements. Moreover, LLMs can generate erroneous content that appears reasonable from a scientific perspective(12). Due to their reliance on textual information, most models are incapable of handling medical images. In addition, they cannot account for non-quantifiable cues involved in medical consultations, such as religious beliefs, sociopsychological characteristics, and emotional shifts(22). These elements, along with the expertise of physicians, play a crucial role in addressing medical issues(23). Therefore, the use of LLMs for clinical consultation requires human intervention to verify the sources and ensure the accuracy of their outputs.

In the foreseeable future, it is imperative to acknowledge that LLMs should never serve as a complete substitute for licensed healthcare providers. Instead, they should be regarded as supplementary tools that

can improve clinical decision-making. Healthcare providers should be aware of these limitations and should use LLMs cautiously.

Conclusion

We assessed four prominent LLMs currently available and demonstrated their competence in performing assigned tasks within the field of urolithiasis consultation. Claude outperformed the other three LLMs in terms of accuracy, comprehensiveness, readability, human care, and case analysis ability. ChatGPT4 ranked second performance. The rapid advancements in artificial intelligence and natural language processing technologies have provided unprecedented prospects for LLMs in medical health consultations and patient education. However, it is important to emphasize that professional review and supervision remain essential in the current process of applying LLMs. Further evaluations and model modifications are required to enhance their effectiveness and strive for an even more ideal level of performance.

Declarations

Author Contributions: B.X and G.Z had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. H.S and Y. X contributed equally to this work and should be considered co–first authors.

Concept and design: H.S, G.Z, B.X.

Acquisition, analysis, or interpretation of data: H.S, Y.X, B.X, Y. S, G.Z, J. L

Drafting of the manuscript: Y.X, H.S, Z.L

Critical revision of the manuscript for important intellectual content: H.L, X.Z and J.L

Statistical analysis: H.S, T.L, G.Zh.

Obtained funding: H.S

Administrative, technical, or material support: Y Xia, X Zeng, and M Chen.

Supervision: J Li, B Xiao, G Zhang.

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

Research involving Human Participants and/or Animals: Not applicable.

Informed Consent: Not applicable.

Funding/Support: This work was supported by the Research Fund of the Tsinghua Changgung Hospital(grants12022C1001).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

1. Zeng G, Zhu W, Robertson WG, Penniston KL, Smith D, Pozdzik A, et al. International Alliance of Urolithiasis (IAU) guidelines on the metabolic evaluation and medical management of urolithiasis. Urolithiasis. 2022;51(1):4.
2. Geraghty RM, Davis NF, Tzelves L, Lombardo R, Yuan C, Thomas K, et al. Best Practice in Interventional Management of Urolithiasis: An Update from the European Association of Urology Guidelines Panel for Urolithiasis 2022. Eur Urol Focus. 2023;9(1):199–208.
3. Baatiah NY, Alhazmi RB, Albathi FA, Albogami EG, Mohammedkhalil AK, Alsaywid BS. Urolithiasis: Prevalence, risk factors, and public awareness regarding dietary and lifestyle habits in Jeddah, Saudi Arabia in 2017. Urol Ann. 2020;12(1):57–62.
4. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023.
5. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Netw Open. 2023;6(6):e2317517.
6. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int J Environ Res Public Health. 2023;20(4).
7. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. 2023;23(4):405–6.
8. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6).
9. Gortz M, Baumgartner K, Schmid T, Muschko M, Woessner P, Gerlach A, et al. An artificial intelligence-based chatbot for prostate cancer education: Design and patient evaluation study. Digit Health. 2023;9:20552076231173304.
10. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med. 2023;21(1):269.
11. Will ChatGPT transform healthcare? Nat Med. 2023;29(3):505–6.
12. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47(1):33.

13. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv. 2023.
14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
15. D'Amico RS, White TG, Shah HA, Langer DJ. I Asked a ChatGPT to Write an Editorial About How We Can Incorporate Chatbots Into Neurosurgical Research and Patient Care. Neurosurgery. 2023;92(4):663–4.
16. Mann DL. Artificial Intelligence Discusses the Role of Artificial Intelligence in Translational Medicine: A JACC: Basic to Translational Science Interview With ChatGPT. JACC Basic Transl Sci. 2023;8(2):221–3.
17. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv. 2023:2023.02.02.23285399.
18. The Lancet Digital H. ChatGPT: friend or foe? Lancet Digit Health. 2023;5(3):e102.
19. Marchandot B, Matsushita K, Carmona A, Trimaille A, Morel O. ChatGPT: the next frontier in academic writing for cardiologists or a pandora's box of ethical dilemmas. Eur Heart J Open. 2023;3(2):oead007.
20. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature. 2023;614(7947):214–6.
21. Lubowitz JH. ChatGPT, An Artificial Intelligence Chatbot, Is Impacting Medical Literature. Arthroscopy. 2023;39(5):1121–2.
22. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation. 2023;185:109729.
23. Anderson LM, Scrimshaw SC, Fullilove MT, Fielding JE, Normand J, Task Force on Community Preventive S. Culturally competent healthcare systems. A systematic review. Am J Prev Med. 2003;24(3 Suppl):68–79.

Figures

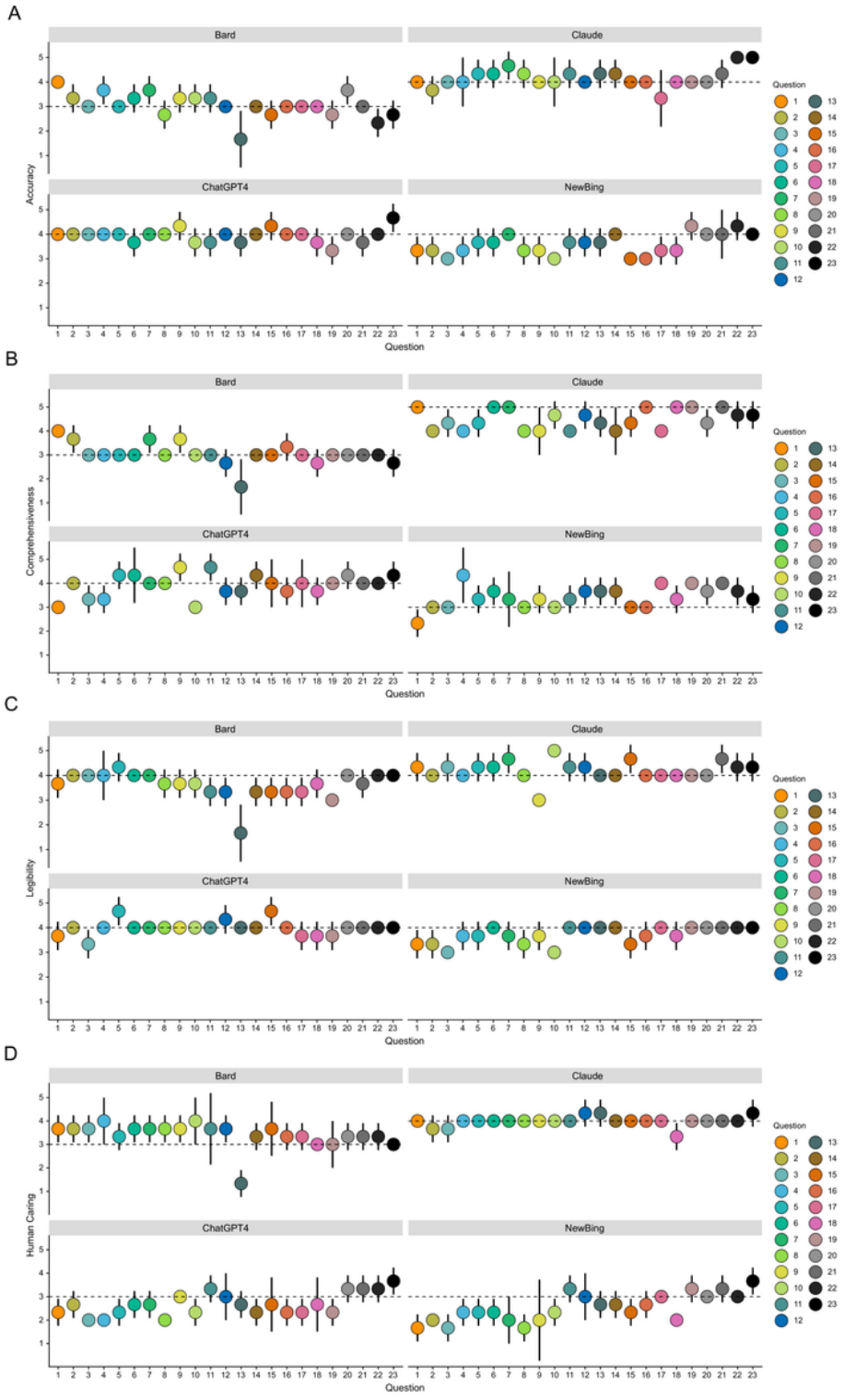

Figure 1

The performance of the four LLMs on various urolithiasis-related questions across different dimensions. A, the performance of accuracy. B. the performance of comprehensiveness. C, the performance of legibility. D, the performance of Human caring. Each color in the figure represents a different question, while the dashed line represents the median score. The vertical bar denotes the extremes observed in different tests.

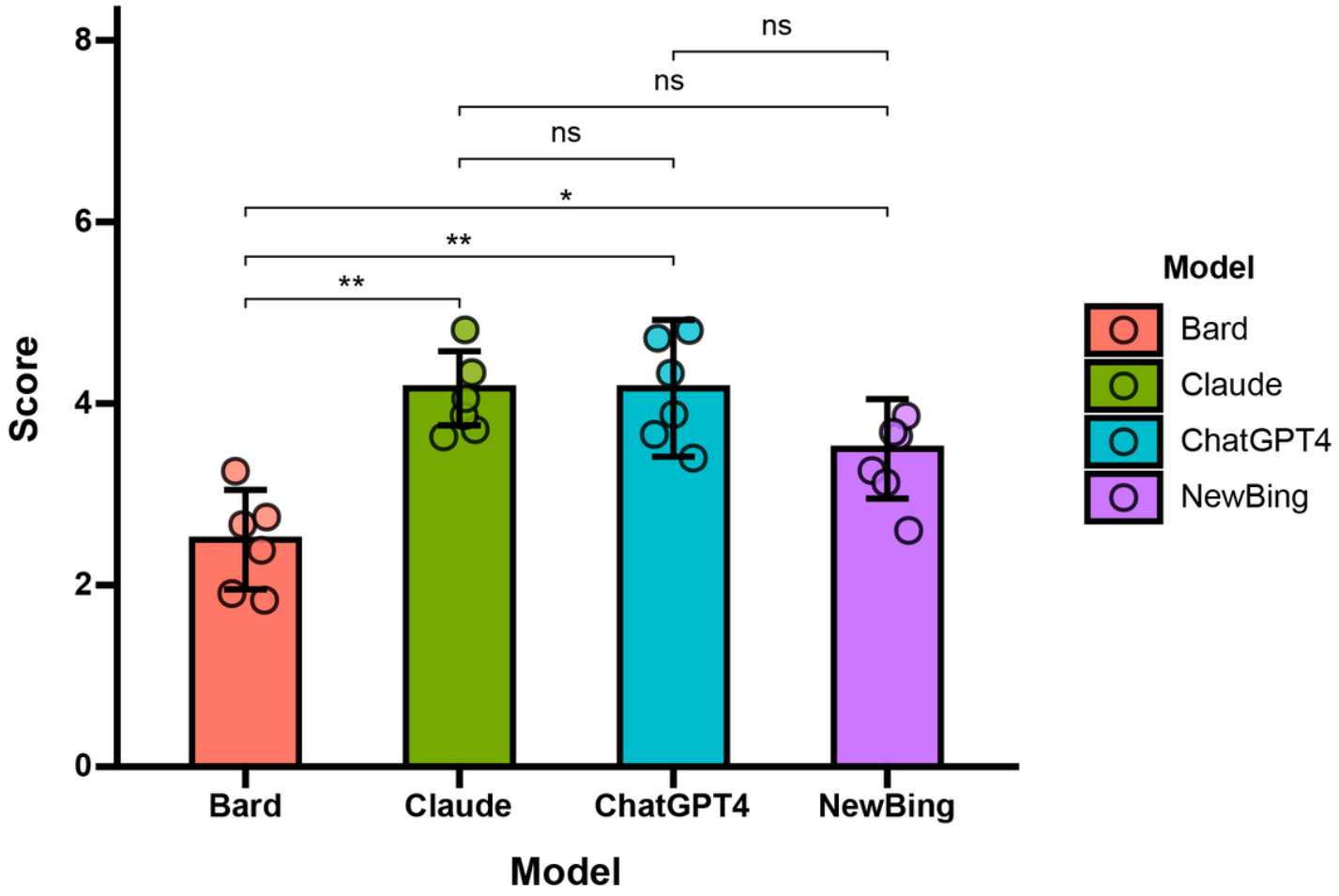

Figure 2

Comparison of the four LLMs regarding the performance in urolithiasis case analysis. The significance levels are indicated as follows: ns: not significant; *: p<0.05; **: p<0.01; ***: p<0.001.

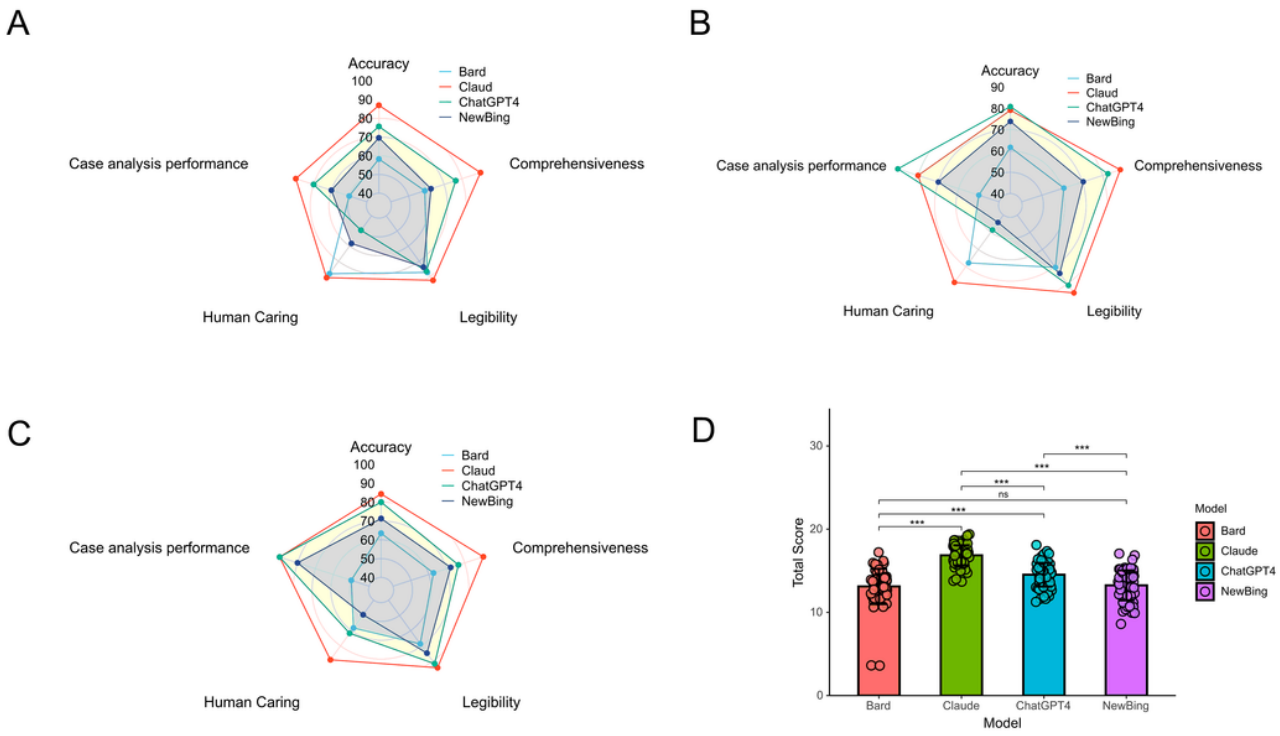

Figure 3

Comparison of overall performances for the models with respect to different dimensions. The score rates of each model for different dimensions were generated by dividing the total score of each model for each question by the full score. A, radar plot of score rates regarding different dimensions for the first test. B, radar plot of score rates regarding different dimensions for the second test. C, radar plot of score rates regarding different dimensions for the third test. D, the comparison of overall scores for different models on each question. Significance levels are indicated as: ns: not significant; *: p<0.05; **: p<0.01; ***: p<0.001.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- LLMsTestSummary.docx