European Spine Journal
https://doi.org/10.1007/s00586-026-09763-x

RESEARCH

Application and efficacy of artificial intelligence in patient education on spinal cord injuries

Jonas Krueckel[1,2] · Melanie Ardelt[3] · David Schiffelholz[1] · Josina Straub[1] · Sebastian Siller[4] · Vanessa Hubertus[5] · Sonja Häckel[6] · Denis Bratelj[7] · Christof Wutte[8,9] · Helena Arias[1,10] · Franz Hilber[2] · Volker Alt[1] · Siegmund Lang[1]

Received: 4 February 2025 / Revised: 14 November 2025 / Accepted: 21 January 2026

Abstract

Introduction/background Spinal cord injuries (SCI) present complex challenges for patients, who increasingly turn to online resources for supplementary information. Large language models (LLMs) like ChatGPT and Google Gemini have emerged as potential tools for patient education. However, concerns about the accuracy, clarity, and comprehensiveness of their responses remain, particularly in specialized fields such as SCI. This study aimed to evaluate the performance of ChatGPT 4, ChatGPT 3.5, and Google Gemini in addressing common patient questions about SCI.

Material and methods A systematic process was used to identify 10 key patient questions related to SCI from online sources, PubMed, and Google Trends. These questions were submitted to ChatGPT 4, ChatGPT 3.5, and Google Gemini using a standardized prompt and a 150-word response cap to elicit expert-like responses. Eight blinded spine surgeons evaluated the chatbot-generated answers for quality, clarity, empathy, and comprehensiveness using a validated rating system. Responses were categorized as "excellent," "satisfactory with minimal clarification," "satisfactory with moderate clarification," or "unsatisfactory."

Results Across all three models, the majority of responses were rated as either excellent or requiring only minimal clarification. ChatGPT 4 achieved the highest proportion of high-quality responses, with up to almost 90% rated as "excellent" or "minimal clarification required." ChatGPT 3.5 and Google Gemini performed similarly, with slightly lower percentages of high-quality responses. No statistically significant differences were observed between the models in overall performance.

Conclusion In a standardized single turn, 150-word setting, publicly available LLMs produced largely satisfactory answers to common SCI questions with comparable performance across models. LLMs can be recommended as adjuncts for general patient education, while their outputs should be reviewed within clinical care. Further studies should test multi turn interactions, include patient and multidisciplinary evaluators, compare chatbot responses with clinician authored answers and evaluate the performance of domain specific medical LLMs.

Level of evidence II.

Keywords Spinal cord injury · Artificial intelligence · Large language models · Spine surgery · Patient education

Introduction

Spinal cord injuries (SCI) represent complex medical conditions that pose significant challenges for patients. Individuals living with spinal cord injury face a multifaceted landscape of physiological, psychological, and social complexities [1, 2]. As individuals seek to better understand their conditions and make informed decisions about their healthcare, many turn to online resources to complement traditional medical advice [3]. In recent years, large language models (LLMs) such as ChatGPT have emerged as powerful tools for obtaining medical information, offering insights across a wide range of domains [4–8]. However, while LLMs hold great promise in making medical knowledge accessible, the quality of the information they provide can vary [9]. The vast scope of data processed by these models means that users risk encountering outdated, oversimplified, or even misleading content, potentially jeopardizing

Extended author information available on the last page of the article

Published online: 27 February 2026

patient outcomes [10, 11]. This variability underscores the need for rigorous evaluation of their reliability, particularly in fields like SCI, where decisions are often nuanced, highly individualized, and deeply consequential.

The role of high-quality educational materials extends beyond individual empowerment. Clear, accurate, and empathetic communication is a cornerstone of partnerships between patients and the multidisciplinary care team [12]. Reliable information not only equips patients with the knowledge needed to make decisions but also fosters trust and mutual understanding between patients and their healthcare providers [13, 14]. As the digital landscape evolves, advanced AI technologies like LLMs offer a unique opportunity to transform these interactions. By generating personalized, scientifically grounded, and easily comprehensible responses, LLMs could bridge existing gaps in patient education and enhance the transparency and efficacy of medical communication.

Despite this potential, critical concerns remain. Questions about the accuracy, comprehensiveness, and contextual appropriateness of LLM-generated outputs are particularly pressing in specialized domains like SCI. Whether LLMs can meet the stringent requirements of this field and fulfill their promise as transformative tools for patient education is a matter that demands careful investigation.

This study seeks to address this gap by evaluating whether LLMs can accurately and effectively respond to common patient questions about SCI. By assessing their ability to provide clear, precise, and empathetic answers, this research aims to explore the potential of LLMs as reliable tools for enhancing patient education and empowering individuals to make informed healthcare decisions.

Materials and methods

Design and data assessment

A structured and thorough approach was undertaken to identify the most common questions related to spinal cord injuries. Searches were conducted on PubMed and Google using the terms "frequently asked questions AND spinal cord injury OR SCI," which returned approximately 28.8 million results as of September 20, 2024. From these, the first 20 Google search results were assessed according to predefined criteria. Eligible sources were required to have been published after December 31, 2017, written in English, and presented in an FAQ or Q&A format. Resources were excluded if they provided non-generalizable content, such as information limited to specific implants or healthcare providers.

To enhance the search process, ChatGPT 4 was used with the prompt: "Suggest a list of the 20 most common frequently asked patient questions about spinal cord injury." Additionally, global data trends were reviewed using Google Trends with the keyword "spinal cord injury" to identify topics of widespread interest and current relevance.

This multi-step strategy yielded a pool of 121 unique questions. These were systematically reviewed and grouped into 10 key themes that reflect the most pressing and commonly raised concerns of individuals affected by SCI. The final list of questions was refined to ensure it addressed critical aspects comprehensively (Table 1). A flowchart detailing this methodology is presented in Fig. 1.

The finalized questions were submitted to the AI chatbots ChatGPT 4, ChatGPT 3.5, and Google Gemini through their

Table 1 Overview of Frequently Asked Questions (FAQs) Provided to Large Language Models (LLMs.), labelled Q1-Q10

FAQS
Q1: What are spinal cord injuries?
Q2: What is the difference between paraplegia and tetraplegia?
Q3: Is there a cure for spinal cord injuries?
Q4: What are the chances of recovery from a spinal cord injury?
Q5: What are the short-term and long-term complications of SCI?
Q6: What kinds of help and treatment are effective in helping people to cope with their new life situation, and to find ways of enjoying life, after a spinal cord injury?
Q7: Can spinal cord injuries cause chronic pain?
Q8: What are the emotional and psychological challenges of living with an SCI?
Q9: What resources and financial assistance are available for SCI patients?
Q10: How do I find trustworthy sources that deal with questions about spinal cord injuries?

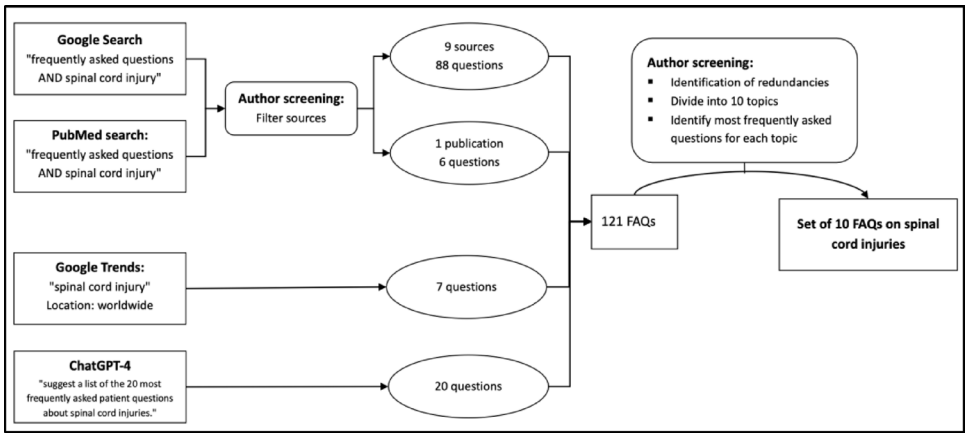

Fig. 1 Selection Process for Identifying the Top 10 Frequently Asked Questions About Spinal Cord Injuries

respective online portals. Each chatbot was engaged using the following standardized prompt: "Act as doctor and expert in the field of spinal cord injuries, who is up to date with the latest scientific research and has years of experience counselling patients with empathy and clarity. Provide a comprehensive and easily understandable answer to the following question about spinal cord injuries (SCI)! Limit your answer to 150 words and focus on the most important aspects to ensure patient information: (…)". To minimize bias from previous responses, a new session was initiated for each question. To facilitate comparisons across models, all interactions were constrained to a single-turn response and a 150-word limit.

ChatGPT, developed by OpenAI, is based on the generative pre-trained transformer (GPT) architecture, pre-trained on diverse internet text to produce versatile and detailed responses. Updates from GPT-3.5 to GPT-4 have enhanced reasoning, factual accuracy, and human-like output.

Google Gemini, rooted in the Pathways Language Model (PaLM) framework, is optimized for conversational applications, integrating reinforcement learning from human feedback (RLHF) to deliver contextually relevant and empathetic responses. These technical differences highlight the potential for variation in how each model handles patient inquiries, with ChatGPT excelling in broad generative tasks and Gemini focusing on real-time conversational relevance and multimodal capabilities, particularly in nuanced medical contexts like SCI.

An expert panel of eight blinded, board-certified spine surgeons in orthopedics or neurosurgery with extensive experience in SCI management evaluated the chatbot responses. Unaware that the answers were AI-generated, the assessors rated each response using a validated rating system [15]. The panel included surgeons practicing in Germany, Austria, and Switzerland, thereby situating the assessment within a Central European healthcare context. Responses were classified into four categories: 'excellent response not requiring clarification,' 'satisfactory requiring minimal clarification,' 'satisfactory requiring moderate clarification,' or 'unsatisfactory requiring substantial clarification.' A satisfactory response was defined as accurate but potentially requiring additional detail or clarification. Responses requiring moderate clarification contained missing or outdated information, while unsatisfactory responses exhibited significant inaccuracies or were overly generic, potentially leading to misinterpretation. For responses that were not rated as "excellent," raters were asked to specify the rationale for their decision. The predefined options included: "off-topic," "clear mistakes," "too much information," "too few information," "language problems," or "other issues."

Additionally, the raters answered four supplemental questions on a five-point Likert scale to evaluate the exhaustiveness, clarity and length of the responses and whether they effectively addressed patient concerns with empathy. All verbatim chatbot responses (per model and question) are provided in Supplementary Material S1.

Statistical analysis

Statistical analysis was conducted using GraphPad Prism (version 10.1, GraphPad Software Inc., San Diego, CA, USA). The distribution of response ratings across predefined categories was analyzed using the Wilcoxon signed-rank test. Differences in ratings between ChatGPT 4, ChatGPT 3.5, and Google Gemini were examined with the Mann-Whitney U-test. Statistical significance was defined as $p < 0.05$. This study was determined to be exempt from Institutional Review Board review.

Results

Across all three models and the 10 evaluated questions, the majority of responses were rated as satisfactory, with only a small proportion categorized as unsatisfactory.

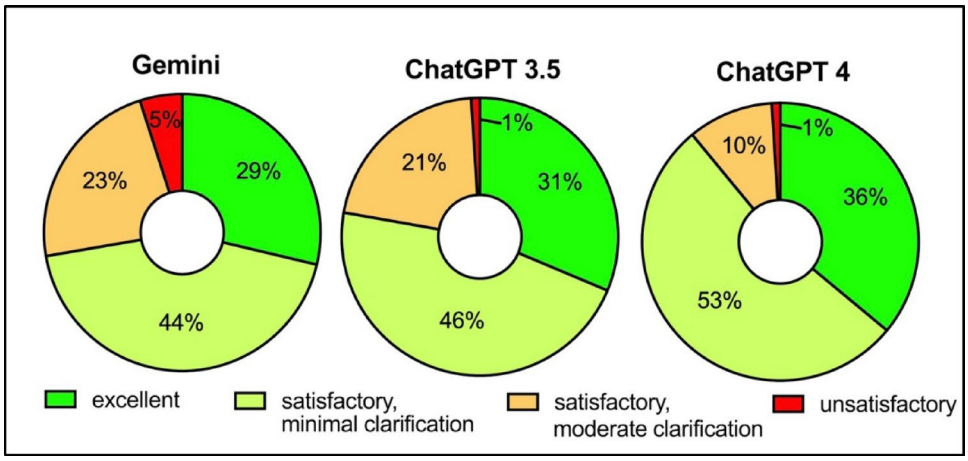

Fig. 2 Pie Chart Depicting the Percentage Distribution of Overall Ratings for the Combined Question Set Across All Three LLMs

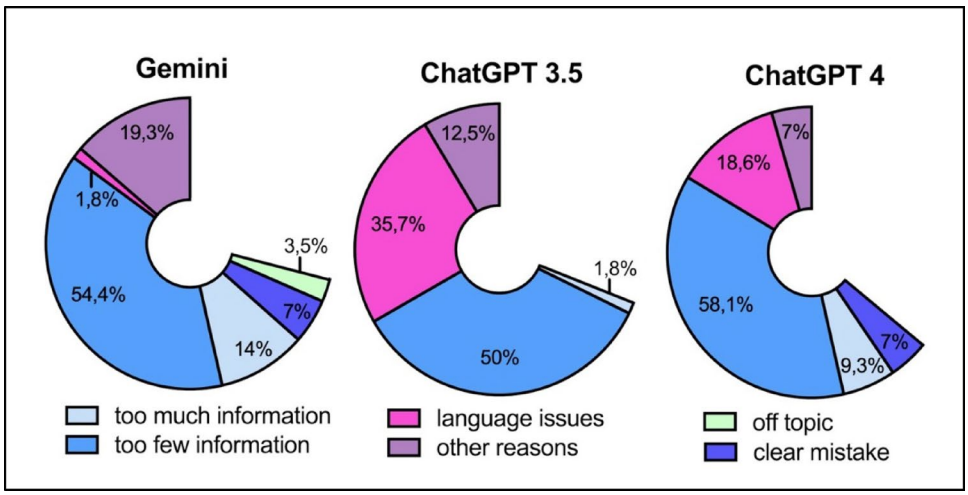

Fig. 3 Distribution of Reasons for Non-Excellent Ratings Across Google Gemini, ChatGPT 3.5, and ChatGPT 4, Expressed as Percentages

Google Gemini provided 29% of responses rated as excellent, 44% requiring minimal clarification, 23% requiring moderate clarification, and 5% categorized as unsatisfactory. ChatGPT 3.5 achieved 31% of responses rated as excellent, 46% requiring minimal clarification, 21% requiring moderate clarification, and only 1% rated as unsatisfactory. ChatGPT 4 demonstrated the highest proportion of excellent ratings, with 36% rated as excellent, 53% requiring minimal clarification, 10% requiring moderate clarification, and 1% categorized as unsatisfactory. no statistically significant differences were observed between the models in terms of overall performance (Fig. 2).

For responses that were not rated as "excellent," raters were asked to specify the primary reasons for their assessments. Among the predefined categories, "too few information" was the most frequently cited issue across all models, accounting for 54.4% of ratings for Google Gemini, 50% for ChatGPT 3.5, and 58.1% for ChatGPT 4. "Language issues," encompassing unclear phrasing or grammar concerns, were reported at 1.8% for Google Gemini, 35.7% for ChatGPT 3.5, and 18.6% for ChatGPT 4. "Too much information," where responses included excessive and unnecessary detail, was noted in 14% of cases for Google Gemini, 1.8% for ChatGPT 3.5, and 9.3% for ChatGPT 4.

The category "off-topic," referring to irrelevant answers, was rare, occurring in 3.5% of Google Gemini responses and none of the responses from ChatGPT 3.5 or ChatGPT 4. "Clear mistakes," such as factual inaccuracies, were identified in 7% of responses from both Google Gemini and ChatGPT 4, while no instances were noted for ChatGPT 3.5. Finally, "other reasons," covering miscellaneous concerns, accounted for 19.3% of ratings for Google Gemini, 12.5% for ChatGPT 3.5, and 7% for ChatGPT 4 (Fig. 3).

The median overall rating across the ten questions did not differ significantly between Google Gemini, ChatGPT 3.5, and ChatGPT 4 (Fig. 4A). Median scores remained largely consistent across all models, with only slight variations observed for specific questions (Fig. 4B). A closer examination of median ratings by individual FAQs further reinforced this observation, indicating comparable performance among the three chatbots across the evaluated questions (Fig. 4).

The evaluation of exhaustiveness, clarity, empathy/professionalism, and response length showed slight variations

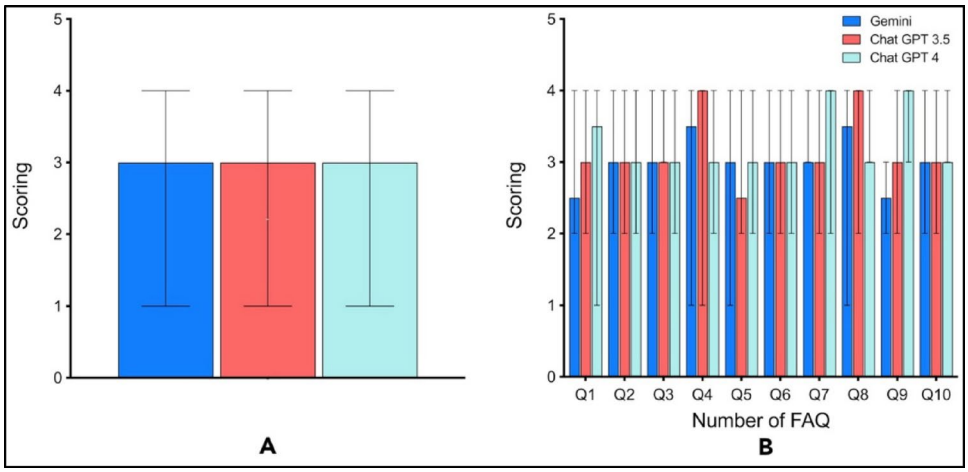

Fig. 4 **A**: Median ratings for the overall quality of responses across Google Gemini, ChatGPT 3.5, and ChatGPT 4. Error bars represent the range (minimum–maximum) of ratings for each model. **B**: Breakdown of median ratings by individual FAQs, illustrating the performance of each chatbot for specific questions

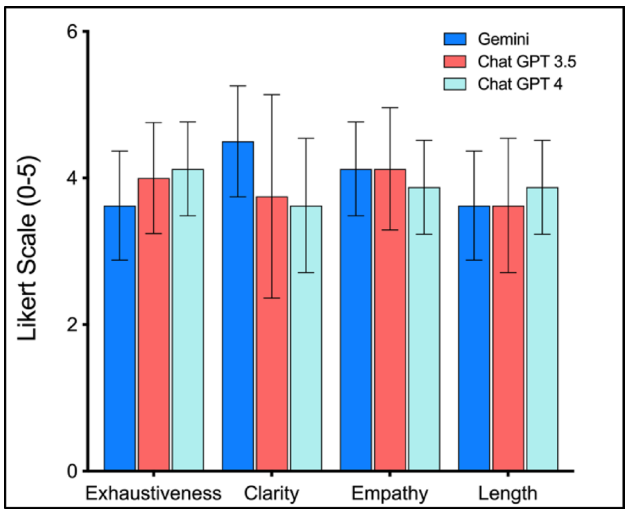

Fig. 5 Mean Ratings with Standard Deviations (SD) for Exhaustiveness, Clarity, Empathy/Professionalism, and Response Length Across Google Gemini, ChatGPT 3.5, and ChatGPT 4

among the three models, but no statistically significant differences were observed. All three chatbots demonstrated similar overall performance, with ratings indicating their ability to provide satisfactory and well-rounded responses. While there were minor differences, such as ChatGPT 4 scoring slightly higher on exhaustiveness and response length, and Google Gemini performing better in clarity, these variations were not significant (Fig. 5).

Discussion

The potential of AI, particularly language models like ChatGPT and Google Gemini, to revolutionize patient education is garnering growing attention in research [5–7, 16–18]. Yet, the broader landscape of online resources poses considerable challenges in healthcare contexts, raising critical concerns about the reliability and accuracy of the information they provide. In light of these considerations, we systematically identified common patient questions about spinal cord injuries and evaluated the responses generated by ChatGPT 4, ChatGPT 3.5, and Google Gemini, based on ratings from blinded experts in spine surgery.

The results demonstrate that all three LLMs achieved satisfactory performance across the majority of evaluated questions, with nearly 90% of ChatGPT 4's responses rated as requiring no clarification or only minimal clarification. Importantly, the differences in overall ratings across the models were not statistically significant, indicating that all three platforms are broadly comparable in their ability to address patient inquiries.

When assessing specific aspects of response quality, slight variations were noted. ChatGPT 4 scored marginally higher in exhaustiveness and response length, suggesting that its additional detail was generally well-received. Conversely, Google Gemini performed better in clarity, indicating that its responses were perceived as easier to understand.

These findings underscore the variability among even the most popular and widely accessible LLMs. It is therefore important to understand that each model has distinct strengths and limitations, even when addressing similar tasks.

Despite these strengths, common challenges were identified across all models. The most prevalent issue was a lack of sufficient information, which accounted for over half of the non-excellent ratings for all three chatbots. This suggests that while the models generally provide relevant information, they may fail to address complex or nuanced aspects of SCI.

The findings align with prior research demonstrating the promise of LLMs in patient education. For example, studies by Ayers et al. and Stroop et al. reported high ratings for accuracy, clarity, and empathy in AI-generated responses, particularly in addressing frequently asked medical questions [19, 20]. However, as noted in earlier research, the risk

of generating plausible but incorrect information ("hallucinations") remains a significant limitation, underscoring the need for expert oversight when applying LLMs in healthcare [21, 22]. Additionally, Lang et al. observed that specific questions related to nuanced or complex topics, such as surgical techniques or long-term outcomes, often receive less satisfactory responses [6, 7]. This aligns with our finding that "too few information" was the most common reason for non-excellent ratings across all three models, particularly for complex inquiries that demand individualized and detailed explanations, as often required in the context of SCI.

One critical factor influencing the quality of LLM-generated responses is the prompt used to guide the models. In this study, a carefully designed prompt instructed the LLMs to act as empathetic, knowledgeable experts in SCI. This was implemented within a standardized 150-word format to ensure comparability of outputs. This approach likely contributed to the high ratings for clarity and professionalism across all models. However, this also highlights the dependence of LLMs on effective prompt engineering, the art of crafting prompts to elicit accurate and contextually appropriate responses. Poorly designed prompts can result in incomplete or irrelevant answers, emphasizing the need for specialized training and expertise when utilizing LLMs in clinical contexts [23, 24].

Our evaluation reflects a predominantly surgical perspective. Responses were rated by board-certified spine surgeons. While this ensures clinical rigor for medical content, it does not capture all perspectives central to everyday life with SCI, such as those of rehabilitation physicians, physiotherapists, occupational therapists, psychologists, nurses, and social workers. These viewpoints may weigh clarity, practicality, empathy, and lived experience differently. Future studies should therefore incorporate multidisciplinary panels to provide a more comprehensive assessment of patient-facing information.

The content accuracy of off-the-shelf LLMs mirrors the unregulated and variable quality of the data from which they are trained. While these general-purpose models can provide a strong starting point, custom LLMs could present an excellent opportunity to integrate specialized, bespoke expertise from specific medical specialties. This customization could ensure greater reliability and relevance in patient education materials. For example, a center-specific LLM trained on outcome data, procedural details, and the expertise of its surgeons could provide tailored, precise, and patient-centered information.

The promising performance of LLMs in this study suggests their potential utility as supplemental tools in patient education. For example, they could provide initial drafts of responses to patient questions, which clinicians can then review and personalize. This hybrid approach could enhance efficiency in patient-physician communication while ensuring accuracy and contextual relevance. Combined with advanced prompt engineering, this underscores the immense potential of AI to transform patient-physician communication and education in highly specialized fields like SCI.

Future research should explore how LLMs perform in real-world scenarios, including comparisons with physician-provided answers. Additionally, integrating advanced prompt engineering techniques, such as chain-of-thought or reasoning-based prompting, may further enhance the quality of LLM-generated responses, particularly for complex or ambiguous inquiries [25]. Research into how patients perceive and interact with AI-generated content will also be crucial for optimizing its role in clinical practice.

Limitations

This study has several limitations that warrant consideration. First, assessor composition was limited to board-certified spine surgeons with expertise in SCI management. Multidisciplinary perspectives (rehabilitation medicine, physiotherapy/occupational therapy, psychology, nursing, social work) and patient raters were not included, which may influence judgments of clarity, practicality, and empathy.

Second, we did not include a comparison with physician-generated answers in this study.

Evaluating LLM responses against a "gold standard" set of physician-generated answers would provide valuable context for interpreting their performance and identifying areas for improvement. Furthermore, the one-question, 150-word setup enabled standardized ratings but differs from typical multi-turn interactions, which may allow more depth and opportunities for correction.

Finally, while we evaluated three widely available general-purpose LLMs, specialized medical LLMs, such as Google's Med-PaLM 2 or ClinicalGPT, were not included. These models may offer more targeted and detailed information for specific medical contexts but are less accessible to the general public. Additionally, it is important to note that the versions of LLMs are continuously evolving, with updates often introducing significant changes that can drastically impact their performance and the quality of their responses. While this variability presents challenges for consistency in evaluation, it also reflects the rapid pace of advancement in this field, underscoring the potential for ongoing improvements in the capabilities of LLMs to support patient education and healthcare delivery.

Conclusion

Within our standardized single turn, 150-word setting, publicly available LLMs produced largely satisfactory answers to common SCI questions with comparable performance across models. LLMs can therefore be recommended as adjuncts for general patient education and pre visit preparation, and to complement it as part of a broader healthcare communication strategy, but not as standalone sources for medical decisions or diagnosis. Outputs should be reviewed within clinical care. Chatbots may be less suitable for complex or highly individualized issues that require tailored clinical judgment. Future research should assess multi turn use, include patient and multidisciplinary evaluators, and benchmark against clinician authored answers, with ongoing reevaluation as models evolve. Specialized medical LLMs may ultimately provide greater reliability for clinical information.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00586-026-09763-x.

Author contributions All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding Open Access funding enabled and organized by Projekt DEAL. We have received no funding for this project.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Anderson KD (2004) Targeting recovery: priorities of the spinal Cord-Injured population. J Neurotrauma Oktober 21(10):1371–1383
2. Manns PJ, Chad KE (2001) Components of quality of life for persons with a Quadriplegic and Paraplegic Spinal Cord Injury
3. Tyrrell Burrus M, Werner BC, Starman JS, Kurkis GM, Pierre JM, Diduch DR (2017) u. a. Patient perceptions and current trends in internet use by orthopedic outpatients. HSS Journal® Musculoskelet J Hosp Spec Surg Oktober 13(3):271–275
4. Ahn C (2023) Exploring chatgpt for information of cardiopulmonary resuscitation. Resuscitation 185:109729
5. Bains SS, Dubin JA, Hameed D, Sax OC, Douglas S, Mont MA (2024) u. a. Use and application of large Language models for patient questions following total knee arthroplasty. J Arthroplasty September 39(9):2289–2294
6. Lang S, Vitale J, Fekete TF, Haschtmann D, Reitmeir R, Ropelato M (2024) u. a. Are large Language models valid tools for patient information on lumbar disc herniation? The spine surgeons' perspective. Brain Spine 4:102804
7. Lang SP, Yoseph ET, Gonzalez-Suarez AD, Kim R, Fatemi P, Wagner K (2024) u. a. Analyzing large Language models' responses to common lumbar spine fusion surgery questions: A comparison between ChatGPT and bard. Neurospine 30 Juni 21(2):633–641
8. Liu J, Wang C, Liu S (2023) Utility of ChatGPT in clinical practice. J Med Internet Res 28 Juni 25:e48568
9. Nwachukwu BU, Varady NH, Allen AA, Dines JS, Altchek DW, Williams RJ (2025) u. a. Currently available large Language models do not provide musculoskeletal treatment recommendations that are concordant with Evidence-Based clinical practice guidelines. Arthrosc J Arthrosc Relat Surg Februar 41(2):263–275e6
10. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA (2023) u. a. Reliability of medical information provided by chatgpt: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 30 Juni 25:e47479
11. Howell MD (2024) Generative artificial intelligence, patient safety and healthcare quality: a review. BMJ Qual Saf 33(11):748–54
12. Wang Y, Wu Q, Wang Y, Wang P (2022) The effects of physicians' communication and empathy ability on Physician–Patient relationship from physicians' and patients' perspectives. J Clin Psychol Med Settings Dezember 29(4):849–860
13. Tan SSL, Goonawardene N (2017) Internet health information seeking and the Patient-Physician relationship: A systematic review. J Med Internet Res 19 Januar 19(1):e9
14. Thom DH, Stanford Trust Study Physicians (2001) Physician behaviors that predict patient trust. J Fam Pract 50(4):323–8
15. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM (2023) Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Jt Surg 4 Oktober 105(19):1519–1526
16. Artioli E, Veronesi F, Mazzotti A, Brogini S, Zielli SO, Giavaresi G (2025) u. a. Assessing ChatGPT responses to common patient questions regarding total ankle arthroplasty. J Exp Orthop Januar 12(1):e70138
17. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS (2023) Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med Dezember 19(12):1989–1995
18. Gibson D, Jackson S, Shanmugasundaram R, Seth I, Siu A, Ahmadi N (2024) u. a. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. J Med Internet Res 14 August 26:e55939

19. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB (2023) u. a. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 1 Juni 183(6):589
20. Stroop A, Stroop T, Zawy Alsofy S, Nakamura M, Möllmann F, Greiner C (2024) u. a. Large Language models: are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? Eur Spine J November 33(11):4135–4143
21. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K (2024) u. a. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inf 31 Juli 12:e54345
22. Zhang P, Shi J, Kamel Boulos MN (2024) Generative AI in medicine and healthcare: moving beyond the 'Peak of inflated expectations'. Future Internet 9 Dezember 16(12):462
23. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C (2024) Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res 10 September 26:e60501
24. Heston T, Khun C (2023) Prompt engineering in medical education. Int Med Educ 31 August 2(3):198–205
25. Ott S, Hebenstreit K, Liévin V, Hother CE, Moradi M, Mayrhauser M (2023) u. a. ThoughtSource: A central hub for large Language model reasoning data. Sci Data 8 August 10(1):528

Authors and Affiliations

Jonas Krueckel[1,2] · Melanie Ardelt[3] · David Schiffelholz[1] · Josina Straub[1] · Sebastian Siller[4] · Vanessa Hubertus[5] · Sonja Häckel[6] · Denis Bratelj[7] · Christof Wutte[8,9] · Helena Arias[1,10] · Franz Hilber[2] · Volker Alt[1] · Siegmund Lang[1]

✉ Jonas Krueckel
jonas.krueckel@ukr.de

Melanie Ardelt
schindler.melanie@gmx.at

David Schiffelholz
d-schiffelholz@gmx.de

Josina Straub
Josina.Straub@klinik.uni-regensburg.de

Sebastian Siller
Sebastian.Siller@klinik.uni-regensburg.de

Vanessa Hubertus
vanessa.hubertus@charite.de

Sonja Häckel
sonja.haeckel@insel.ch

Denis Bratelj
denis.bratelj@paraplegie.ch

Christof Wutte
Christof.Wutte@bgu-murnau.de

Helena Arias
Helena.Arias@BGU-Frankfurt.de

Franz Hilber
hilber@sporthopaedicum.de

Volker Alt
Volker.Alt@klinik.uni-regensburg.de

Siegmund Lang
Siegmund.Lang@klinik.uni-regensburg.de

1 Department of Trauma Surgery, University Medical Centre Regensburg, Regensburg, Germany

2 Sporthopaedicum Regensburg, Regensburg, Germany

3 Division of Orthopaedics and Traumatology, University Hospital Krems, Karl Landsteiner University of Health Sciences, Krems, Austria

4 Department of Neurosurgery, University Medical Centre Regensburg, Regensburg, Germany

5 Department of Neurosurgery, Charité University Medicine, Berlin, Germany

6 Department of Orthopaedic Surgery and Traumatology, University Hospital Bern, Bern, Switzerland

7 Spine and Orthopedic Surgery, Swiss Paraplegic Center, Nottwil, Switzerland

8 Spinal Cord Injury Center, Traumacenter Murnau, Murnau am Staffelsee, Germany

9 Institute for Biomechanics of Traumacenter Murnau and PMU Salzburg, Murnau am Staffelsee, Germany

10 Department for Spinal Surgery and Neurotraumatology, BG Trauma Center Frankfurt am Main, Frankfurt am Main, Germany