Brain and Spine 4 (2024) 102804

Contents lists available at ScienceDirect

Brain and Spine

journal homepage: www.journals.elsevier.com/brain-and-spine

Are large language models valid tools for patient information on lumbar disc herniation? The spine surgeons' perspective

Siegmund Lang [a,*], Jacopo Vitale [b], Tamás F. Fekete [b], Daniel Haschtmann [b], Raluca Reitmeir [b], Mario Ropelato [b], Jani Puhakka [b], Fabio Galbusera [b], Markus Loibl [b]

[a] Department of Trauma Surgery, University Hospital Regensburg, Regensburg, Germany
[b] Spine Center, Schulthess Klinik, Zurich, Switzerland

ARTICLE INFO

Handling Editor: Prof F Kandziora

Keywords:
Lumbar disc herniation
Patient education
Large language models
ChatGPT
Google bard
AI evaluation

ABSTRACT

Introduction: Generative AI is revolutionizing patient education in healthcare, particularly through chatbots that offer personalized, clear medical information. Reliability and accuracy are vital in AI-driven patient education.
Research question: How effective are Large Language Models (LLM), such as ChatGPT and Google Bard, in delivering accurate and understandable patient education on lumbar disc herniation?
Material and methods: Ten Frequently Asked Questions about lumbar disc herniation were selected from 133 questions and were submitted to three LLMs. Six experienced spine surgeons rated the responses on a scale from "excellent" to "unsatisfactory," and evaluated the answers for exhaustiveness, clarity, empathy, and length. Statistical analysis involved Fleiss Kappa, Chi-square, and Friedman tests.
Results: Out of the responses, 27.2% were excellent, 43.9% satisfactory with minimal clarification, 18.3% satisfactory with moderate clarification, and 10.6% unsatisfactory. There were no significant differences in overall ratings among the LLMs ($p = 0.90$); however, inter-rater reliability was not achieved, and large differences among raters were detected in the distribution of answer frequencies. Overall, ratings varied among the 10 answers ($p = 0.043$). The average ratings for exhaustiveness, clarity, empathy, and length were above 3.5/5.
Discussion and conclusion: LLMs show potential in patient education for lumbar spine surgery, with generally positive feedback from evaluators. The new EU AI Act, enforcing strict regulation on AI systems, highlights the need for rigorous oversight in medical contexts. In the current study, the variability in evaluations and occasional inaccuracies underline the need for continuous improvement. Future research should involve more advanced models to enhance patient-physician communication.

1. Introduction

Generative artificial intelligence (AI) is reshaping patient education, offering personalized, understandable content that demystifies complex medical conditions (Topol, 2019; Rajpurkar et al., 2022). In healthcare, AI-driven platforms, including chatbots, are increasingly employed to provide real-time, tailored patient education, a trend that is gaining momentum (Bickmore et al., 2018; Crook et al., 2023). AI and machine learning in healthcare can improve decision-making at the patient level by enabling quantification and communication of uncertainty, engendering trust, and providing safeguards against known failure modes (Kompa et al., 2021). The rapid evolution of medical knowledge and AI capabilities necessitates stringent quality control over AI-generated patient information. This is particularly true as AI systems are tasked with interpreting the latest research findings and guidelines into patient-friendly language. It's important to acknowledge that general AI tools are not specifically designed for medical advice and typically include disclaimers against using them for direct health-related inquiries. The new EU AI Act, enforcing strict regulation on AI systems, highlights the need for rigorous oversight in medical contexts (EU AI Act: European Parliament and Council Reach Agreement). Despite these regulations, patients might still consult these AI models for personal health issues, underscoring the importance of clear guidelines and caution in the use of AI for healthcare advice. It's important to evaluate

* Corresponding author. Department of Trauma Surgery, University Hospital Regensburg, Franz-Joseph-Strauss-Allee 11, 93053, Regensburg, Germany.
E-mail addresses: Siegmund.lang@ukr.de (S. Lang), jacopo.vitale@kws.ch (J. Vitale), Tamas.Fekete@kws.ch (T.F. Fekete), daniel.haschtmann@kws.ch (D. Haschtmann), raluca.reitmeir@kws.ch (R. Reitmeir), Mario.Ropelato@kws.ch (M. Ropelato), jani.puhakka@kws.ch (J. Puhakka), Fabio.Galbusera@kws.ch (F. Galbusera), markus.loibl@kws.ch (M. Loibl).

https://doi.org/10.1016/j.bas.2024.102804
Received 27 December 2023; Received in revised form 19 February 2024; Accepted 4 April 2024
Available online 6 April 2024

bias and fairness. LLMs trained on diverse data can still produce biased responses and continuous attempts have to be undertaken to censor such biased outputs and to finetune training to minimize them (Perlis and Fihn, 2023; Ghassemi et al., 2019). Monteith et al. pointed out that recent advancements in LLMs can be potentially misleading and patients might encounter inaccurate online information (Monteith et al., 2024). If these tools are to be integrated into patient education, it's essential that they not only adhere to regulatory guidelines but also consistently deliver information that is accurate, transparent, reliable, and easily understandable by patients. (Matheny et al., 2020). The development of publicly available chatbots like ChatGPT has significant implications, particularly in private patient education. ChatGPT, a dialogue-based chatbot utilizing the GPT-3.5 and subsequently the GPT-4 large language model (LLM), was released in November 2022 and has since been utilized in various fields due to its ability to provide detailed, human-like responses (Chow et al., 2023).

Lumbar disc herniation (LDH) plays a significant role in public health due to its prevalence and substantial impact on patients' quality of life. LDH is one of the most common causes of functional disability and is particularly prevalent in the age groups of 40–49 and 50–59 years, often affecting the L3/L4 and L4/L5 levels of the spine (Prevalence of Lumbar Disk Herniation in Adult Patients with Low Back Pain Based in Magnetic Resonance Imaging Diagnosis, 2023). This condition leads to various degrees of physical, mental, economic, and social challenges for patients, severely impacting their quality of life. However, the long-term impact of LDH, especially post-surgery, shows that patients generally report a quality of life comparable to the age and gender-matched general population (Roiha et al., 2023). Importantly, cognitive and psychological factors play a crucial role in the disability and quality of life associated with LDH (Engel-Yeger et al., 2018). Factors such as pain catastrophizing (exaggerated negative thoughts about pain) and unfulfilled psychological needs like autonomy, competence, and relatedness can influence pain perception and overall quality of life (Ionescu et al., 2023). Therefore, patient information and education about LDH are critical. Educating patients about the nature of their condition, and treatment options, and managing expectations can help in better managing the pain and disability associated with LDH, ultimately leading to improved outcomes (Ionescu et al., 2023). A plethora of treatment options and recommendations may exacerbate patient insecurity. The complexity of treating LDH arises from the variety of its manifestations and the need to tailor treatment to individual patient needs. Treatment options, spanning from conservative measures to a variety of surgical procedures, lack a unanimous consensus due to their diversity (Wan et al., 2022).

By leveraging LLM, chatbots have been anticipated to close the information gap in patient education for various health conditions. These advanced models may be capable of providing intelligently tailored, understandable, and up-to-date medical insights, helping to clarify the diverse treatment options, and alleviating patient insecurity. Therefore, this study aims to answer the research question: "How do three publicly available LLM-based chatbots perform in responding to common questions on lumbar disc herniation?"

2. Methods

2.1. Identification of relevant FAQs

To identify frequently asked questions (FAQs) of general patient interest, a comprehensive Google search was conducted using the search term: '*FAQ OR frequently asked questions AND Lumbar AND Disc herniation OR herniated disc OR Sequestrectomy OR Laminotomy OR Discectomy OR Surgery*' yielding approximately 800.000 results within 0.55 s (August 25th, 2023; region: USA). For this study, the first 20 Google hits were checked, and the following inclusion and exclusion criteria were applied (Table 1).

The search results were screened by the authors using these criteria, and from the array of sources available, questions from eleven institutions (Suppl. Material 1) were used to define the most recurrent FAQs. In addition, ChatGPT-4 was directly engaged with the prompt: '*Please create a list of 20 common questions that patients are likely to ask about lumbar herniated discs. Prioritize questions that are specifically related to surgical treatment options.*' to generate a list of questions relevant to our study. ChatGPT-4 was employed for the methodology due to its increased capacity, while ChatGPT-3.5 was utilized to answer patient questions, as it was deemed important to use a publicly available and popular chatbot for practical application.

Table 1
Inclusion and exclusion criteria for questions.

Inclusion criteria	Exclusion criteria
Published after January 1st, 2017	Non-generalizable information e.g., provider or implant-specific details
Published in English language	Emphasis on non-spine-surgical aspects, e.g., anesthesiologic information
Information presented in FAQ or Q&A sections	

This two-step approach resulted in a consolidated pool of 133 questions. From this, a ranking was derived, highlighting the top 10 most frequently recurring topics (Table 2). The authors then carefully reviewed this ranked list of topics and subsequently, they crafted new questions, synthesizing the essence of these 10 identified topics, resulting in a list of 10 questions to be presented to the LLMs. (Table 3). In instances of discord, the authors collaboratively agreed on a consensus in the formulation of the final question set.

Following this, the questions were submitted to the publicly accessible AI chatbot ChatGPT-3.5 through its online portal (https://chat.openai.com/chat) on September 8th, 2023 (**Answer Set #1**). Second, the questions were relayed to ChatGPT-3.5, with the subsequent prompt used before each question (**Answer Set #2**):

> "*Act as an expert spine surgeon who is up to date with the latest scientific research and has years of experience counseling patients with empathy and clarity. Provide comprehensive and easily understandable answers to the following question about disc herniation and disc herniation surgery! Ensure the responses are timely, incorporate the most recent advancements, and address potential concerns patients might have. Limit your answer to 150 words and focus on the most important aspects to ensure patient information*: (…)".

Third, the identical questions were presented to Google's chatbot "Bard" without prompting (https://bard.google.com/chat) on the same date (**Answer Set #3**). For each question, a new window of the respective chatbot was created to avoid any biases from the prior questions. After the answers were generated, they were recorded verbatim in our database.

2.2. Raters and rating of LLM responses

The answer set for each LLM was provided to the raters using the online Google Forms application. Each response was subjected to a rigorous evaluation by six independent raters from the same Institution (Schulthess Klinik, Zürich, Switzerland), each with extensive experience in spine surgery. A previously published rating system was instituted to categorize the responses as "excellent response not requiring clarification," "satisfactory requiring minimal clarification," "satisfactory requiring moderate clarification," or "unsatisfactory requiring substantial clarification" (Mika et al., 2023). In the case of unsatisfaction (partial or total) with the answers, the raters were asked to identify the reason among one of these categories: 1) "Off topic, the answer is not pertinent to the question", 2) "Clear mistakes in the answer", 3) "Too much information, not all necessary", 4) "Too few information, not enough for an exhaustive answer.", 5) "Other reasons". The evaluative

Table 2
Top 10 recurrent topics about LDH.

Ranking	Topic and sample Questions	Frequency
1	**Post-Operative Care** Is the post-operative period painful? What is the post-operative time until I am recovered? What is the discectomy recovery process like?	16
2	**Surgical Techniques** How many times have you done this lumbar discectomy procedure? What is the surgical procedure? Can you provide me a step-by-step description of the discectomy procedure?	13
3	**Clinical Manifestations/Symptoms** What are the symptoms of a herniated disc? Which are the most frequent symptoms? What are the symptoms of a herniated disc in the lumbar spine?	12
4	**Indications for Surgery** When is Spine Surgery an Option for Pain Relief from a Disc Herniation? When should a patient have surgery? When is surgery recommended for a lumbar herniated disc?	12
5	**Underlying Mechanisms/Pathomechanism** Why Does a Disc Herniation Cause Back Pain? What causes it? What causes a herniated disc?	11
6	**Therapeutic Options** Have we exhausted all appropriate nonsurgical treatment options? What are the non-surgical treatment options for a lumbar herniated disc? What is the treatment for lumbar disc herniation?	10
7	**Definition/Terminology** What is a lumbar disc herniation? What is lumbar disc herniation or slipped disc or prolapsed intervertebral disc? What is a herniated disc?	8
8	**Post- Diagnostic Measures** How is a lumbar disc herniation diagnosed? How is a lumbar herniated disc diagnosed? How are slipped discs diagnosed?	8
9	**Surgical Outcomes and Prognosis** How successful is back surgery for herniated discs? What are the overall results & outcome of surgery done for slipped disc? What is the recurrence rate of herniation after surgical treatment?	8
10	**Demographics, Risk Factors, and Contributing Factors** Who belongs to the risk group for a lumbar disc herniation? What are the risk factors for slipped disc? What increases my risk of having a herniated disc?	6

Table 3
List of ten Questions presented to the LLMs.

1	What should I expect during the recovery period after lumbar disc herniation surgery?
2	Can you explain the different surgical techniques used to treat lumbar disc herniation?
3	What are the most common symptoms of lumbar disc herniation, and how are they different from other back issues?
4	Under what circumstances should I consider surgery for lumbar disc herniation?
5	What causes lumbar disc herniation to occur in the first place?
6	What are the various treatment options for managing lumbar disc herniation without surgery?
7	Could you clarify the medical terminology associated with lumbar disc herniation?
8	What diagnostic tests are recommended to confirm lumbar disc herniation?
9	What are the success rates and long-term outcomes for surgery on lumbar disc herniation?
10	Who is most at risk for developing lumbar disc herniation, and what lifestyle factors contribute?

framework was augmented with the subsequent four inquiries (Table 4), wherein participants were provided with a 5-point Likert scale extending from 'I strongly disagree' to 'I strongly agree'. The participants were asked to answer these four inquiries referring to each of the three answer sets.

Finally, the raters were presented with seven inquiries aimed at eliciting their preference for the best set of three responses, followed by additional questions designed to collect their general perspective on the utilization of AI tools in patient care (Table 5). A 5-point Likert scale has been used to answer these questions.

2.3. Statistical analysis

Data are presented using absolute values, percentages, mean, and standard deviations (SD) for descriptive purposes. The interrater reliability was assessed using Fleiss Kappa. Chi-square tests ($\chi 2$) were applied to test differences in ratings among LLMs, raters, and questions. A Friedman test was applied to test differences among LLMS in exhaustiveness, clarity, empathy, and length. All statistical procedures were performed using GraphPad Prism 9.5.1. The level of statistical significance was set at $p < 0.05$.

3. Results

The distribution of ratings for the combined question set across all three models is as follows: 10.6% (n = 19) of responses were rated as "unsatisfactory requiring substantial clarification", 18.3% (n = 33) as "satisfactory requiring moderate clarification", 43.9% (n = 79) as "satisfactory requiring minimal clarification", and 27.2% (n = 49) as "excellent" responses not requiring any clarification (Fig. 1).

For ChatGPT-3.5, 8% of responses were rated as "unsatisfactory requiring substantial clarification", 18% as "satisfactory requiring moderate clarification", 45% as "satisfactory requiring minimal clarification", and 28% were rated as "excellent" not requiring any clarification. The prompted version of ChatGPT-3.5 had 13% of responses rated as "unsatisfactory requiring substantial clarification", 17% as "satisfactory requiring moderate clarification", 45% as "satisfactory requiring minimal clarification", and 25% as "excellent", not requiring any clarification. Bard had 10% of responses rated as "unsatisfactory requiring substantial clarification", 20% as "satisfactory requiring moderate clarification", 42% as "satisfactory requiring minimal clarification", and a similar proportion of 28% rated as "excellent", not requiring any clarification. No statistically significant differences among LLMs were detected in the overall rating ($p = 0.90$; Fig. 2). Overall, the reasons for unsatisfaction with the answers were: 1.4% off-topic, 36.4% clear mistakes, 6.8% too much information, 28.4% too few information, and 27.0% other reasons.

The $\chi 2$ test highlighted a significant difference in the distribution of ratings among questions ($p = 0.043$, Fig. 3). Overall, Q1 and Q4 both have the highest rate of "excellent" or "minimal clarification required" at 89% each. Q7 and Q9 have the next highest "excellent" or "minimal clarification required" rates at 78%. Q5, Q8, and Q10 all share the same distribution with "excellent" or "minimal clarification required" rates at 72%. Q2 has the highest "unsatisfactory" or "moderate clarification" rate at 61%. In detail, keeping each LLM separate, the answer to Q8 received the best ratings for ChatGPT-3.5 and ChatGPT-3.5 prompted

Table 4
Subsequent inquiries for the Likert scale rating.

The overall content of all answers is comprehensive and covers all necessary aspects.	The answers are easy to understand and are communicated clearly.	The answers address patient concerns empathetically and professionally.	The overall length and detail of each answer are appropriate for the target audience.

Table 5
Questions for final evaluation and general opinion.

<!–Col Count:1–>1. In your opinion, which of the above 3 sets contained the highest quality answers and answered the 10 FAQs in the most appropriate and professional way?
2. In general, have the above responses met your expectations of the performance of currently available LLMs?
3. Based on your experience with the scored responses above, would you consider integrating LLM or AI-based patient information into any aspect of your clinical practice in the future?
4. In your opinion, how could the utilization of LLMs improve the patient experience, especially in streamlining the information process before and after surgical procedures?
5. Do you think the integration of LLMs in healthcare could alleviate some of the workload on medical staff, particularly in providing initial information to patients?
6. How do you foresee the role of AI/LLMs in optimizing the patient-physician relationship and communication, particularly in ensuring patients are well-informed and prepared for their surgical procedures?
7. What is your general attitude toward the development of AI/LLMs in healthcare?

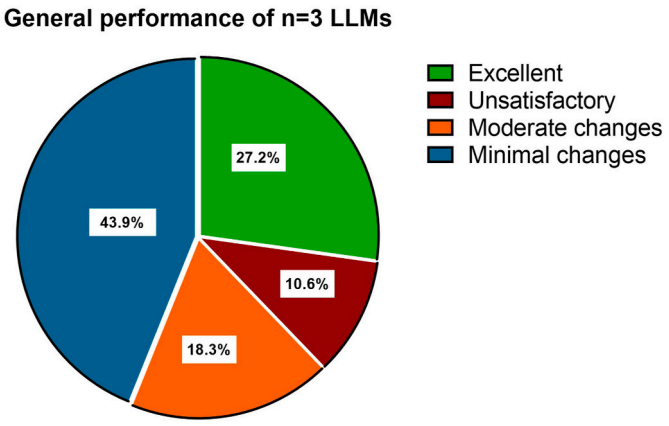

Fig. 1. Pie chart with the distribution of overall ratings, expressed in percentages, for the combined question set across all three LLMs.

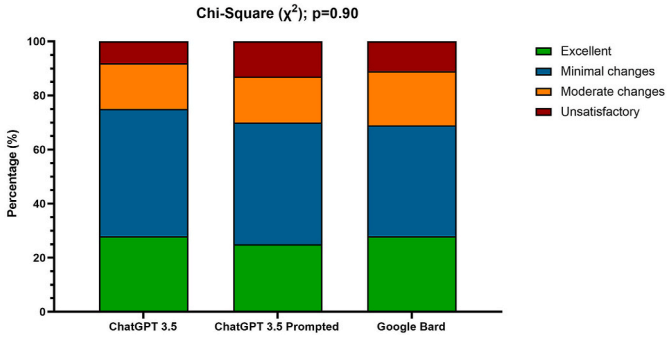

Fig. 2. Histograms with the rating distribution, expressed in percentages, for ChatGPT-3.5, ChatGPT-3.5 prompted, and Google Bard. The χ2 highlighted no differences among LLMs.

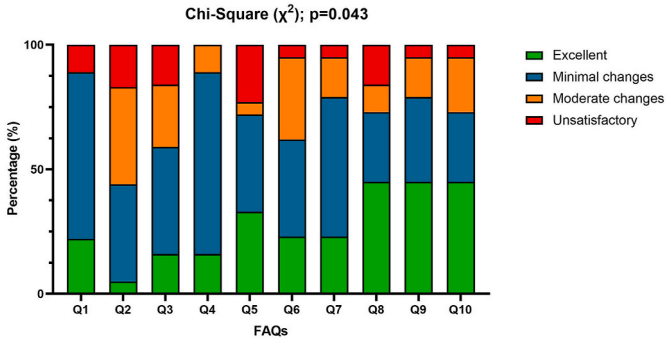

Fig. 3. Histograms with the ratings distribution, expressed in percentages, for each FAQ, from Q1 to Q10. The χ2 test highlighted a significant difference ($p = 0.043$) in the rating distribution among questions.

whereas answers to Q7, Q9, and Q10 were the best rated for Google Bard.

Exhaustiveness, clarity, empathy, and length of the answers were rated >3.5 on average however the Friedman test did not show any significant differences among LLMs (Fig. 4).

Nevertheless, inter-rater reliability was not achieved, and large differences among raters were detected in the frequency of answers distribution for ChatGPT-3.5 ($p = 0.023$), prompted ChatGPT-3.5 ($p < 0.0001$) and Google Bard ($p = 0.007$), separately and mixed ($p < 0.0001$; Fig. 5).

The final ratings of the additional questions reported in Table 5, are shown in Table 6. Overall, the scores by raters were always >4.0. Three raters indicated ChatGPT-3.5 as the best LLM, two raters preferred Google Bard and only one rater indicated prompted ChatGPT-3.5 as the best LLM.

4. Discussion

The use of LLMs is gaining attention for its potential in patient education. Recent research has started to shed light on how LLMs perform in various medical domains, particularly in patient education (PatientEngagementHIT, 2023; Hornung et al., 2022; Samaan et al., 2023). Patients complement doctors' advice with various sources, notably online platforms like social media, health websites, forums, and blogs, which allow for information sharing and questions (Daraz et al., 2019; McMullan, 2006). However, the varying reliability of these online resources poses challenges for their use in health-related contexts. It's essential to note that websites from scientific societies and professional organizations are often more reliable, as they are continuously revised and updated by experts, providing a more dependable source for health-related information. In our present study, we systematically identified typical questions from patients regarding lumbar spine disc herniation and assessed the responses provided by ChatGPT-3.5, ChatGPT-3.5 using a prompt and Google Bard. These responses were evaluated based on ratings from experienced spine surgeons.

A combined 71.1% of responses from the three evaluated LLMs were deemed either "excellent" or "satisfactory with minimal changes required." This high rate of satisfactory responses underscores the potential of LLMs in providing quality initial information in patient education contexts. Ayers et al. found that patients generally preferred responses from AI chatbots over those from physicians, citing higher quality and empathy in the AI responses (Ayers et al., 2023). A current study by Stoop et al. evaluated the effectiveness of ChatGPT in providing information, particularly in the context of LDH (Stroop et al., 2023). In their study ChatGPT's responses were largely understandable (97%), specific (86% satisfactory), and medically correct (96%) (Stroop et al., 2023). This aligns with the high ratings in our study. However, Stoop et al. also noted isolated instances of incorrect information, highlighting the need for caution and oversight when using AI for medical advice. In line The EU's AI Act, effective from 2026, will mandate strict regulations for high-risk AI systems, reinforcing the need for expert oversight in AI use (EU AI Act: European Parliament and Council Reach Agreement). This resonates with the current study's findings where a portion of responses required clarification, underlining the importance of expert review in the use of AI tools. Our study, however, focused exclusively on

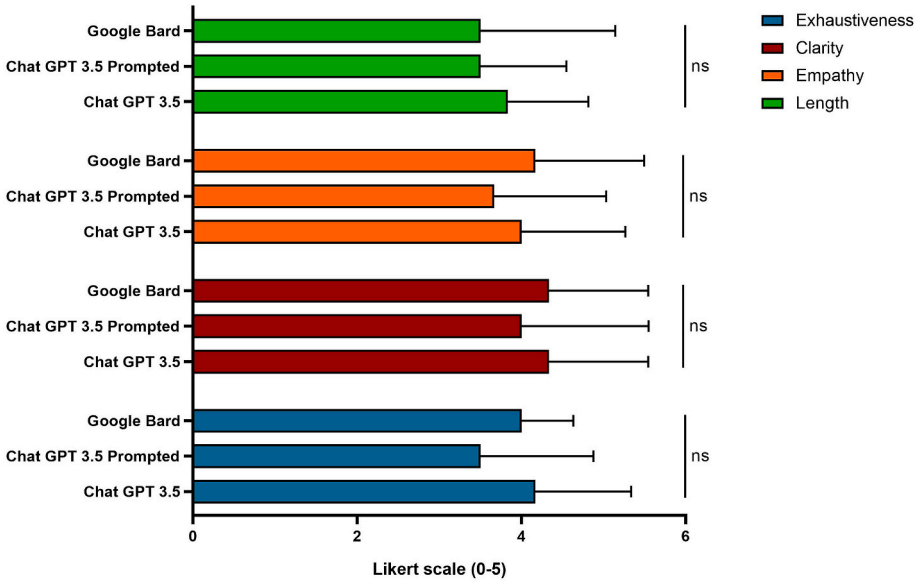

Fig. 4. Histograms with mean and SD for the scores reported by raters, on a Likert scale from 1 to 5, of exhaustiveness, clarity, empathy, and length of the answers. The Friedman test did not show any significant differences among LLMs. *Legend*: ns, non significant.

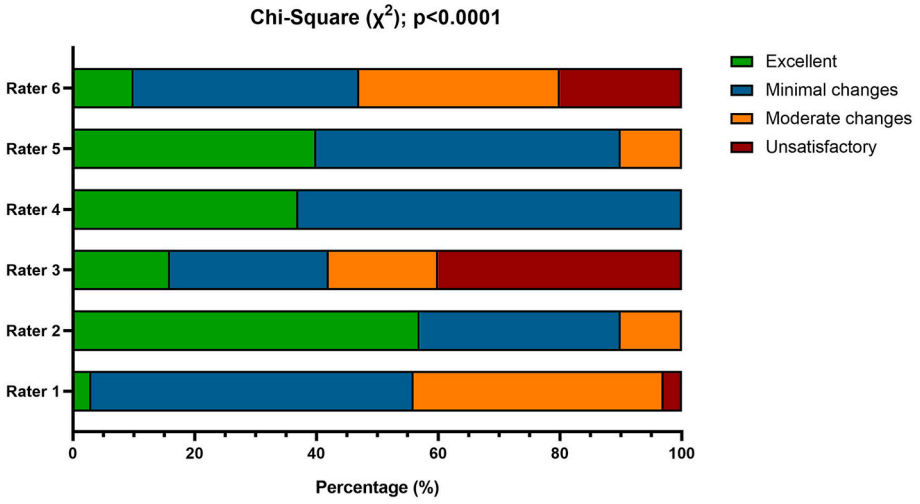

Fig. 5. Histograms with the rating distribution, expressed in percentages, for each rater. The χ2 test highlighted a significant difference ($p < 0.0001$) in the rating distribution among raters.

Table 6
Ratings for the questions on the final evaluation and general opinion by raters.

Q1	n = 3: ChatGPT 3.5 n = 2: Google Bard n = 1 ChatGPT 3.5
Q2	4.2 ± 0.4
Q3	4.2 ± 0.7
Q4	4.0 ± 0.6
Q5	4.2 ± 0.7
Q6	4.3 ± 0.5
Q7	4.8 ± 0.4

Data are reported as mean ± SDs.

evaluating the precision and dependability of responses from LLMs, without comparing them to physician-provided answers. The comparison with physicians' recommendations but also the investigation of AI-aided patient-physician communication has to be elucidated in future studies. The lack of inter-rater reliability in this study, indicated by significant differences in how raters evaluated the responses from ChatGPT-3.5, its prompted version, and Google Bard, suggests a subjective element in assessing AI-generated answers. Such variability can stem from differing expectations, experiences, or understanding of the subject matter among raters. The subjective nature of rater evaluations could also play a role. Given the variability in how individual raters perceive and score AI-generated responses, it's possible that biases or preconceived notions about AI capabilities influenced their ratings, leading to a lack of significant differences among the LLMs. This inconsistency in evaluation poses a challenge for objectively measuring the performance of LLMs, highlighting the need for more standardized criteria or automated methods for assessing AI response quality in future studies.

In our survey, the questions on postoperative care (Q1) and indications for LDH surgery (Q4) received the best ratings with "excellent" or "minimal clarification required" at 89% each. LDH and its surgical treatment are common topics in medical education and patient inquiries. This prevalence might mean that LLMs are more frequently

exposed to such content in their training data, resulting in better responses (Pearson et al., 2012; Kögl et al., 2021). Further questions that received good ratings were Q7 on the clarification of the terminology around LDH and Q9 regarding the success rate of LDH surgery. While Q7's answer contained basic, well-defined information, Q9's diversity makes the high-quality results from all three LLMs notable. ChatGPT-3.5 reported a success rate of "80–90%", its prompted version "about 90%", and Bard "80–90%". This is in line with meta-analyses, that find the success rates of different surgical techniques for LDH to be as high as 90% (Bai et al., 2021; Feng et al., 2017). Bard described microdiscectomy as "the most common type of surgery for LDH." Only the prompted ChatGPT-3.5 failed to elaborate on long-term follow-ups, and Bard did not emphasize individual patient factors adequately (Suppl. Material 2). The finding that Q2, which inquired about various surgical techniques for LDH, had the highest rate of "unsatisfactory"or "moderate clarification" responses may be attributable to the complexity and diversity of the surgical options available. For instance, a study by Wei et al. highlights the multitude of eight surgical interventions for LDH, and this network meta-analysis of 27 randomized controlled trials emphasized the differences in complications, operation time, and blood loss among these procedures (Wei et al., 2021).

Notably, we did not detect a statistically significant difference in the ratings between the three models despite the use of prompting in ChatGPT-3.5. This finding can be interpreted through several lenses: The high scores across the board suggest a potential ceiling effect, where the quality of responses was already at a high level, leaving little room for noticeable improvement through prompting. This aligns with the observed high rate of satisfactory responses, indicating that even baseline LLM performance was adequate for most inquiries. The role of prompting in LLMs in general has been reported to be crucial. It involves instructing or questioning the model, a process that can enhance response quality and reasoning capabilities (Raheja, 2023). However, this is not always straightforward. LLMs can sometimes generate plausible but incorrect information ("hallucination") or struggle with complex reasoning, impacting the effectiveness of prompted responses (Large language models are zero). Advanced prompting strategies like Chain-of-Thought, Self-Consistency, Least-to-Most prompting, Tree of Thoughts, and Reasoning via Planning have been developed to address these challenges and improve LLM performance in complex tasks (Ott et al., 2023). As outlined in a survey on language model prompting, the revolutionary development of pre-training and scaling up LLMs has endowed these models with a range of reasoning abilities, which can be further enhanced by prompting strategies (Reasoning with Language Model Prompting). Chain-of-thought prompting, proposed by Wei et al., involves adding intermediate reasoning steps into few-shot prompts, effectively guiding LLMs to generate a reasoning process before answering (Reasoning with Language Model Prompting). This technique has proven to enhance the performance of LLMs, particularly in tasks requiring complex reasoning. It must be mentioned that some patients, seeking straightforward information might be overwhelmed by the complexity of effective prompting. The impact of prompting on LLM responses may be more significant for complex questions than for the FAQs in our study, suggesting that the absence of substantial differences in our results could be due to the relatively simple nature of the queries examined. Anyhow, the scientific literature on the role of prompting in LLMs for patient education remains sparse, necessitating further research.

In the final evaluation of the current survey, raters provided their opinions on the use of LLMs in healthcare, specifically in the context of patient information and physician-patient communication. The ratings (Likert scale 1–5, with 5 being the best) were generally positive, with scores above 4.0 for all questions, indicating a favorable view of LLMs' potential in clinical practice. The Ethics Guidelines for Trustworthy AI emphasize that AI, including LLMs used in healthcare, must be lawful, ethical, and technically robust, with a focus on human agency, safety, privacy, transparency, and accountability (Ethics guidelines for trustworthy AI, 2019). These principles, as outlined in the final Assessment List for Trustworthy AI (ALTAI), provide a crucial framework for ensuring that AI systems in clinical settings are not only accurate but also adhere to broader ethical and societal standards (Ethics guidelines for trustworthy AI, 2019).

Questions regarding the integration of LLMs in healthcare, their role in improving patient experience, and alleviating medical staff workload received scores around 4.2, suggesting strong support for their use. The highest score (4.8) was for the general attitude towards AI/LLMs development in healthcare, reflecting a very positive outlook on the future role of AI in this field. In the study by Stroop et al., 88% of respondents believed patients would use ChatGPT to learn about their health conditions, with 58% considering it useful for enhancing patient information. However, opinions were mixed regarding its impact on doctor-patient communication and the informed consent process, with 63% viewing it positively for conversations, 42% expecting it to shorten informed consent times, but a significant proportion (46% and 42%, respectively) seeing no effect or remaining undecided (Stroop et al., 2023). A survey of 114 physicians across various specialties by Al-Medfa et al. revealed a generally positive attitude towards AI in clinical practice, with specific support for its role in diagnosis and patient care planning. Concerns were noted about AI's impact on employment, regardless of the respondents' demographics or AI knowledge (Al-Medfa et al., 2023). In a cross-sectional study by Fritsch et al., involving 452 patients, over 90% were aware of AI, but only 24% had good knowledge of it. Most viewed AI in healthcare positively yet emphasized the need for physician oversight, with older patients and those with less education being more cautious about AI's role in healthcare (Fritsch et al., 2022). In summary, there's strong support for AI and LLMs' role in enhancing patient information, diagnosis, and care planning. Our study suggests that patients seeking information about lumbar disc herniation can expect valid information provided by publicly available LLMs.

In a potential clinical practice scenario, LLM-driven patient education on LDH could involve personalized digital materials explaining causes, symptoms, treatments, and self-management strategies. This approach could enhance patient understanding, engagement, and decision-making, fostering collaborative relationships and improving healthcare outcomes. However, concerns about AI's impact on employment and the necessity for physician oversight are prevalent, highlighting the need for a balanced integration of AI in clinical practice and strict adherence to regulations and ethical guidelines (EU AI Act: European Parliament and Council Reach Agreement; Ethics guidelines for trustworthy AI, 2019). The integration of LLMs in the daily practice, supplementing the patient-physician communication needs careful adherence to ethical and data security guidelines and faces regulatory and practical challenges, dependent on the individual structure of the medical center (Gottlieb and Silvis, 2023; Clusmann et al., 2023). Ensuring that the-generated content aligns with expert medical knowledge remains a critical and ongoing research challenge, particularly in the sensitive area of healthcare.

5. Limitations

Our study's limitations include the rapid evolution of LLMs, making our analysis potentially outdated due to frequent updates, such as Google's recent integration of the Gemini LLM. Additionally, the transparency of these LLMs is limited, as their answers are refined by human raters, especially on sensitive topics, and this process is not well-documented. It's unclear whether responses are AI-generated or human-adjusted, introducing potential biases. Additionally, the raters' discernible enthusiasm for LLMs in the additional question responses may have influenced their overall ratings, potentially leading to more favorable evaluations. A further limitation of our study is the small number of raters and the lack of a standardized rating method which necessitates cautious interpretation of the statistical analysis. The small number of raters further limits the options of statistical testing to the

methods that have been applied in this study. For future studies, not only a higher rater number and diversity would enrich the analysis, but also a mixed-methods approach, including physicians' answers as ground truth and a qualitative evaluation of LLMs' answers by patients and physicians. Finally, while general-purpose LLMs are widely used, specialized LLMs, such as those for medical applications, offer more targeted information but are less accessible to the public, e.g. ClinicalGPT and Google's Med-PaLM 2.

6. Conclusion

Chatbots based on LLMs like ChatGPT show promise for patient education in lumbar spine surgery, as evidenced by their high ratings for accuracy and relevance in our study. However, the variability in rater evaluations and occasional inaccuracies underscore the need for continuous quality control and training of these models. With fast-evolving models like GPT-4 and Gemini gaining broader availability, future research should include them to further explore their utility in patient-physician communication and patient information.

Generative AI disclosure statement

During the preparation of this work, the authors used ChatGPT 4.0 to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bas.2024.102804.

References

Al-Medfa, M.K., Al-Ansari, A.M.S., Darwish, A.H., Qreeballa, T.A., Jahrami, H., 2023. Physicians' attitudes and knowledge toward artificial intelligence in medicine: benefits and drawbacks. Heliyon 9, e14744. https://doi.org/10.1016/j.heliyon.2023.e14744.

Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., et al., 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern. Med. 183, 589–596. https://doi.org/10.1001/jamainternmed.2023.1838.

Bai, X., Lian, Y., Wang, J., Zhang, H., Jiang, M., Zhang, H., et al., 2021. Percutaneous endoscopic lumbar discectomy compared with other surgeries for lumbar disc herniation. Medicine (Baltim.) 100, e24747. https://doi.org/10.1097/MD.0000000000024747.

Bickmore, T.W., Trinh, H., Olafsson, S., O'Leary, T.K., Asadi, R., Rickles, N.M., et al., 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. J. Med. Internet Res. 20, e11510 https://doi.org/10.2196/11510.

Chow, J.C.L., Sanders, L., Li, K., 2023. Impact of ChatGPT on medical chatbots as a disruptive technology. Front. Artif. Intell. 6, 1166014 https://doi.org/10.3389/frai.2023.1166014.

Clusmann, J., Kolbinger, F.R., Muti, H.S., Carrero, Z.I., Eckardt, J.-N., Laleh, N.G., et al., 2023. The future landscape of large language models in medicine. Commun. Med. 3, 1–8. https://doi.org/10.1038/s43856-023-00370-1.

Crook, B.S., Park, C.N., Hurley, E.T., Richard, M.J., Pidgeon, T.S., 2023. Evaluation of online artificial intelligence-generated information on common hand procedures. J. Hand. Surg. Am. https://doi.org/10.1016/j.jhsa.2023.08.003. S0363-5023(23)00414-8.

Daraz, L., Morrow, A.S., Ponce, O.J., Beuschel, B., Farah, M.H., Katabi, A., et al., 2019. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. J. Gen. Intern. Med. 34, 1884–1891. https://doi.org/10.1007/s11606-019-05109-0.

Engel-Yeger, B., Keren, A., Berkovich, Y., Sarfaty, E., Merom, L., 2018. The role of physical status versus mental status in predicting the quality of life of patients with lumbar disk herniation. Disabil. Rehabil. 40, 302–308. https://doi.org/10.1080/09638288.2016.1253114.

Ethics guidelines for trustworthy AI | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019–. (Accessed 15 December 2023).

EU AI Act: European Parliament and Council Reach Agreement | Perspectives. Events. Mayer Brown. n.d. https://www.mayerbrown.com/en/perspectives-events/publications/2023/12/eu-ai-act-european-parliament-and-council-reach-agreement (accessed December 15, 2023)..

Feng, F., Xu, Q., Yan, F., Xie, Y., Deng, Z., Hu, C., et al., 2017. Comparison of 7 surgical interventions for lumbar disc herniation: a network meta-analysis. Pain Physician 20, E863–E871.

Fritsch, S.J., Blankenheim, A., Wahl, A., Hetfeld, P., Maassen, O., Deffge, S., et al., 2022. Attitudes and perception of artificial intelligence in healthcare: a cross-sectional survey among patients. Digit. Health. 8, 20552076221116772 https://doi.org/10.1177/20552076221116772.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y., Ranganath, R., 2019. Practical guidance on artificial intelligence for health-care data. Lancet. Digit. Health. 1, e157–e159. https://doi.org/10.1016/S2589-7500(19)30084-6.

Gottlieb, S., Silvis, L., 2023. How to safely integrate Large Language Models into health care. JAMA. Health. Forum. 4, e233909 https://doi.org/10.1001/jamahealthforum.2023.3909.

Hornung, A.L., Hornung, C.M., Mallow, G.M., Barajas, J.N., Rush, A., Sayari, A.J., et al., 2022. Artificial intelligence in spine care: current applications and future utility. Eur. Spine J. 31, 2057–2081. https://doi.org/10.1007/s00586-022-07176-0.

Ionescu, D., Iacob, C.I., Brehar, F.M., Avram, E., 2023. The role of catastrophizing and basic psychological needs satisfaction on health-related quality of life and pain in patients with lumbar disc herniation. Front. Psychol. 14.

Kögl, N., Brawanski, K., Girod, P.-P., Petr, O., Thomé, C., 2021. Early surgery determines recovery of motor deficits in lumbar disc herniations—a prospective single-center study. Acta Neurochir. 163, 275–280. https://doi.org/10.1007/s00701-020-04614-0.

Kompa, B., Snoek, J., Beam, A.L., 2021. Second opinion needed: communicating uncertainty in medical machine learning. Npj. Digit. Med. 4, 1–6. https://doi.org/10.1038/s41746-020-00367-3.

Large language models are zero-shot reasoners - insights of a pre-trained language model | Data Science Dojo n.d. https://datasciencedojo.com/blog/llms-zero-shot-reasoners/(accessed December 14, 2023).

Matheny, M.E., Whicher, D., Thadaney Israni, S., 2020. Artificial intelligence in health care: a report from the national academy of medicine. JAMA 323, 509–510. https://doi.org/10.1001/jama.2019.21579.

McMullan, M., 2006. Patients using the Internet to obtain health information: how this affects the patient–health professional relationship. Patient Educ. Counsel. 63, 24–28. https://doi.org/10.1016/j.pec.2005.10.006.

Mika, A.P., Martin, J.R., Engstrom, S.M., Polkowski, G.G., Wilson, J.M., 2023. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J. Bone. Joint. Surg. Am. https://doi.org/10.2106/JBJS.23.00209.

Monteith, S., Glenn, T., Geddes, J.R., Whybrow, P.C., Achtyes, E., Bauer, M., 2024. Artificial intelligence and increasing misinformation. Br. J. Psychiatry 224, 33–35. https://doi.org/10.1192/bjp.2023.136.

Ott, S., Hebenstreit, K., Liévin, V., Hother, C.E., Moradi, M., Mayrhauser, M., et al., 2023. ThoughtSource: a central hub for large language model reasoning data. Sci. Data 10, 528. https://doi.org/10.1038/s41597-023-02433-3.

PatientEngagementHIT, 2023. ChatGPT continues to prove useful for patient education. PatientEngagementHIT. https://patientengagementhit.com/news/chatgpt-continues-to-prove-useful-for-patient-education. (Accessed 5 October 2023).

Pearson, A., Lurie, J., Tosteson, T., Zhao, W., Abdu, W., Mirza, S., et al., 2012. Who should have surgery for an intervertebral disc herniation? Spine 37, 140–149. https://doi.org/10.1097/BRS.0b013e3182276b2b.

Perlis, R.H., Fihn, S.D., 2023. Evaluating the application of Large Language Models in clinical research contexts. JAMA Netw. Open 6, e2335924. https://doi.org/10.1001/jamanetworkopen.2023.35924.

Prevalence of Lumbar Disk Herniation in Adult Patients with Low Back Pain Based in Magnetic Resonance Imaging Diagnosis, 2023. Open. Access. Macedonian J. Med. Sci.

Raheja, T., 2023. Mastering the art of prompting for Large Language Models. Cisco. Tech. Blog. https://techblog.cisco.com/blog/mastering-the-art-of-prompting-for-large-language-models-reducing-hallucination-and-improving-reasoning. (Accessed 14 December 2023).

Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. Nat. Med. 28, 31–38. https://doi.org/10.1038/s41591-021-01614-0.

Reasoning with Language Model prompting: a survey. Ar5iv n.d. https://ar5iv.labs.arxiv.org/html/2212.09597. (Accessed 13 December 2023).

Roiha, M., Marjamaa, J., Siironen, J., Koskinen, S., Koski-Palkén, A., 2023. Favorable long-term health-related quality of life after surgery for lumbar disc herniation in young adult patients. Acta Neurochir. 165, 797–805. https://doi.org/10.1007/s00701-023-05522-9.

Samaan, J.S., Yeo, Y.H., Rajeev, N., Hawley, L., Abel, S., Ng, W.H., et al., 2023. Assessing the accuracy of responses by the Language Model ChatGPT to questions regarding bariatric surgery. Obes. Surg. 33, 1790–1796. https://doi.org/10.1007/s11695-023-06603-5.

Stroop, A., Stroop, T., Zawy Alsofy, S., Nakamura, M., Möllmann, F., Greiner, C., et al., 2023. Large language models: are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? Eur. Spine J. https://doi.org/10.1007/s00586-023-07975-z.

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25, 44–56. https://doi.org/10.1038/s41591-018-0300-7.
Wan, Z.Y., Shan, H., Liu, T.F., Song, F., Zhang, J., Liu, Z.H., et al., 2022. Emerging issues questioning the current treatment strategies for lumbar disc herniation. Front. Surg. 9.
Wei, F.-L., Li, T., Gao, Q.-Y., Yang, Y., Gao, H.-R., Qian, J.-X., et al., 2021. Eight surgical interventions for lumbar disc herniation: a network meta-analysis on complications. Front. Surg. 8.