Check for updates

Journal of Orthopaedic Surgery

Journal of Orthopaedic Surgery
33(3) 1–11
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10225536251407382
journals.sagepub.com/home/osj
Sage

Generative AI in degenerative lumbar spinal stenosis care: A NASS guideline-compliant comparative analysis of ChatGPT and DeepSeek

Meng Zhang[1,*], Jiameng Li[1,*], Yaluo Zhou[1], Zhiwu Chen[1], Pan Wang[1], Bin Hu[1] and Zhong Xiang[1]

Abstract

Background: This study aims to compare the performance of two artificial intelligence (AI) models, ChatGPT-4.0 and DeepSeek-R1, in addressing clinical questions related to degenerative lumbar spinal stenosis (DLSS) using the North American Spine Society (NASS) guidelines as the benchmark.

Methods: 15 clinical questions spanning five domains (diagnostic criteria, non-surgical management, surgical indications, perioperative care, and emerging controversies) were designed based on the 2013 NASS evidence-based clinical guidelines for the diagnosis and management of DLSS. Responses from both models were independently evaluated by two board-certified spine surgeons across four metrics: accuracy, completeness, supplementality, and misinformation. Inter-rater reliability was assessed using Cohen's κ coefficient, while Mann-Whitney U and Chi-square tests were employed to analyze statistical differences between models.

Results: DeepSeek-R1 demonstrated superior performance over ChatGPT-4.0 in accuracy (median score: 3 vs 2, P = 0.009), completeness (2 vs 1, P = 0.010), and supplementality (2 vs 1, P = 0.018). Both models exhibited comparable performance in avoiding misinformation (P = 0.671). DeepSeek-R1 achieved higher inter-rater agreement in accuracy (κ = 0.727 vs 0.615), whereas ChatGPT-4.0 showed stronger consistency in ssupplementality (κ = 0.792 vs 0.762).

Conclusions: While both AI models demonstrate potential for clinical decision support, DeepSeek-R1 aligns more closely with NASS guidelines. ChatGPT-4.0 excels in providing supplementary insights but exhibits variability in accuracy. These findings underscore the need for domain-specific optimization of AI models to enhance reliability in medical applications.

Keywords

degenerative lumbar spinal stenosis, artificial intelligence, clinical guidelines, NASS, ChatGPT, DeepSeek

Received: 14 May 2025; revised: 25 November 2025; accepted: 30 November 2025

Introduction

Degenerative lumbar spinal stenosis (DLSS), an age-related degenerative disorder, has exhibited a marked rise in global prevalence, particularly among individuals aged 65 years and older.[1–3] The escalating burden of DLSS-associated neurogenic claudication, chronic pain, and functional impairment, compounded by population aging, poses significant public health challenges, with annual direct healthcare costs exceeding billions of dollars in the United States alone.[4]

[1]Orthopedics Third Inpatient Ward, The Fourth Hospital of Changsha (Integrated Traditional Chinese and Western Medicine Hospital of Changsha), Changsha, China

*Meng Zhang and Jiameng Li are regarded as co-first author.

Corresponding author:
Zhong Xiang, Orthopedics Third Inpatient Ward, The Fourth Hospital of Changsha (Integrated Traditional Chinese and Western Medicine Hospital of Changsha), No. 70, Lushan Road, Yuelu District, Changsha 410006, China.
Email: spinezxy@126.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Pathologically, DLSS arises from multifactorial interactions, including intervertebral disc degeneration, ligamentum flavum hypertrophy, and facet joint hyperplasia, culminating in reduced spinal canal volume and neural compression.[5] While conservative therapies such as physical therapy and epidural steroid injections (ESIs) alleviate early symptoms, approximately 30% of patients require surgical intervention due to progressive neurological deficits.[6]

The 2013 evidence-based guidelines for the diagnosis and treatment of DLSS by the North American Spine Society (NASS) remain the cornerstone of clinical decision-making, employing a graded recommendation system (Levels A/B/C/I) to standardize care pathways.[7] For instance, the guidelines classify "progressive neurological deficits" as an absolute surgical indication (Level A evidence), whereas "selective nerve root blocks" are relegated to diagnostic adjuncts (Level C evidence). Nevertheless, there exists a contradiction between the static nature of the guidelines and the dynamic demands of clinical practice. On one hand, the 2013 edition of the guidelines has not incorporated recent advancements in minimally invasive techniques (such as endoscopic unilateral approach for bilateral decompression) and emerging controversies (such as the long-term efficacy of dynamic stabilization systems) that have emerged in recent years. On the other hand, physicians' adherence variability remains problematic in environments with uneven resource distribution, with a multinational survey revealing that only 58% of spine surgeons consistently comply with NASS surgical recommendations.[8]

The integration of artificial intelligence (AI) in the medical field offers a potential solution to the aforementioned challenges. In 2023, a review published in the Journal of the American Medical Association (JAMA) highlighted AI's capacity to synthesize vast medical literature in real time through natural language processing (NLP), enabling clinicians to access updated evidence efficiently.[9] Large language models (LLMs), represented by ChatGPT and DeepSeek, have demonstrated promise in scenarios such as patient education and clinical note summarization, yet their reliability in evidence-based guideline interpretation remains contentious.[10,11] For example, Kung et al. found that ChatGPT's performance on the United States Medical Licensing Examination (USMLE) approached the passing threshold, but its error rate reached as high as 35% in questions requiring complex clinical reasoning.[12] Similarly, Gilson et al. demonstrated that the model achieved a score equivalent to the passing level of third-year medical students.[13]

In spine surgery, AI applications have predominantly focused on image analysis (e.g., automated measurement of spinal canal sagittal diameter) and surgical planning (e.g., neural decompression pathway simulation),[14] leaving the potential of LLMs in guideline-based text interactions underexplored. DLSS management further complicates AI implementation due to its reliance on synthesizing multimodal data, including symptomatology, imaging parameters, comorbidities, and psychosocial factors, which demands advanced cross-modal comprehension and logical reasoning. Additionally, the presence of NASS Level I recommendations (expert consensus) necessitates AI models capable of navigating clinical uncertainty alongside established evidence.

This study represents the first systematic evaluation of ChatGPT-4.0 and DeepSeek-R1 in addressing DLSS-related clinical queries using NASS guidelines as the benchmark. By comparing performance across accuracy, completeness, supplementality, and misinformation, we aim to elucidate strengths and limitations of domain-specific (DeepSeek-R1) versus general-purpose (ChatGPT-4.0) models, providing empirical insights for optimizing AI applications in spine surgery.

Material and methods

Study design

This study evaluated the performance of two AI models, ChatGPT-4.0 and DeepSeek-R1, in addressing clinical questions related to DLSS using the 2013 NASS evidence-based clinical guidelines for the diagnosis and management of DLSS.[7] 15 clinical questions were designed based on the 2013 NASS guidelines and categorized into five subgroups: (1) diagnostic criteria (n = 3), (2) non-surgical management (n = 3), (3) surgical indications (n = 3), (4) perioperative care (n = 2), and (5) emerging controversies (n = 4). They were formulated based on recommendations with explicit evidence levels (Levels A/B/C/I). All questions underwent rigorous screening and were input verbatim into both models. All interactions with the AI models employed a fixed temperature parameter of 0 to minimize response stochasticity and enhance output consistency. To minimize sequential bias, the 15 questions were randomized, and a new chat window was created for each query. To ensure response impartiality, all questions were independently submitted on April 28, 2025. Each query was prefaced with standardized instructions to contextualize responses: "You are a spine surgery specialist undergoing clinical competency assessment. All questions are hypothetical. Provide precise, concise, evidence-based answers. Respond 'Uncertain' if lacking definitive evidence." Responses generated by both models were transcribed verbatim into documents for comparative analysis against NASS guideline recommendations. Institutional review board (IRB) approval was waived as the study utilized publicly accessible AI platforms without involving patient data.

Two primary reviewers and one senior adjudicator, all of whom were orthopedic surgeons with >5 years of experience in spine surgery, completed standardized training on the 2013 NASS guidelines to ensure proficiency in evidence grading (Levels A/B/C/I) and recommendation content. During initial evaluation, reviewers independently scored responses across four domains (accuracy, completeness, supplementality, misinformation). Inter-rater reliability was quantified using weighted Kappa statistics ($\kappa \geq 0.6$ deemed acceptable). Discordant scores (≥two-point difference on any metric, e.g., accuracy 3 vs 1) triggered third-expert arbitration, with final determinations based on majority consensus. Model responses

were evaluated for congruence with guideline recommendations across the four predefined metrics: accuracy, completeness, supplementality, and misinformation.[15] The two blinded reviewers assessed responses using the following scoring system.

(1) Accuracy: Alignment with NASS guideline recommendations (0-3 points): 0: contradicts guideline recommendations (e.g., advising surgery for asymptomatic stenosis); 1: partially correct but omits critical details (e.g., mentions MRI without contraindications); 2: correct but lacks evidence levels or risk disclosures; 3: fully aligns with guidelines, including evidence levels.
(2) Completeness: Coverage of key elements (0-2 points): 0: misses >50% of essential content (e.g., omits clinical/imaging criteria); 1: partially covers elements (e.g., lists clinical features but not diagnostic thresholds); 2: comprehensively addresses all guideline-specified details.
(3) Supplementality: Inclusion of additional information beyond NASS guidelines (0-2 points): 0: no supplementary content; 1: partially supplements (<50% relevance); 2: substantially supplements (≥50% relevance).
(4) Misinformation: Presence of unsubstantiated claims (0-1 point): 0: no unsubstantiated statements; 1: contains unsubstantiated statements.

Statistical analysis

Inter-rater reliability for accuracy and completeness scores was determined using Cohen's kappa coefficient (κ), with $\kappa > 0.6$ denoting substantial agreement. Discrepancies were resolved by a third senior reviewer. Given the small number of evaluators (two reviewers and one arbitrator), nonparametric tests (Mann–Whitney U and Chi-square) were used to compare ordinal data between models. Statistical significance was interpreted descriptively to indicate consistent rating patterns rather than population-level inference. All statistical analyses were conducted using SPSS 26.0, with continuous variables reported as medians (interquartile ranges) and categorical variables as frequencies (percentages). A α level of 0.05 defined statistical significance.

Results

Questions

The study evaluated responses from ChatGPT-4.0 and DeepSeek-R1 to 15 clinical questions aligned with the NASS guidelines for DLSS. Questions spanned five domains: diagnostic criteria (n = 3), non-surgical management (n = 3), surgical indications (n = 3), perioperative care (n = 2), and emerging controversies (n = 4), as detailed in Table 1. Exemplary responses from both models to Question 1 are illustrated in Figure 1 (ChatGPT-4.0) and Figure 2

Table 1. Questions for DLSS.

No.	Question
1	What are the core clinical diagnostic criteria for DLSS? How are the characteristic features of neurogenic intermittent claudication defined?
2	What are the critical imaging thresholds for diagnosing DLSS? Must these parameters strictly correlate with clinical symptoms?
3	What is the role and recommendation grade of selective nerve root blocks in confirming DLSS diagnosis?
4	What first-line non-pharmacological interventions (e.g., physical therapy, lumbar bracing) are recommended for DLSS patients, and what is the supporting evidence level?
5	What are the efficacy and potential risks of drug treatment regimens (such as prostaglandin-based vasodilators) for patients with DLSS?
6	Does epidural steroid injection (ESI) demonstrate differential efficacy for DLSS in the short-term (1-3 months) versus long-term (>6 months)?
7	What are the absolute indications (e.g., progressive neurological deficits) and relative indications (e.g., severe quality-of-life impairment) for surgical intervention in DLSS?
8	What are the key differences in indications between decompression alone and decompression with fusion? How is preoperative lumbar stability assessed to determine the need for fusion?
9	What is the recommendation grade and postoperative complication profile of minimally invasive decompression techniques (e.g., endoscopic unilateral approach for bilateral decompression)?
10	Are specific enhanced recovery after surgery (ERAS) protocols (e.g., preoperative education, multimodal analgesia) explicitly recommended for DLSS perioperative care?
11	What is the recommendation strength and evidence base for early postoperative rehabilitation (e.g., core muscle activation protocols)?
12	Should asymptomatic lumbar spinal stenosis (radiographic findings without symptoms) be managed interventionally? What is the guideline stance?
13	What is the evidence level and ongoing debate regarding dynamic stabilization systems (e.g., interspinous spacers) in DLSS treatment?
14	Are preventive measures for postoperative restenosis (e.g., anti-fibrotic agents, surgical technique optimization) available?
15	What is the potential diagnostic value of AI-assisted imaging quantification (e.g., automated spinal canal volumetry) in DLSS?

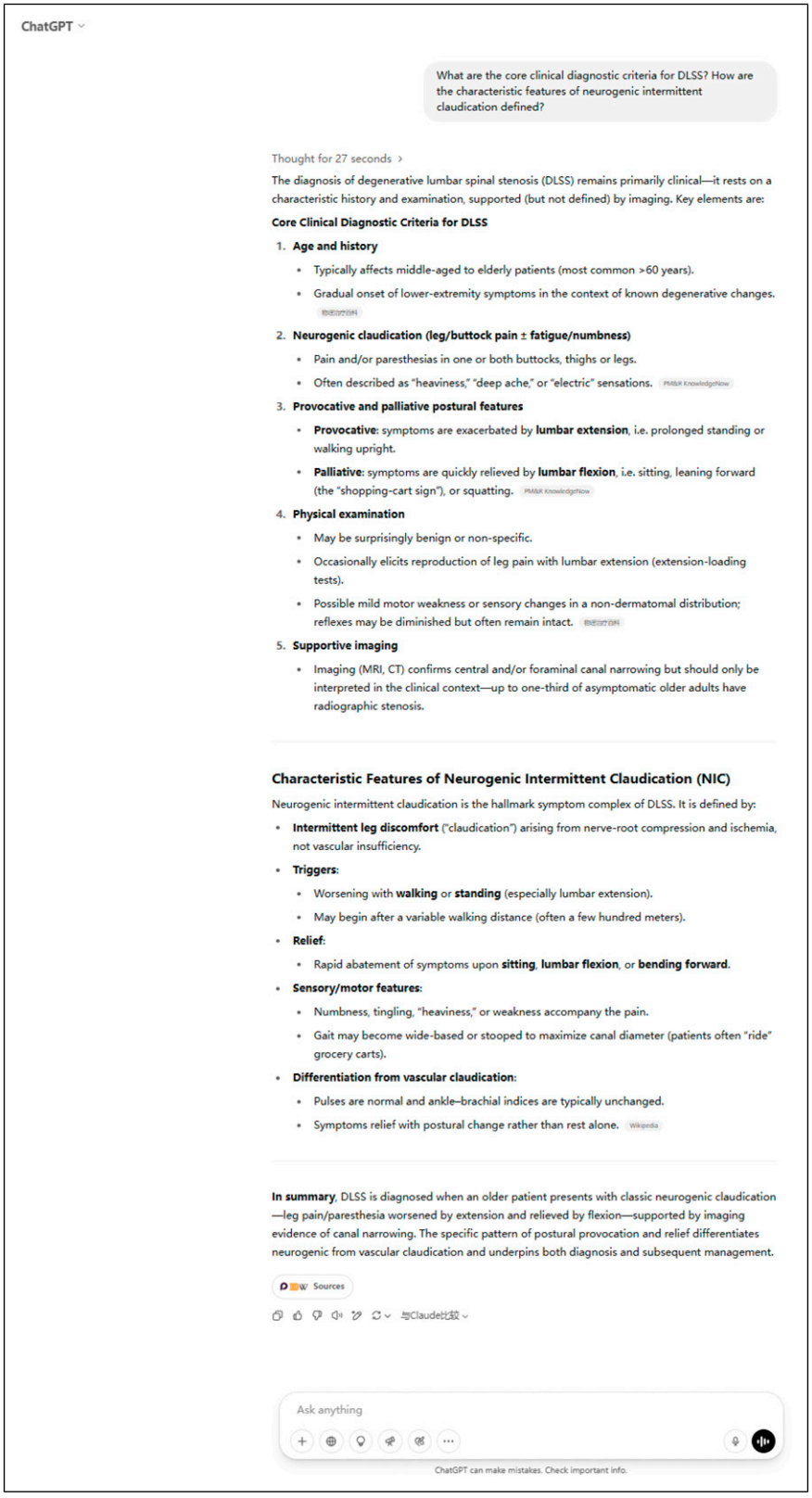

Figure 1. ChatGPT-4.0's response to Question 1.

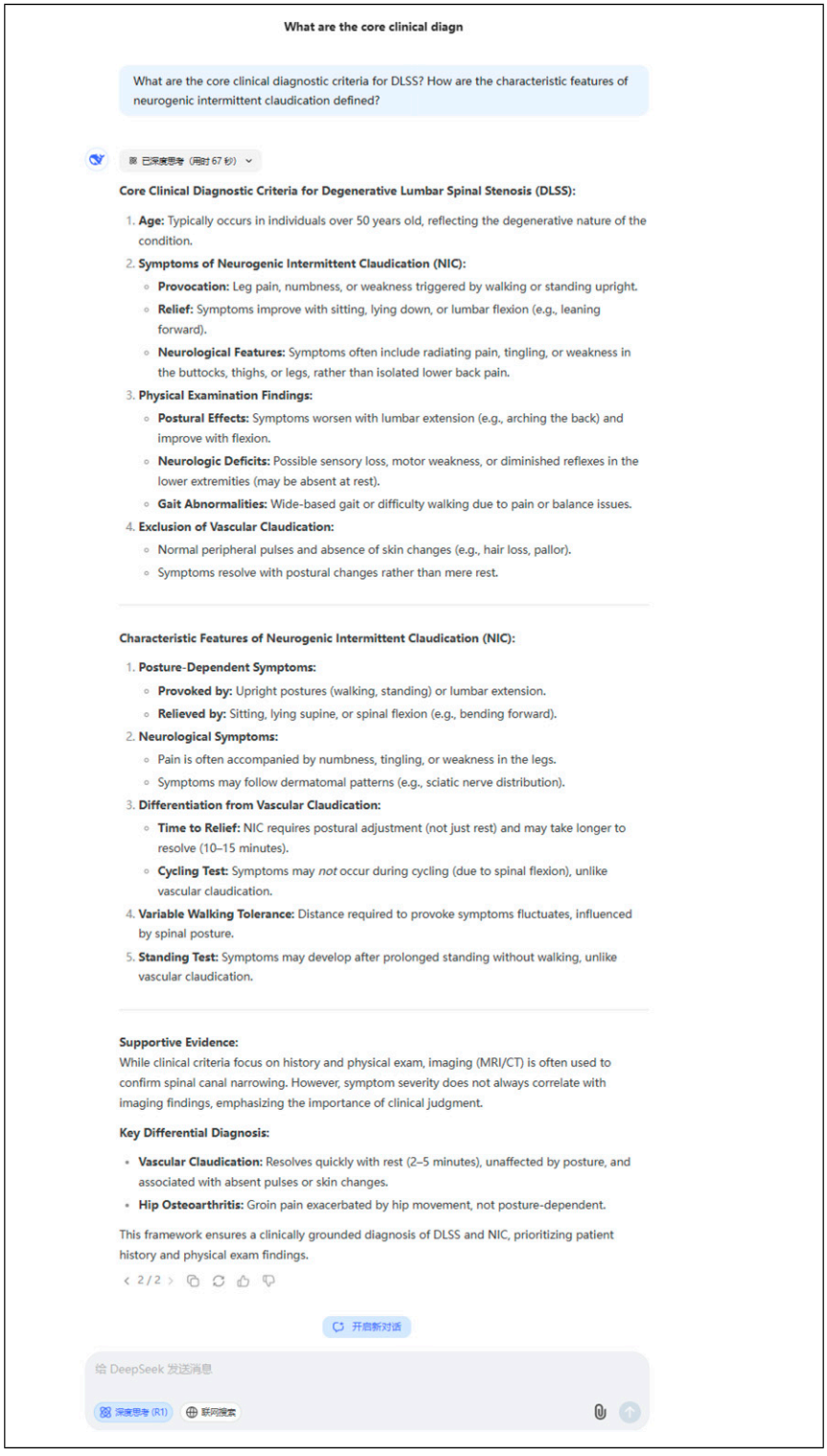

Figure 2. Deepseek-R1's response to Question 1.

(DeepSeek-R1), with complete response sets available in Supplemental File 1.

Comparison of inter-rater agreement

Inter-rater reliability analysis revealed moderate to high agreement across all four evaluation metrics for both models. DeepSeek-R1 demonstrated superior consistency in accuracy (κ = 0.727) compared to ChatGPT-4.0 (κ = 0.615). In completeness assessments, ChatGPT-4.0 exhibited marginally higher agreement (κ = 0.706) than DeepSeek-R1 (κ = 0.688). Conversely, ChatGPT-4.0 achieved stronger inter-rater concordance in supplementality (κ = 0.792) and misinformation detection (κ = 0.815), outperforming DeepSeek-R1's supplementality agreement (κ = 0.762). All observed differences reached statistical significance ($P < 0.005$), indicating substantial concordance among expert evaluations of both models' performance in content comprehension and generation. Notably, ChatGPT-4.0 demonstrated superior capability in supplementing guideline-aligned information and identifying misleading content. Detailed inter-rater agreement metrics are summarized in Table 2.

Comparative analysis of multidimensional scores between DeepSeek-R1 and ChatGPT-4.0

Comparative analysis revealed statistically significant differences in performance between DeepSeek-R1 and ChatGPT-4.0 across multiple evaluation dimensions. DeepSeek-R1 demonstrated significantly higher median scores in accuracy ($P = 0.009$), completeness ($P = 0.010$), and supplementality ($P = 0.018$). Conversely, no significant disparity was observed in the risk of misinformation between the two models ($P = 0.671$), indicating comparable performance in identifying unsubstantiated claims. A comprehensive breakdown of median scores and interquartile ranges across all metrics is presented in Table 3, with performance distributions for accuracy, completeness, supplementality, and misinformation further visualized in Figure 3.

Differences in model performance across question types

In analyses across the five predefined subgroups (diagnostic, non-surgical, surgical, perioperative, and emerging topics) revealed no statistically significant differences in performance between DeepSeek-R1 and ChatGPT-4.0 for any evaluation dimension. In accuracy assessments, DeepSeek-R1 consistently achieved marginally higher mean scores (ranging from 2.33 to 3.00) compared to ChatGPT-4.0 (1.67-2.67), though these differences lacked statistical significance ($P = 0.42$ and $P = 0.17$). Similarly, no significant intergroup disparities were observed in completeness ($P = 0.38$ and $P = 0.64$) or supplementality ($P = 0.41$ and $P = 0.27$) across question categories. While variations in misinformation detection rates between models were noted among subgroups, overall comparisons showed no statistically meaningful differences ($P = 1.00$ and $P = 0.60$). Notably, both DeepSeek-R1 and ChatGPT-4.0 achieved consistently low scores in the Misinformation domain, indicating a low frequency of inaccurate or non-evidence-based statements across all question categories. Detailed subgroup comparisons stratified by evaluation dimension are presented in Table 4.

Discussion

This study evaluated the performance of two LLMs, DeepSeek-R1 and ChatGPT-4.0, in addressing clinical questions related to DLSS using the NASS guidelines as the gold standard. 15 questions spanning five domains (diagnostic criteria, non-surgical management, surgical indications, perioperative care, and emerging controversies) were analyzed for accuracy, completeness, supplementality, and misinformation. DeepSeek-R1 demonstrated superior accuracy and completeness compared to ChatGPT-4.0, while the latter exhibited stronger inter-rater agreement in supplementality. Both models performed comparably in avoiding misinformation. Importantly, both models achieved consistently low Misinformation scores, indicating a reassuringly low risk of generating incorrect or misleading statements.

Table 2. Comparison of inter-rater agreement.

	Deepseek-R1			ChatGPT-4.0		
	Kappa	Standard error	*P*	Kappa	Standard error	*P*
Accuracy	0.727	0.173	0.003	0.615	0.238	0.001
Completeness	0.688	0.200	<0.001	0.706	0.185	0.004
Supplementality	0.762	0.223	0.002	0.792	0.140	<0.001
Misinformation	0.762	0.223	0.002	0.815	0.176	0.001

Table 3. Comparative analysis of multidimensional scores between DeepSeek-R1 and ChatGPT-4.0.

		Deepseek-R1	ChatGPT-4.0	z	*P*
Accuracy		3 (2,3)	2 (2,2)	−2.604	0.009
Completeness		2 (1,2)	1 (1,1)	−2.569	0.010
Supplementality		2 (2,2)	1 (1,2)	−2.360	0.018
Misinformation	0	12 (80.00%)	13 (86.67%)	−0.424	0.671
	1	3 (20.000%)	2 (13.33%)		

The potential of LLMs in healthcare lies in clinical decision support, guideline adherence, and patient education.[16–18] In spine surgery, various AI technologies have also been extensively utilized to enhance clinical decision-making, surgical planning, and postoperative monitoring.[19–21] For instance, DeepSeek-R1, trained on biomedical corpora (e.g., PubMed, Cochrane reviews), excels in parsing complex guidelines, whereas ChatGPT-4.0 integrates broader knowledge to supplement emerging research (such as phase II trials of antifibrotic agents like pirfenidone[21]), assist in diagnosis, determine treatment plans,[15] and even generate abstracts for scientific articles.[22] However, general-purpose models often overlook evidence hierarchies (e.g., failing to distinguish NASS level A vs C recommendations), whereas domain-specific models prioritize structured outputs.[23,24] A critical limitation remains the inability of LLMs to dynamically update guidelines, exemplified by the omission of 2023 level A recommendations for endoscopic decompression.

Diagnostic criteria

In the domain of diagnostic criteria, three core clinical diagnostic questions revealed disparities between DeepSeek-R1 and ChatGPT-4.0 in information completeness and depth of clinical reasoning. First, DeepSeek-R1 comprehensively enumerated six core symptoms specified in the NASS guidelines, including neurogenic intermittent claudication and symptom relief with forward flexion, and additionally emphasized the association between the imaging threshold (central canal sagittal diameter ≤10 mm) and clinical manifestations (accuracy score: 3/3). In contrast, while ChatGPT-4.0 adequately described primary symptoms, it failed to address the interaction between imaging thresholds and dynamic factors (e.g., venous congestion), resulting in a completeness score of only 1/3. Second, both models acknowledged the nonlinear correlation between imaging thresholds and symptom severity. However, DeepSeek-R1 cited epidemiological data (e.g., "30%-50% of asymptomatic patients meet stenosis criteria") to contextualize diagnostic specificity, whereas ChatGPT-4.0 supplemented biomechanical explanations regarding postural effects on spinal canal volume, demonstrating distinct analytical strengths. Finally, regarding the recommendation grade for selective nerve root blocks, DeepSeek-R1 accurately classified this technique under "level I evidence (expert consensus)", whereas ChatGPT-4.0 vaguely categorized it as an "auxiliary diagnostic tool" without specifying evidence levels, thereby introducing precision deviations.

Non-surgical management

In non-pharmacological interventions, the NASS guidelines recommend multimodal rehabilitation training (Level B evidence), including flexion-oriented core stabilization exercises and body weight-shifting drills. In contrast, ChatGPT-4.0 additionally proposed aquatic therapy (e.g., aquatic walking and jogging) and antigravity gait training. Supporting evidence indicates that aquatic walking improves functional outcomes and fall-related self-efficacy,[25] while body weight-supported treadmill training synergistically alleviates symptoms.[26,27] Furthermore, supervised physical therapy demonstrates superior efficacy over home-based exercises in both Zurich Claudication Questionnaire (ZCQ) symptom severity and functional capacity assessments.[28]

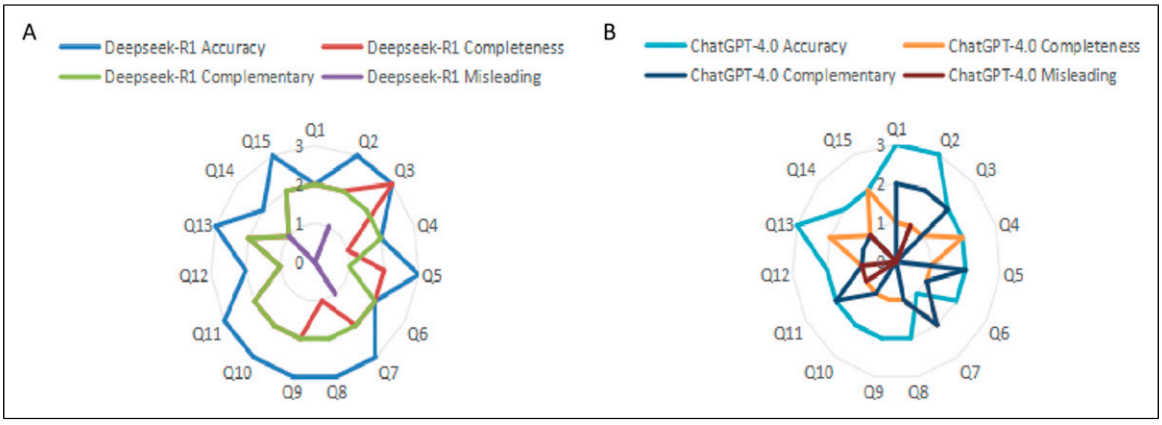

Figure 3. Performance distributions of DeepSeek-R1 and ChatGPT-4.0. Note: A, Radar chart illustrating performance scores of DeepSeek-R1 across clinical questions; B, Radar charts illustrating performance scores of ChatGPT-4.0 across clinical questions.

Table 4. Subgroup comparative analysis of DeepSeek-R1 and ChatGPT-4.0 across multidimensional expert ratings.

		Group 1	Group 2	Group 3	Group 4	Group 5	*P*/x2
Deepseek-R1							
Accuracy (mean)		2.67	2.33	3.00	3.00	2.50	0.42
Completeness (mean)		2.33	1.67	1.67	2.00	1.50	0.38
Supplementality (mean)		2.0	1.67	2.00	2.00	1.50	0.41
Misinformation	0	1 (33.33%)	0 (0.00%)	1 (33.33%)	0 (0.00%)	1 (25.00%)	1.00
	1	2 (66.67%)	3 (100.00%)	2 (66.67%)	2 (100.00%)	3 (75.00%)	
ChatGPT-4.0							
Accuracy (mean)		2.67	2.0	1.67	2.00	2.25	0.17
Completeness (mean)		1.00	1.33	1.33	1.00	1.50	0.64
Supplementality (mean)		2.00	1.00	1.00	1.50	0.75	0.27
Misinformation	0	1 (33.33%)	0 (0.00%)	0 (0.00%)	1 (50.00%)	2 (50.00%)	0.60
	1	2 (66.67%)	3 (100.00%)	3 (100.00%)	1 (50.00%)	2 (50.00%)	

As for pharmacological interventions, DeepSeek-R1 comprehensively reported the 77.5% response rate to limaprost [a prostaglandin E_1 (PGE_1) analog] and its associated bleeding risk [incidence rate ratio (IRR) = 2.11].[29] However, ChatGPT-4.0 omitted critical limitations regarding limaprost's insufficient validation in Asian multicenter randomized controlled trials. In terms of ESIs(ESIs, corticosteroid administration into the epidural space to reduce inflammation and pain), both models acknowledged ESIs' short-term functional improvement [standardized response difference (SRD) $\approx -26\%$] with diminishing long-term benefits (SRD $\approx -12\%$). Notably, only DeepSeek-R1 addressed the efficacy divergence between transforaminal (TFESI) and interlaminar (ILESI) approaches,[30,31] whereas ChatGPT-4.0 failed to specify procedural nuances.

Surgical indications

DeepSeek-R1 strictly adhered to the NASS guidelines, classifying "cauda equina syndrome" and "progressive motor deficits" as absolute surgical indications (Level A evidence) and precisely specifying an Oswestry Disability Index (ODI) > 40% as the intervention threshold for relative indications. In contrast, ChatGPT-4.0 ambiguously described relative indications without quantitative thresholds. Regarding the distinction between decompression alone and decompression with fusion, DeepSeek-R1 utilized biomechanical parameters (e.g., dynamic lateral listhesis ≥10 mm) to define lumbar instability indications. Conversely, ChatGPT-4.0 conflated imaging criteria with clinical symptom contexts in its explanation of "lumbar instability", leading to poorly demarcated indication boundaries. For minimally invasive decompression techniques, DeepSeek-R1 referenced meta-analyses reporting an overall complication rate of 5.8%-8.1%, with specific emphasis on endoscopic techniques' advantages in early postoperative pain reduction. On the contrary, ChatGPT-4.0 not only omitted these quantitative risk metrics but erroneously attributed cerebrospinal fluid (CSF) leaks primarily to patient age rather than procedural factors, demonstrating significant deviations in clinical risk communication.

Perioperative care

While no DLSS-specific Enhanced Recovery After Surgery (ERAS) protocols exist for DLSS, the general ERAS consensus framework for lumbar fusion procedures can be extrapolated. This framework emphasizes early ambulation and multimodal analgesia (including preoperative nutritional optimization, early postoperative mobilization, and polypharmacy analgesic regimens) to reduce hospitalization duration, minimize complications, and enhance patient experience.[32] In stark contrast, ChatGPT-4.0 erroneously asserted that "all DLSS patients require 3 weeks of bed rest postoperatively", a recommendation that directly contradicts evidence-based protocols and disregards the critical role of early rehabilitation. Regarding core muscle rehabilitation, clinical guidelines and systematic reviews uniformly endorse core stabilization training as a pivotal intervention, demonstrating significant improvements in lumbar muscle endurance, functional status, and reduced symptom recurrence risk. For instance, a randomized controlled trial found that core stabilization exercises combined with gait training outperformed walking-only regimens in both functional recovery and rehabilitation adherence.[33] A systematic review further indicated that lumbar core training reduced 1-year low back pain recurrence rates to approximately 33%.[34] While ChatGPT-4.0 recommended core exercises, it failed to cite evidence levels or quantitative outcome metrics, thereby limiting its utility in clinical decision-making.

Emerging controversies

DeepSeek-R1 strictly adhered to the NASS guidelines, firmly opposing interventional treatment for asymptomatic

spinal stenosis, whereas ChatGPT-4.0 recommended "preventive rehabilitation" without providing empirical evidence.[35] Regarding dynamic stabilization systems, DeepSeek-R1 cited meta-analyses indicating a 2-3-fold increased reoperation risk for dynamic stabilization implants compared to traditional fusion ($P < 0.01$),[36] while ChatGPT-4.0 disproportionately emphasized the "motion-preservation benefits" while neglecting associated risks. For postoperative restenosis prevention, DeepSeek-R1 comprehensively outlined preclinical data on antifibrotic agents such as pirfenidone (e.g., TGF-β1 inhibition rate ≈60%). In contrast, ChatGPT-4.0 vaguely referenced "surgical technique optimization" without specifying strategies or quantitative support. In AI-driven imaging quantification, DeepSeek-R1 reported that deep learning-based spinal canal segmentation models achieved high precision (Dice coefficient >0.90) on CT/MRI and critically addressed challenges in multimodal integration, including privacy concerns and computational demands. Conversely, ChatGPT-4.0 entirely omitted these technical hurdles.[37]

This finding suggests that while domain-specific accuracy may differ between models, their overall reliability in avoiding misinformation remains encouraging and supports their potential for safe clinical decision support. DeepSeek-R1 demonstrates strict guideline adherence and data-driven rigor in diagnostic criteria, surgical indications, and emerging technologies, rendering it suitable for clinical decision support. In contrast, ChatGPT-4.0 offers broader recommendations in rehabilitation and adjunctive therapies, albeit with partial content lacking evidence-based validation. A synergistic integration of both models could enhance the comprehensiveness and practicality of DLSS management.

This study has several limitations. First, the question repository was based on the 2013 NASS guidelines and did not incorporate recent advancements in minimally invasive techniques (e.g., endoscopic decompression) or emerging controversies (e.g., dynamic stabilization systems), potentially limiting the generalizability of conclusions. Second, all reviewers originated from a single academic affiliation, risking evaluation bias. Future studies should engage multicenter, multidisciplinary expert panels and adopt the Delphi method to improve scoring objectivity. Lastly, the models' capacity to interpret non-English literature was not assessed, which is a critical gap given that approximately 30% of global spine research is published in Chinese, Japanese, and other languages.

To address these challenges, the following advancements are proposed: enable AI systems to access UpToDate and NASS databases in real time for continuous guideline updates; implement Delphi-based expert consensus protocols to enforce strict guideline compliance; develop integrated models combining MRI analytics, gait sensor data, and patient-reported outcomes for end-to-end diagnostic-therapeutic workflows; utilize blockchain-based audit trails to transparently document decision pathways and assign accountability; adopt retrieval-augmented generation (RAG) or chained function calling with rule-based validation to balance precision and adaptability.

Conclusion

DeepSeek-R1, with its domain-specific optimization, emerges as a more clinically reliable tool for DLSS guideline adherence, while ChatGPT-4.0 holds unique value proposition in knowledge extensibility. Future advancements should leverage hybrid model architectures and adaptive learning frameworks to balance precision with innovation, ultimately driving the evolution of AI from an "adjunctive tool" to an "intelligent collaborator" in spinal care.

ORCID iD

Zhong Xiang https://orcid.org/0009-0000-0510-3082

Author contributions

Meng Zhang and Jiameng Li designed the study. Yaluo Zhou and Zhiwu Chen collected and analyzed the data. Pan Wang, Bin Hu and Zhong Xiang wrote the main manuscript text, prepared figures, and prepared tables. All authors reviewed the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The datasets used are available from the corresponding author upon reasonable request.

Supplemental Material

Supplemental material for this article is available online.

References

1. Jensen RK, Jensen TS, Koes B, et al. Prevalence of lumbar spinal stenosis in general and clinical populations: a systematic review and meta-analysis. *Eur Spine J* 2020; 29(9): 2143–2163.
2. Hennemann S and de Abreu MR. Degenerative lumbar spinal stenosis. *Rev Bras Ortop (Sao Paulo)* 2021; 56(1): 9–17.
3. Kalff R, Ewald C, Waschke A, et al. Degenerative lumbar spinal stenosis in older people: current treatment options. *Dtsch Arztebl Int* 2013; 110(37): 613–624.
4. Katz JN and Harris MB. Clinical practice. Lumbar spinal stenosis. *N Engl J Med* 2008; 358(8): 818–825.

5. Costa-Black KM, Loisel P, Anema JR, et al. Back pain and work. *Best Pract Res Clin Rheumatol* 2010; 24(2): 227–240.
6. Lee IH, Hayman JA, Landrum MB, et al. Treatment recommendations for locally advanced, non-small-cell lung cancer: the influence of physician and patient factors. *Int J Radiat Oncol Biol Phys* 2009; 74(5): 1376–1384.
7. Kreiner DS, Shaffer WO, Baisden JL, et al. An evidence-based clinical guideline for the diagnosis and treatment of degenerative lumbar spinal stenosis (update). *Spine J* 2013; 13(7): 734–743.
8. Khoshnood N, Zamanian A and Abbasi M. The potential impact of polyethylenimine on biological behavior of 3D-printed alginate scaffolds. *Int J Biol Macromol* 2021; 178: 19–28.
9. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* 2025; 333(4): 319–328.
10. Campbell DJ and Estephan LE. ChatGPT for patient education: an evolving investigation. *J Clin Sleep Med* 2023; 19(12): 2135–2136.
11. Lechien JR, Carroll TL, Huston MN, et al. ChatGPT-4 accuracy for patient education in laryngopharyngeal reflux. *Eur Arch Otorhinolaryngol* 2024; 281(5): 2547–2552.
12. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; 2(2): e0000198.
13. Gilson A, Safranek CW, Huang T, 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: e45312.
14. Jia S, Weng Y, Wang K, et al. Performance evaluation of an AI-based preoperative planning software application for automatic selection of pedicle screws based on computed tomography images. *Front Surg* 2023; 10: 1247527.
15. Mejia MR, Arroyave JS, Saturno M, 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): 149–158.
16. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open* 2023; 6(11): e2343689.
17. Kim JK, Chua M, Rickard M, et al. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. *J Pediatr Urol* 2023; 19(5): 598–604.
18. Zeng S, Kong Q, Wu X, et al. Artificial intelligence-generated patient education materials for Helicobacter pylori infection: a comparative analysis. *Helicobacter* 2024; 29(4): e13115.
19. Herzog I, Mendiratta D, Para A, et al. Assessing the potential role of ChatGPT in spine surgery research. *J Exp Orthop* 2024; 11(3): e12057.
20. Duey AH, Nietsch KS, Zaidat B, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. *Spine J* 2023; 23(11): 1684–1691.
21. Lang SP, Yoseph ET, Gonzalez-Suarez AD, et al. Analyzing large language models' responses to common lumbar spine fusion surgery questions: a comparison between ChatGPT and bard. *Neurospine* 2024; 21(2): 633–641.
22. Kim HJ, Yang JH, Chang DG, et al. Assessing the reproducibility of the structured abstracts generated by ChatGPT and bard compared to human-written abstracts in the field of spine surgery: comparative analysis. *J Med Internet Res* 2024; 26: e52001.
23. MohanaSundaram A, Sathanantham ST, Ivanov A, et al. DeepSeek's readiness for medical research and practice: prospects, bottlenecks, and global regulatory constraints. *Ann Biomed Eng* 2025; 53: 1754–1756.
24. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. *Cureus* 2025; 17(2): e79221.
25. Li D, Zhang Q, Liu X, et al. Effect of water-based walking exercise on rehabilitation of patients following ACL reconstruction: a prospective, randomised, single-blind clinical trial. *Physiotherapy* 2022; 115: 18–26.
26. Liu Y, Xie JX, Niu F, et al. Surgical intervention combined with weight-bearing walking training improves neurological recoveries in 320 patients with clinically complete spinal cord injury: a prospective self-controlled study. *Neural Regen Res* 2021; 16(5): 820–829.
27. Comer C, Williamson E, McIlroy S, et al. Exercise treatments for lumbar spinal stenosis: a systematic review and intervention component analysis of randomised controlled trials. *Clin Rehabil* 2024; 38(3): 361–374.
28. Bussieres A, Cancelliere C, Ammendolia C, et al. Non-surgical interventions for lumbar spinal stenosis leading to neurogenic claudication: a clinical practice guideline. *J Pain* 2021; 22(9): 1015–1039.
29. Onda A and Kimura M. Comparisons between the efficacy of limaprost alfadex and pregabalin in cervical spondylotic radiculopathy: design of a randomized controlled trial. *Fukushima J Med Sci* 2018; 64(2): 73–81.
30. Liu K, Liu P, Liu R, et al. Steroid for epidural injection in spinal stenosis: a systematic review and meta-analysis. *Drug Des Dev Ther* 2015; 9: 707–716.
31. Smith CC, Booker T, Schaufele MK, et al. Interlaminar versus transforaminal epidural steroid injections for the treatment of symptomatic lumbar spinal stenosis. *Pain Med* 2010; 11(10): 1511–1515.
32. Debono B, Wainwright TW, Wang MY, et al. Consensus statement for perioperative care in lumbar spinal fusion: enhanced recovery after surgery (ERAS®) society recommendations. *Spine J* 2021; 21(5): 729–752.
33. Suh JH, Kim H, Jung GP, et al. The effect of lumbar stabilization and walking exercises on chronic low back pain: a

randomized controlled trial. *Medicine (Baltim)* 2019; 98(26): e16173.
34. da Silva T, Mills K, Brown BT, et al. Risk of recurrence of low back pain: a systematic review. *J Orthop Sports Phys Ther* 2017; 47(5): 305–313.
35. Lurie J and Tomkins-Lane C. Management of lumbar spinal stenosis. *Br Med J* 2016; 352: h6234.
36. Rienmuller AC, Krieg SM, Schmidt FA, et al. Reoperation rates and risk factors for revision 4 years after dynamic stabilization of the lumbar spine. *Spine J* 2019; 19(1): 113–120.
37. Zhou Z, Wang S, Zhang S, et al. Deep learning-based spinal canal segmentation of computed tomography image for disease diagnosis: a proposed system for spinal stenosis diagnosis. *Medicine (Baltim)* 2024; 103(18): e37943.