Wang *et al.*
European Journal of Medical Research (2025) 30:45
https://doi.org/10.1186/s40001-025-02296-x

European Journal of Medical Research

RESEARCH

Open Access

Assessing the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in managing lumbar disc herniation

Suning Wang[1†], Ying Wang[3], Linlin Jiang[2,3], Yong Chang[2,3], Shiji zhang[2,3], Kun Zhao[1*], Lu Chen[1,2*] and Chunzheng Gao[1*]

Abstract

Purpose This study evaluated and compared the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in diagnosing and treating lumbar disc herniation (LDH) with radiculopathy.

Methods Twenty-one questions (across 5 categories) from NASS Clinical Guidelines were input into ChatGPT 4o and ChatGPT 4o mini. Five orthopedic surgeons assessed their responses using a 5-point Likert scale for accuracy and completeness, and a 7-point scale for reliability. Flesch Reading Ease scores were calculated to assess readability. Additionally, ChatGPT 4o analyzed lumbar images from 53 patients, comparing its recognizable agreement with orthopedic surgeons using Kappa values.

Results Both models demonstrated strong clinical support capabilities with no significant differences in accuracy or reliability. However, ChatGPT 4o provided more comprehensive and consistent responses. The Flesch Reading Ease scores for both models indicated that their generated content was "very difficult to read," potentially limiting patient accessibility. In evaluating lumbar disc herniation images, ChatGPT 4o achieved an overall accuracy of 0.81, with LDH recognition precision, recall, and F1 scores exceeding 0.80. The AUC was 0.80, and the Kappa value was 0.61, indicating moderate agreement between the model's predictions and actual diagnoses, though with room for improvement.

Conclusion While both models are effective, ChatGPT 4o offers more comprehensive clinical responses, making it more suitable for high-integrity medical tasks. However, the difficulty in reading AI-generated content and occasional use of misleading terms, such as "tumor," indicate a need for further improvements to reduce patient anxiety.

Keywords ChatGPT, Lumbar disc herniation, Clinical guidelines, Artificial intelligence, Spine

[†]Suning Wang is the independent first author of this article.

*Correspondence:
Kun Zhao
zhaokun0227@163.com
Lu Chen
793428124@qq.com
Chunzheng Gao
gaochunzheng1964@sina.com
[1] Department of Orthopedics, The Second Hospital of Shandong University, Qilu Hospital of Shandong University, Shandong University, Jinan 250000, China
[2] Department of Orthopedics, Qilu Hospital of Shandong University, The Second Hospital of Shandong University, Jinan 250012, China
[3] Shandong University, NO 44, Wenhuaxi Road, Jinan 250012, China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Low back pain (LBP) is a common condition, affecting approximately 80% of individuals during their life time span [1]. In the United States, healthcare costs for treating LBP exceed $100 billion annually [2]. Lumbar disc herniation (LDH) is one of the most common causes of LBP, most frequently affecting individuals aged 30 to 50, with a male-to-female ratio of approximately 2:1 [3]. It is also one of the most common causes of LBP. The primary symptoms of LDH include radicular pain, sensory disturbances, and weakness affecting one or more lumbosacral nerve roots [4, 5]. Managing LBP clinically requires multidisciplinary care and consideration of various prognostic factors.

The North American Spine Society (NASS) has issued an evidence-based clinical guideline on lumbar disc herniation with radiculopathy [6]. The guideline addresses a series of questions concerning the diagnosis and treatment of lumbar disc herniation with radiculopathy. Each question is answered by a panel of experts following a comprehensive review of the relevant literature, with expert recommendations included when necessary [6].

ChatGPT (Chat Generation Pre-Training Transformer, OpenAI) is an advanced artificial intelligence (AI) system that uses natural language processing (NLP) to understand text and simulate human-like responses [7]. It has demonstrated potential in offering clear answers to complex medical questions [8, 9]. ChatGPT has successfully passed Steps 1 and 2 of the U.S. Medical Licensing Examination (USMLE), achieving over 60% accuracy, the general passing standard [10]. OpenAI released ChatGPT 4o in May, and gave it some new capabilities to process audio and visual data compared to the previous ChatGPT 4.

ChatGPT has attracted interest from researchers and clinicians, who believe it can serve as an "online counseling" tool to help both clinicians and patients better understand diseases. In public health communication, ChatGPT demonstrates significant potential by providing accurate information on HIV-related topics and responding to oncology inquiries in a reliable manner consistent with oncology professionals' expertise [11, 12]. As for in medical education, medical learners can use their powerful natural language generation and comprehension to acquire methods and skills for clinical reasoning and decision making, but this learning style may deprive students of their innate opportunity for self-reflection [13, 14]. ChatGPT offers more accurate and comprehensive responses to open-ended questions than residents and specialists; however, there are limitations in its clinical case responses and in selecting additional tests and treatments [15, 16].

The purpose of this experiment is to compare and evaluate the clinical support capabilities of two AI models, ChatGPT 4o and ChatGPT 4o mini, in the context of lumbar disc herniation with radiculopathy, using questions from NASS Clinical Guidelines. The study will assess their performance in terms of accuracy, completeness, and reliability. Additionally, it will explore ChatGPT 4o's ability to recognize LDH in medical images. Ultimately, this experiment aims to provide empirical evidence for the application of AI in spine care and offer guidance for future optimization and improvement of AI in healthcare.

Methods

AI selection and question categorization

ChatGPT was selected for this study to enable direct comparison and scoring between ChatGPT 4o and ChatGPT 4o mini versions. Additionally, ChatGPT is publicly accessible and has demonstrated relevance in current medical literature, showing potential in supporting clinical workflows [17–19]. The input questions for OpenAI's ChatGPT were sourced from the 2012 NASS Clinical Guidelines for diagnosing and treating lumbar disc herniation with radiculopathy. These questions were developed by orthopedic and spine surgery specialists in the fields of orthopedics and neurosurgery and address the natural history, diagnosis, and treatment of lumbar disc herniation [20]. We qualitatively classified the clinical guidelines into five categories: Group 1: Definition and History, Group 2: Diagnosis, Group 3: Non-Surgical Interventions, Group 4: Surgical Interventions, and Group 5: Prognosis. A total of 21 questions were retained, and the screening process is illustrated in Supplement Fig. 1.

Questions input and assessment

The 21 guiding questions were used as input for OpenAI's ChatGPT software. To ensure consistency, a single investigator separately input all questions into the ChatGPT 4o and ChatGPT 4o mini versions. Each ChatGPT response was evaluated by five independent orthopedic surgeons with at least three years of experience. The complete set of questions and answers is available in the Supplementary Materials.

Recognition of images

We randomly selected 53 patient MRIs from the inpatient case database of the Second Hospital of Shandong University, and divided them into two groups based on the primary diagnosis: lumbar disc herniation (LDH, $n=31$) and non-LDH (N-LDH, $n=22$). Two independent orthopedic surgeons evaluated each patient's MRI, selecting the image with the most severe lesion, which

was saved in PNG format. If discrepancies arose, a third physician resolved them. The images were then input into ChatGPT 4o to generate responses.

Data safety

To safeguard the confidentiality and integrity of patient data used in our study, we implemented several comprehensive measures. We established a dedicated in-network database using a high-performance NAS system, ensuring that only authorized users can access sensitive information through role-based access management. Detailed logs of data access and modifications are maintained and regularly audited to detect unauthorized access. Data de-identification techniques were employed, removing sensitive details to ensure anonymity (SupplementFig. 2). Additionally, robust network security measures, including firewalls and intrusion detection systems, protect against external threats. All data are encrypted during storage and transmission to prevent unauthorized interpretation. Regular data backups ensure rapid recovery in case of system failures, and a monitoring mechanism is in place to promptly detect and respond to potential security incidents. These measures collectively uphold ethical standards and protect patient privacy throughout the study.

Evaluation metrics and statistical analysis

A 5-point Likert scale was used to assess the accuracy and completeness of ChatGPT responses. A 7-point Likert scale was used to assess reliability. Flesch Reading Ease scores and Flesch–Kincaid reading levels were calculated for both NASS Clinical Guidelines and ChatGPT responses to evaluate readability. Higher Flesch Reading Ease scores indicate better readability, while lower Flesch–Kincaid levels reflect easier reading. The Flesch–Kincaid reading level provides a grade-level score that indicates the school grade necessary to understand the text. This metric is useful for determining whether the language and complexity of the text are appropriate for the intended audience. SPSS version 27 was used for statistical analysis. All samples were tested for normality using the Shapiro–Wilk tests, proving that the samples were non-normally distributed (Supplementary Table 2). The Mann–Whitney *U* test was used to compare the two models. The Kruskal–Wallis test was used to compare different groups within the same model. The *P* value of <0.05 was deemed statistically significant. Python 3.10.11 was used for Kappa statistics and ROC curve analysis.

The grading criteria are described in detail below:

Accuracy:

1. Completely incorrect
2. More incorrect than correct [>75% incorrect]
3. Approximately equal correct and incorrect
4. More correct than incorrect [>75% correct]
5. Completely correct

Completeness:

1. Very incomplete [0–25%]
2. Incomplete [25–50%]
3. Moderate [50–75%]
4. Complete [>75%]
5. Very complete [100%]

Reliability:

1. Totally insecure: None of the information provided could be verified from medical sources or contained inaccurate and incomplete information.
2. Very insecure: Most of the information provided is not verifiable from medical sources or is partially correct, but contains significant inaccurate or incomplete information.
3. Relatively reliable: Most of the information provided is verified from medical scientific sources, but contains some important incorrect or incomplete information.
4. Reliable: Most of the information provided has been verified by medical-scientific sources, but there is some inaccurate or incomplete information.
5. Relatively very reliable: Most of the information provided has been verified by medical-scientific sources, with few inaccuracies or incomplete information.
6. Very secure: Most of the information provided has been verified by medical-scientific sources and there is little inaccurate or incomplete information.
7. Absolutely secure: All information provided has been verified by medical scientific sources and there is no inaccurate or incomplete information or missing information.

Results

Comparison of ChatGPT 4o and ChatGPT 4o mini

We input 21 questions from the NASS Clinical Guidelines on the diagnosis and treatment of lumbar disc herniation with radiculopathy into ChatGPT 4o mini and ChatGPT 4o, comparing their accuracy, completeness, and reliability (Fig. 1, Table 1). A comprehensive list of the NASS guidelines and the corresponding responses from both ChatGPT versions were documented (Supplementary Table 1).

Using a 5-point Likert scale, ChatGPT 4o mini had a mean accuracy rating of 4.63, while ChatGPT 4o

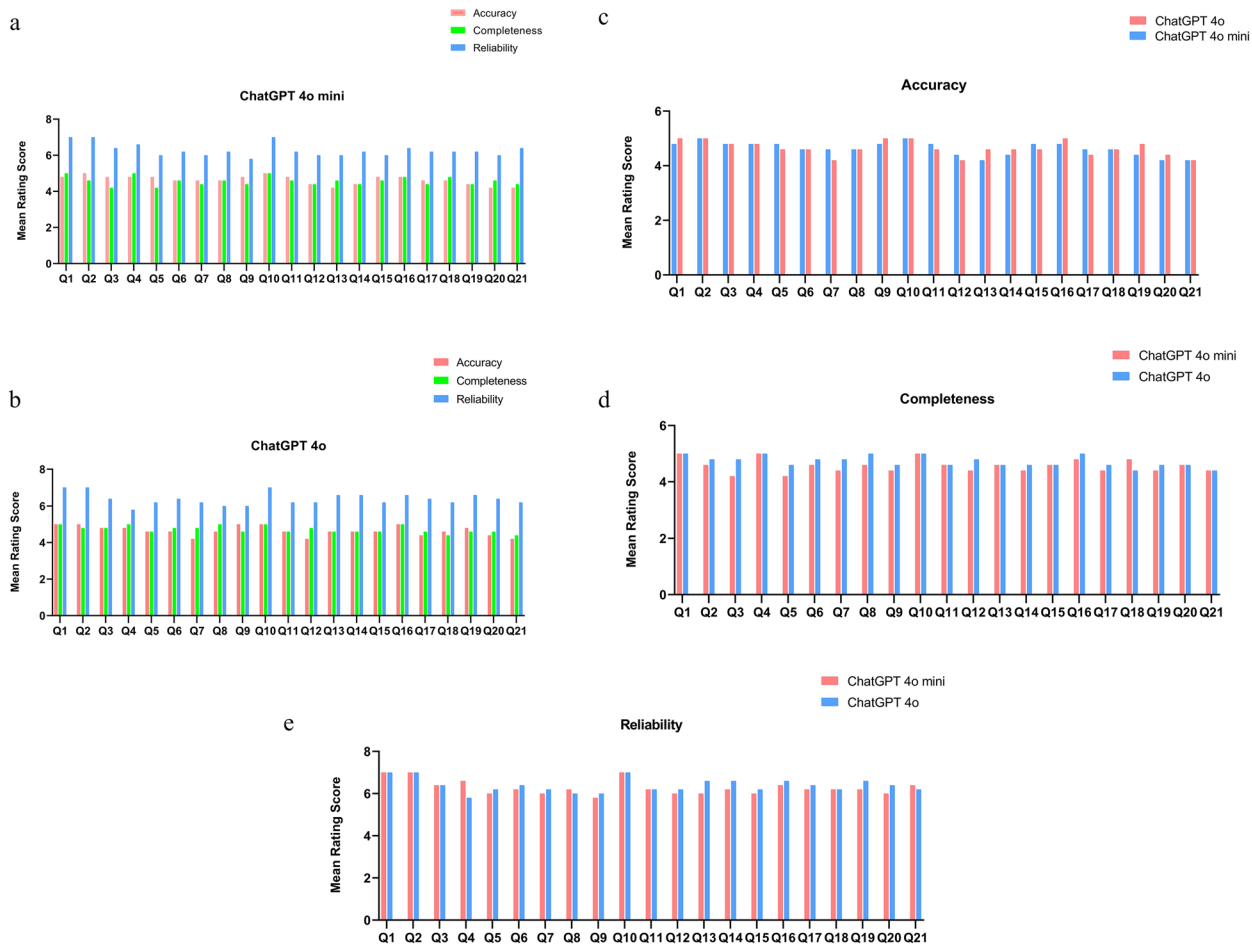

Fig. 1 The accuracy, completeness and reliability of ChatGPT 4o and ChatGPT 4o mini

Table 1 ChatGPT 4o mini vs. ChatGPT 4o

	ChatGPT 4o min	ChatGPT 4o	*P*
Mean accuracy	4.63	4.65	0.77
Median accuracy [IQR]	5 [1]	5 [1]	–
Mean completeness	4.57	4.72	0.04
Median completeness [IQR]	5 [1]	5 [1]	–
Mean reliability	6.29	6.43	0.11
Median reliability [IQR]	6 [1]	6 [1]	–

IQR interquartile range

scored 4.65, with both models exceeding 75% accuracy. Despite ChatGPT 4o's slightly higher mean score, the *P* value of 0.77 indicated no statistically significant difference (Table 1). The completeness score for ChatGPT 4o mini was 4.57, while ChatGPT 4o achieved 4.72, with a significant difference ($P=0.04$) favoring ChatGPT 4o (Fig. 1d). Reliability ratings were also similar (7-point Likert scale), with ChatGPT 4o mini at 6.29 and ChatGPT 4o at 6.43, with no significant difference ($P=0.11$).

Both models achieved a median accuracy and median completeness of 5, with an IQR of 1, indicating that most ratings were concentrated around a score of 5, demonstrating consistency in performance for these metrics. For reliability, both models had a median of 6 and an IQR of 1, suggesting that ratings in this category were also stable and closely clustered around 6. The similarity in median and IQR between ChatGPT 4o mini and ChatGPT 4o further indicates that the overall performance distribution for both models is comparable, with limited variability, ensuring stability in their performance (Table 1).

Intergroup differences in the two models

We categorized the 21 questions into five groups based on content: Group 1 (Definition and History), Group 2 (Diagnosis), Group 3 (Non-Surgical Interventions),

Group 4 (Surgical Interventions), and Group 5 (Prognosis) (Supplement Fig. 1, Supplementary Table 1).

Group 1 had the highest mean scores for accuracy (4.90), completeness (4.80), and reliability (7.00) among the five groups. Group 5 had the lowest mean accuracy (4.35). Group 3 had the lowest scores for completeness (4.44) and reliability (6.04). Among the five groups, ChatGPT 4o mini responses showed no significant difference in completeness ($P > 0.05$). However, in terms of accuracy and reliability, there was a statistically significant difference between groups ($P < 0.05$, Table 2).

For accuracy and completeness, both models generally had a median score of 5 across all groups, with IQRs mostly between 0 and 1. This indicates that ratings for these two metrics were tightly clustered around the median, demonstrating consistent performance and minimal variability across groups. In terms of reliability, the median scores were typically 6 or 7, but with some variation in the IQR (ranging from 0 to 2) across different groups. This variability in IQR for reliability suggests that there was greater fluctuation in the reliability ratings within certain groups, pointing to potential group-specific factors affecting the models' reliability consistency (Table 2).

In the ChatGPT 4o model, Group 1 had the highest mean scores for accuracy (5.00), completeness (4.90), and reliability (7.00) among the five groups. Group 5 had the lowest scores in both accuracy (4.50) and completeness (4.50), but the differences between groups were not statistically significant ($P > 0.05$). Group 3 had the lowest mean reliability score (6.16), which was statistically significant ($P < 0.05$, Table 2).

Readability test

ChatGPT 4o mini had a Flesch Reading Ease score of 19.72, corresponding to a Flesch–Kincaid Grade Level described as "very difficult to read". ChatGPT 4o had a similar Flesch Reading Ease score of 17.41, also rated as "very difficult to read". The required education level for both models was a college graduate. However, the NASS Clinical Guidelines showed readability at the "Professional" education level, with a Flesch Reading Ease score of 5.89 (Table 3).

Recognition of lumbar disc herniation

ChatGPT 4o's precision, recall, and F1 scores for N-LDH classification were 0.80, 0.73, and 0.76, respectively. For LDH identification, the precision, recall, and F1 scores

Table 2 Comparison between different groups in two models

	Group 1 ($n=2$)	Group 2 ($n=2$)	Group 3 ($n=5$)	Group 4 ($n=8$)	Group 5 ($n=4$)	*P*
ChatGPT 4o mini						
Mean accuracy	4.90	4.80	4.68	4.63	4.35	0.02
Median accuracy [IQR]	5 [0]	5 [0]	5 [1]	5 [1]	4 [1]	–
Mean completeness	4.80	4.60	4.44	4.60	4.55	0.48
Median completeness [IQR]	5 [0]	5 [1]	4 [1]	5 [1]	5 [1]	–
Mean reliability	7.00	6.50	6.04	6.25	6.20	< 0.01
Median reliability [IQR]	7 [0]	6.5 [1]	6 [2]	6 [1]	6 [1]	–
ChatGPT 4o						
Mean accuracy	5.00	4.80	4.6	4.63	4.50	0.07
Median accuracy [IQR]	5 [0]	5 [0]	5 [1]	5 [1]	4.5 [1]	–
Mean completeness	4.90	4.90	4.76	4.73	4.50	0.08
Median completeness [IQR]	5 [0]	5 [0]	5 [0]	5 [1]	4.5 [1]	–
Mean reliability	7.00	6.50	6.16	6.48	6.35	< 0.01
Median reliability [IQR]	7 [0]	6.5 [1]	6 [0]	7 [1]	6 [1]	–

IQR interquartile range

Table 3 Flesch reading ease scores of the NASS clinical guidelines and the responses from ChatGPT 4o min and ChatGPT 4o to NASS questions

	Reading Score	Flesch–Kincaid Grade Level	Education level required
NASS guideline	5.89	Extremely difficult to read	Professional
ChatGPT-4o min	19.72	Very difficult to read	College graduate
ChatGPT-4o	17.41	Very difficult to read	College graduate

* Flesch Reading Ease scores were utilized to evaluate interpretability and accessibility to the public

were 0.82, 0.87, and 0.84. The F1 score results further indicate that the model's overall performance was strong in the LDH category. The model's overall accuracy was 0.81, sensitivity was 0.87, and specificity was 0.73. (Table 4). The area under the ROC curve (AUC) value of the model's ROC curve was 0.80, indicating good performance in distinguishing between LDH and N-LDH. The Kappa value of 0.61 demonstrated a moderate level of agreement with physicians (Fig. 2).

Discussion

In this study, we confirm that artificial intelligence platforms (in this case ChatGPT 4o mini, ChatGPT 4o), show potential for providing accurate, comprehensive, and reliable medical information in the field of LDH, with the possibility of even replacing doctors in the future.

We first analyzed the mean accuracy, completeness, and reliability of ChatGPT 4o and ChatGPT 4o mini's responses to 21 questions from the NASS Clinical Guidelines. The mean accuracy and completeness of all responses exceeded 4 in both models (Fig. 1). For reliability, the mean scores for all questions were greater than 6 (Fig. 1). This indicates that for LDH-related questions, the AI platform responses were highly accurate, complete, and reliability. Although ChatGPT 4o and ChatGPT 4o mini did not differ significantly in mean accuracy and reliability, ChatGPT 4o demonstrated a slight advantage in completeness (Table 1). This suggests that ChatGPT 4o may be preferable in scenarios where information integrity is crucial, such as detailed patient counseling or educational materials.

It is worth noting that in previous studies, the authors chose to use NASS answers as criteria to evaluate the AI platform responses for accuracy, over-conclusiveness, supplementary, and incompleteness [20, 21]. These results illustrate the differences between AI responses and NASS guideline answers but overlook certain issues. First, the NASS Clinical Guidelines have been published for over a decade, and many new technologies and methods currently used in clinical practice are not covered by the guidelines. Second, the commenters may not be specialized orthopedic surgeons, which may not accurately reflect true clinical scenarios. Finally, the commenters knew that the answers were AI-generated, raising the possibility of bias against the AI responses. To better reflect real clinical scenarios and clinicians' attitudes, we selected five specialized orthopedic surgeons for scoring, without informing them that the answers were AI generated. This approach aimed to reflect actual clinical situations and provide a more objective evaluation of the AI platform more closely.

In this study, we divided the questions into five categories (Supplement Fig. 1). When analyzing responses from different groups, both models demonstrated variations in performance. Group 1, which covered the definition and history of the disease, had the highest scores

Table 4 Performance of ChatGPT 4o in recognizing LDH

	LDH ($n=31$)	N-LDH ($n=22$)
Correct identification	n = 27	$n=16$
Precision	0.82	0.80
Recall	0.87	0.73
F1 score	0.84	0.76
Accuracy	0.81	
Sensitivity	0.87	
Specificity	0.73	

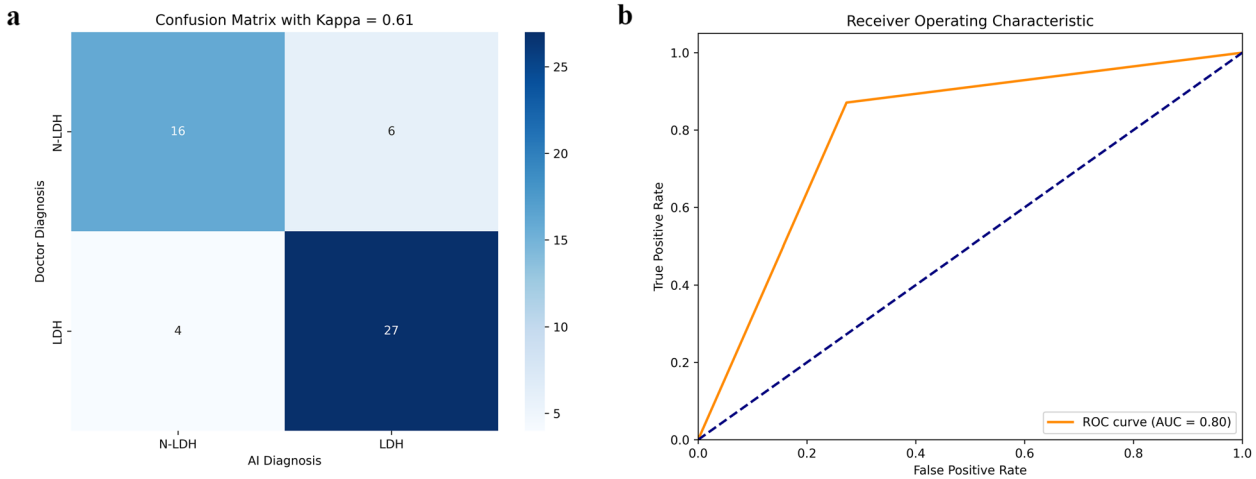

Fig. 2 Confusion Matrix with Kappa Score and ROC Curve Evaluation

in accuracy, completeness, and reliability among the five groups (Table 2). The findings of Ankur Kayastha et al. are consistent with our results [20]. In general, the natural history of lumbar disc herniation with radiculopathy is well studied and relatively basic, suggesting that the AI performs well in delivering foundational knowledge [5]. In contrast, Group 5 (prognosis) typically received the lowest scores, particularly for ChatGPT 4o mini. In clinical practice, prognosis-related questions are often more challenging for doctors to answer. In the clinic, these may also be more difficult questions for doctors to answer. A statistically significant difference ($P<0.05$) in mean accuracy and reliability was observed between groups for ChatGPT 4o mini, while ChatGPT 4o showed significant differences in reliability only, suggesting that the 4.o model is somewhat more stable (Table 2). The between-group differences in reliability may be due to the fact that orthopedic surgeons tend to be more cautious when assessing the treatment and prognosis of lumbar disc herniation.

The median values for both models across accuracy and completeness metrics are consistently high (typically 5) in all groups, with relatively narrow IQRs (0 to 1). This indicates that the performance for these two metrics remains stable and is not influenced by outliers or extreme values. The small IQRs suggest that most ratings fall close to the median, demonstrating consistent outputs across groups in Table 2. This consistency is particularly valuable, as it implies that both ChatGPT 4o mini and ChatGPT 4o provide reliable accuracy and completeness. However, reliability shows a different trend. While the median reliability is generally 6 or 7 across groups for both models, the IQR varies of ChatGPT 4o mini slightly more (0 to 2), indicating some variability in this metric across groups.

The readability of the output from both models was assessed using the Flesch Reading Ease score. The Flesch Reading Ease scores were used to assess interpretability and public accessibility, with higher scores indicating easier readability and comprehension [22, 23]. Both models were rated as “very difficult to read,” with scores of 19.72 for ChatGPT 4o mini and 17.41 for ChatGPT 4o (Table 3). Although both models are below the "professional" readability level of the NASS guidelines, they are still quite difficult to read, equivalent to the reading level of a college graduate. This finding highlights a potential barrier for the general public, particularly for individuals without a medical background or with lower levels of education. Improving the readability of AI responses could enhance their usability, particularly in scenarios targeting less-educated patients, thereby benefiting broader public health communication and disease prevention.

Additionally, ChatGPT 4o was equipped with the ability to process audio and visual data, a feature not present in the previous version 4.0[24]. In previous studies, ChatGPT’s performance in the medical field has predominantly focused on textual data [25]. Although ChatGPT is not designed to diagnose diseases, we were curious about ChatGPT 4o’s ability to recognize diseases in images. To explore this, we randomly selected 53 patients and input the image with the most severe lesion site into ChatGPT 4o. ChatGPT 4o performed well in identifying and classifying LDH versus N-LDH, with precision, recall, and F1 scores all above 0.80 for LDH (Table 4). The discrepancy between precision and recall resulted in an F1 score of 0.76 for N-LDH, suggesting that while the model performed reasonably well, there is significant room for improvement, particularly in correctly identifying more true N-LDH cases. The model’s sensitivity for LDH was 0.87, higher than its specificity for N-LDH (0.73), indicating that while the model is effective at identifying LDH cases, it is less reliable at ruling out N-LDH cases. The model’s overall accuracy was 0.81, indicating that 81% of the predictions were correct. While this accuracy is acceptable, it highlights that nearly 20% of the predictions were incorrect, suggesting that the model’s decision-making process could benefit from further refinement. This imbalance could result in a higher rate of false positives in practical applications, where accurately identifying non-cases is just as important as identifying true cases.

The Kappa value of 0.61 indicates moderate agreement between the model’s predictions and the actual diagnoses, suggesting that while the predictions align with the ground truth, they are not highly reliable (Fig. 2). In a clinical setting, this moderate level of agreement may require further validation or the use of complementary diagnostic tools to ensure patient reliability and diagnostic accuracy. Additionally, the AUC was 0.80, indicating the model had a good ability to distinguish between LDH and N-LDH cases (Fig. 2). An AUC of 0.80 is generally considered to indicate good discriminatory ability, though it is not exceptional. This suggests that while the model is effective, there is still room for improvement, particularly in reducing false positives and enhancing recall for N-LDH cases.

One of ChatGPT’s strengths is its ability to process large amounts of information and generate responses in a conversational, easy-to-understand format. ChatGPT’s content is updated much more frequently than hospital patient information leaflets and other traditional sources, as shown by Johnson et al. [26]. Additionally, an increasing number of patients are searching for their conditions online, which can be misleading and exacerbate anxiety due to the presence of irrelevant

or inaccurate information. In two cases of lumbar disc herniation, ChatGPT 4o mentioned the word "tumor" in the responses. Although ChatGPT 4o mentioned "tumor" only as a possibility, this can still increase anxiety and fear, especially for patients with low levels of education or no medical background.

This study has several limitations. First, the questions were based on NASS guidelines and may not fully reflect typical outpatient scenarios, though they allow for an assessment of ChatGPT's recommendations for lumbar disc herniation with radiculopathy. Second, orthopedists' evaluations of ChatGPT's responses are subjective and may differ from the evidence-based NASS guidelines, despite generally aligning with spine surgeons' opinions. Third, this study only examines lumbar disc degeneration using ChatGPT 4o mini and ChatGPT 4o, leaving uncertainty about other models' performance for different conditions. Lastly, the MRI image provided to ChatGPT 4o showed the most prominent lesion, but patients may struggle to understand such images without professional guidance. This limitation may affect a patient's ability to use AI for self-assessment.

Conclusion

With the rapid growth of the Internet and the vast availability of accessible medical information, more patients are taking an increasingly active role in managing their healthcare. This study demonstrates that both ChatGPT 4o and ChatGPT 4o mini exhibit strong clinical service capabilities. While the difference in accuracy does not significantly diminish the utility of ChatGPT 4o mini, ChatGPT 4o generally provides more complete and comprehensive answers. For questions requiring a higher level of completeness and security, ChatGPT 4o is the preferred choice. Although ChatGPT 4o is effective in identifying lumbar disc herniation in images, its diagnoses may occasionally increase patient anxiety.

Abbreviations

LBP	Low back pain
LDH	Lumbar disc herniation
NASS	The North American Spine Society
ChatGPT	Chat generation pre-training transformer
USMLE	Medical licensing examination

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40001-025-02296-x.

Supplementary material 1

Supplementary material 2

Supplementary material 3

Supplementary material 4

Acknowledgements

We are grateful to all the doctors who participated in the study. And we also thank the study's investigators.

Author contributions

Suning Wang: first author, analysed the data, wrote the first draft of the manuscript, and revised it. Ying Wang: assist in data analysis and contributed to data curation. Linlin Jiang: assist in data analysis. Yong Chang: contributed to software operation. Shiji Zhang: assist in data analysis. All authors have read and approved the final manuscript. Kun Zhao: contributed to article modification. Lu Chen: contributed to software operation and assist in data analysis. Chunzheng Gao: corresponding author, agreed to be accountable for all aspects of the work, thereby ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and also resolved the final approval of the version to be published.

Funding

Not applicable.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. For human experiments, the trial was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by our hospital's ethics committee (KYLL2024762).

Competing interests

The authors declare no competing interests.

Received: 20 September 2024 Accepted: 13 January 2025
Published online: 22 January 2025

References

1. Andersson GBJ. Epidemiological features of chronic low-back pain. Lancet. 1999;354(9178):581–5.
2. Martin BI. Expenditures and health status among adults with back and neck problems. JAMA. 2008;299(6):656. https://doi.org/10.1001/jama.299.6.656.
3. Pojskic M, Bisson E, Oertel J, Takami T, Zygourakis C, Costa F. Lumbar disc herniation: epidemiology, clinical and radiologic diagnosis WFNS spine committee recommendations. World Neurosurg X. 2024;22:100279.
4. Vroomen PCAJ. Diagnostic value of history and physical examination in patients suspected of lumbosacral nerve root compression. J Neurol Neurosurg Psychiatry. 2002;72(5):630–4. https://doi.org/10.1136/jnnp.72.5.630.
5. Zhang AS, Xu A, Ansari K, Hardacker K, Anderson G, Alsoof D, et al. Lumbar disc herniation: diagnosis and management. Am J Med. 2023;136(7):645–51.
6. Kreiner DS, Hwang SW, Easa JE, Resnick DK, Baisden JL, Bess S, et al. An evidence-based clinical guideline for the diagnosis and treatment of lumbar disc herniation with radiculopathy. Spine J. 2014;14(1):180–91.
7. Unveiling the Cognitive Capacity of ChatGPT. Assessing its human-like reasoning abilities. Int Res J Mod Eng Technol Sci. 2024;10:15.
8. Waisberg E, Ong J, Masalkhi M, Kamran SA, Zaman N, Sarker P, et al. GPT-4: a new era of artificial intelligence in medicine. Ir J Med Sci. 1971;192(6):3197–200. https://doi.org/10.1007/s11845-023-03377-8.
9. Lee P. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. 2023. N Engl J Med. https://doi.org/10.1056/NEJMsr2214184.
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States medical licensing

examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:45312.
11. De Vito A, Colpani A, Moi G, Babudieri S, Calcagno A, Calvino V, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav. 2024;28(8):2746–54. https://doi.org/10.1007/s10461-024-04391-2.
12. Cè M, Chiarpenello V, Bubba A, Felisaz PF, Oliva G, Irmici G, et al. Exploring the role of ChatGPT in oncology: providing information and support for cancer patients. BioMedInformatics. 2024;4(2):877–88.
13. Van De Ridder JMM, Shoja MM, Rajput V. Finding the place of ChatGPT in medical education. Acad Med. 2023;98(8):867–867. https://doi.org/10.1097/ACM.0000000000005254.
14. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for medical education: exploring its impact on doctors and medical students. JMIR Med Educ. 2024;10:e52483.
15. Lechien JR, Naunheim MR, Maniaci A, Radulesco T, Saibene AM, Chiesa-Estomba CM, et al. Performance and consistency of Chatgpt-4 versus otolaryngologists: a clinical case series. Otolaryngol-Head Neck Surg Off J Am Acad Otolaryngol-Head Neck Surg. 2024;170(6):1519–26.
16. De Vito A, Geremia N, Marino A, Bavaro DF, Caruana G, Meschiari M, et al. Assessing ChatGPT's theoretical knowledge and prescriptive accuracy in bacterial infections: a comparative study with infectious diseases residents and specialists. Infection. 2024. https://doi.org/10.1007/s15010-024-02350-6.
17. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño 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. https://doi.org/10.1371/journal.pdig.0000198.
18. Duey AH, Nietsch KS, Zaidat B, Ren R, Ndjonko LCM, Shrestha N, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. Spine J. 2023;23(11):1684–91.
19. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT Throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2024;25:48659.
20. Kayastha A, Lakshmanan K, Valentine MJ, Nguyen A, Dholakia K, Wang D. Lumbar disc herniation with radiculopathy: a comparison of NASS guidelines and ChatGPT. North Am Spine Soc J NASSJ. 2024;19:100333.
21. Mejia MR, Arroyave JS, Saturno M, Ndjonko LCM, Zaidat B, Rajjoub R, et al. Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a north american spine society guideline comparison. Neurospine. 2024;21(1):149–58. https://doi.org/10.14245/ns.2347052.526.
22. Bald A, Richardson H, Al Samaraee A, Fasih T. Quality and readability of online information and materials on post-surgery breast seroma. Br J Hosp Med. 2024;85(6):1–9. https://doi.org/10.12968/hmed.2024.0058.
23. Michel C, Dijanic C, Abdelmalek G, Sudah S, Kerrigan D, Gorgy G, et al. Readability assessment of patient educational materials for pediatric spinal conditions from top academic orthopedic institutions. J Child Orthop. 2023;17(3):284–90. https://doi.org/10.1177/18632521231156435.
24. Holmlund M, Hagelbäck J, Lundström O. Bachelor's degree Project.
25. Shanahan M. Talking About Large Language Models. arXiv; 2023. http://arxiv.org/abs/2212.03551
26. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT Model. 2023. https://www.researchsquare.com/article/rs-2566942/v1

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.