Turkish Journal of Geriatrics
2026; 29(2):191–202

DOI: 10.29400/tjgeri.2026.491

Uğur SÖZLÜ[1]
Selim Mahmut GÜNAY[2]
Sevda Demir TÜRE[3]
Gül ÖZKAZANÇ[4]

[1] Tokat Gaziosmanpasa University, Department of Physical Therapy and Rehabilitation, Tokat, Türkiye

[2] Hatay Mustafa Kemal University, Department of Physical Therapy and Rehabilitation, Hatay, Türkiye

[3] Bursa Uludag University, Department of Physical Medicine and Rehabilitation, Faculty of Medicine, Bursa, Türkiye

[4] Bursa Technical University, Graduate School of Education, Bursa, Türkiye

Correspondence

Uğur SÖZLÜ
Phone : +905368921936
e-mail : ugur.sozlu@gop.edu.tr

Received : Feb 20, 2026
Accepted: Apr 27, 2026

Cite this article as:
Sözlü U, Günay SM, Türe SD et al. Most Frequently Asked Questions by Older Adults in Geriatric Rehabilitation: Evaluating Large Language Models As a Source of Information. Turkish Journal of Geriatrics2026;29(2):191–202.doi:10.29400/tjgeri.2026.491

ORIGINAL ARTICLE

MOST FREQUENTLY ASKED QUESTIONS BY OLDER ADULTS IN GERIATRIC REHABILITATION: EVALUATING LARGE LANGUAGE MODELS AS A SOURCE OF INFORMATION

Abstract

Introduction: Older adults in geriatric rehabilitation are increasingly turning to internet-based resources and large language models for health-related information outside of clinical follow-up. The aim of this study was to compare the reliability, clinical accuracy, quality, usefulness, and readability of ChatGPT-5.2, Gemini 3, and DeepSeek V3.2 responses to patient questions regarding geriatric rehabilitation.

Materials and Method: In this cross-sectional comparative content analysis, 24 predefined questions on geriatric rehabilitation were developed from YouTube comments, relevant literature, and clinical experience. Each question was submitted to three large language models under standardized conditions. Anonymized responses were independently evaluated by two experienced physiotherapists for reliability, clinical accuracy, quality, and usefulness, with disagreements resolved by consensus with a specialist physician. Readability was assessed using the Flesch Reading Ease scores.

Results: No statistically significant difference was found among the three large language models in terms of reliability scores ($p = 0.097$). However, significant differences were observed for clinical accuracy, quality, usefulness, readability, and text characteristics (all $p = 0.001$). ChatGPT-5.2 and DeepSeek V3.2 showed the highest clinical accuracy scores, while ChatGPT-5.2 was superior in terms of quality and usefulness. For readability, ChatGPT-5.2 and Gemini 3 outperformed DeepSeek V3.2.

Conclusion: Although all three large language models generally produced reliable content, ChatGPT-5.2 and DeepSeek V3.2 showed stronger clinical accuracy performance. Nevertheless, because of the risk of incorrect information being generated, the use of large language models by the older population should preferably be done under expert supervision.

Keywords: Geriatrics; Rehabilitation; Artificial Intelligence; Patient Education as Topic; Health Literacy.

INTRODUCTION

Geriatric rehabilitation is a multidisciplinary process aimed at enabling older adults to regain independence, improve their quality of life, and support their social participation (1). The rapid global increase in the older population has further enhanced the strategic importance of this specialized field (2). However, the rehabilitation process is often long-term, and patients require continuous access to health-related information outside of clinical visits. With the increase in digital health literacy, this need is now largely met through internet-based resources rather than traditional methods (3). Accordingly, older adults and caregivers frequently use video-sharing websites and social media platforms to obtain information on symptom management, exercise, and prognosis (4). However, older adults with low health literacy may face risks such as incorrect exercise practices and injury, or they may not adhere to treatment (5). At the same time, serious concerns exist regarding the accuracy, scientific basis, and suitability of the content on these platforms for the specific needs of older adults (4).

In recent years, artificial intelligence systems referred to as large language models (LLMs) (ChatGPT, Gemini, DeepSeek, etc.) have created a significant transformation in the field of health counseling (6-8). Recent studies have indicated that although the responses provided by LLMs to medical questions may at times contain a level of detail comparable to those of specialist physicians (9), some information generated by artificial intelligence may carry a risk of "hallucination" (the production of factually incorrect information) and may not always align with clinical guidelines (10). Additionally, LLMs may generate incorrect or incomplete health information owing to biases in their training data and limitations in source availability (11). This situation has the potential to pose particular risks for the older population.

To the best of our knowledge, the reliability, clinical accuracy, quality, and readability of LLM-generated responses to questions by older adults have not yet been sufficiently examined within the specific context of geriatric rehabilitation. The data obtained through the evaluation of these models will provide critical evidence regarding whether artificial intelligence can serve as a reliable supportive resource in the geriatric rehabilitation process. The primary hypothesis of the study is that the reliability, clinical accuracy, and quality levels of responses provided to patient-centered questions related to geriatric rehabilitation can differ significantly among different LLMs. Accordingly, the aim of this study was to comparatively evaluate the performance of current LLMs in addressing frequently asked patient questions in geriatric rehabilitation in terms of the reliability, clinical accuracy, quality, usefulness, and readability of the information they generate.

MATERIALS AND METHOD

Study Design and Question Development Process

This study was conducted as a cross-sectional and comparative content analysis. A blinded approach was adopted, where the evaluators were unaware of which model had generated each response.

No previously validated standardized questionnaire was available for this specific topic, and formal pilot testing was not conducted. Therefore, the final set consisted of 24 predefined patient-centered questions purpose-developed to reflect common information needs in geriatric rehabilitation. Three primary sources were used in the development of the question set. First, as a result of a search conducted on the YouTube platform using the keyword "geriatric rehabilitation," the 30 most-viewed videos were identified, and approximately 1,500 user comments associated with these videos were systematically examined. Recurrent questions, concerns, and information requests in video titles and user comments were analyzed to identify dominant themes. Second, literature on geriatric rehabilitation was reviewed, with current guidelines



and previous scientific studies taken into account (12, 13). Third, clinical experience was incorporated, in that clinical observations and patient feedback from researchers working in the field of geriatric rehabilitation were integrated into the question development process.

Large Language Models and Response Generation

The question submission and response generation processes were performed by a researcher with expertise in the field of information technology but no professional training in the health sciences. Responses to the developed questions were generated using three different LLMs: ChatGPT-5.2 (OpenAI, San Francisco, CA, USA), Gemini 3 (Google, Mountain View, CA, USA), and DeepSeek V3.2 (DeepSeek Inc., Hangzhou, China). Each question was posed separately to each model and only once. To prevent previous inputs from influencing the responses, the session of the respective model was reset before each question, and the chat history was disabled. During response generation, no additional prompts, explanations, or follow-up questions were used, and a standardized and consistent procedure was applied across all models. This approach ensured methodological consistency and minimized potential bias related to response selection (14). All responses were generated in English to maintain linguistic consistency and enhance international comparability. The questions, model responses, and relevant links are presented in Table 1.

Expert Evaluation Process

The responses generated by the models were evaluated by two independent physiotherapists (US, SMG) with doctoral degrees and experience in the field of geriatric rehabilitation. Any disagreement was mediated by a physician (SDT) specialized in physical medicine and rehabilitation, and decisions were reached by consensus. This multi-stage approach was implemented to enhance the objectivity and reliability of the evaluations. The readability assessments were performed independently by an information technology specialist (GÖ).

Blinding Procedure

A single-blind design was applied during the evaluation process. For this purpose, all responses generated by the LLMs were anonymized prior to evaluation, and the model names and their identifying cues were removed from the text. The responses were randomly coded and presented to the evaluators, who were unaware of which model had produced each response. This approach served to minimize the influence of model-specific expectations, biases, or preferences on the evaluation outcomes.

Assessment Tools

Reliability: Information reliability was assessed using the modified DISCERN (mDISCERN) tool, which was developed to evaluate the reliability and transparency of written health information. It is used to assess the clarity of the information aims, the balanced and unbiased presentation of content, and the presence of misleading elements. In this study, each mDISCERN item was scored dichotomously, with 1 point for "yes" and 0 points for "no," and analyses were performed based on total scores (15). A higher mDISCERN score indicates a higher level of information reliability, with scores of 0–1 reflecting low reliability, 2–3 moderate reliability, and 4–5 high reliability (Table 2).

Clinical Accuracy Assessment: Because mDISCERN evaluates reliability rather than direct clinical accuracy, an additional instrument was developed for this purpose. Moreover, no previously validated questionnaire was available to assess the clinical accuracy of LLM-generated responses in the specific context of geriatric rehabilitation. Therefore, a clinical accuracy questionnaire was developed specifically for this study. The preliminary version was reviewed by an expert panel of eight professionals from the fields of Physical

Turkish Journal of GERIATRICS 2026; 29(2):191–202

Table 1. Most Frequently Asked Geriatric Rehabilitation Questions and Links to LLM-Generated Responses

Questions	ChatGPT 5.2	Gemini 3	DeepSeek V3.2
1.As an older adult, which exercises are most beneficial for me?	https://chatgpt.com/share/694d0b65-87f8-8011-a23f-c1891585c80d	https://gemini.google.com/share/7c314a76b04e	https://chat.deepseek.com/share/m276h1dvxp2up0wcla
2.As an individual aged 65 and over, how many days per week should I exercise?	https://chatgpt.com/share/694d0bfb-3cf8-8011-93fe-285e284f5f07	https://gemini.google.com/share/3f8963701f1e	https://chat.deepseek.com/share/rkambbxyg8muruptuu
3.As an older adult, which strength exercises should I perform?	https://chatgpt.com/share/694d0c2e-e604-8011-9b07-c9976f6eb7a9	https://gemini.google.com/share/6934eb295d80	https://chat.deepseek.com/share/30yvu1xutbmg2u4s6g
4.As an older adult living at home, which exercises can I perform independently?	https://chatgpt.com/share/694d0c85-16ec-8011-a179-9df5280591a2	https://gemini.google.com/share/637b8e4e4864	https://chat.deepseek.com/share/r0elb8fyk00iepkkd8
5.As an older adult, is walking alone sufficient for my physical health?	https://chatgpt.com/share/694d0cae-1fb8-8011-a12a-5d4324cdaf76	https://gemini.google.com/share/1bfbbf77c8ec	https://chat.deepseek.com/share/zs34ucn6zzj44kdevr
6.As an older adult, which balance exercises can I perform?	https://chatgpt.com/share/694d0cdc-65fc-8011-bb98-3d54e829b2be	https://gemini.google.com/share/0898b66bf7e6	https://chat.deepseek.com/share/y917yk4yzwafmfqrhk
7.As an older adult, can exercising prevent muscle loss?	https://chatgpt.com/share/694d0d1c-2704-8011-95ee-bbf4a7b20698	https://gemini.google.com/share/5f40ee6f395d	https://chat.deepseek.com/share/7zgqjlss4wecoynq7u
8.As an older adult, is it normal to feel pain while exercising?	https://chatgpt.com/share/694d0d6c-54a8-8011-932d-742906109b91	https://gemini.google.com/share/a9cb73efe55d	https://chat.deepseek.com/share/nvlv4wku8phs76dojy
9.As a geriatric individual, how long should one exercise session last?	https://chatgpt.com/share/694d0db3-f578-8011-b25c-b2bf55e9c166	https://gemini.google.com/share/773ad9244883	https://chat.deepseek.com/share/o27db5h4kfwe0vx9vd
10.As an older adult, how can I reduce my risk of falling?	https://chatgpt.com/share/694d0e00-9c44-8011-bf1b-c8ae0e668177	https://gemini.google.com/share/9da45650ebc5	https://chat.deepseek.com/share/vex30m9f99s5hfzfvj
11.As an older adult, can exercising reduce my tendency to feel fatigued quickly?	https://chatgpt.com/share/694d0e3a-9fb4-8011-9c7c-49208ee9d250	https://gemini.google.com/share/33f48bea576d	https://chat.deepseek.com/share/849ba4v45li7nh3clw
12.As an older adult, which exercises can I do to improve my posture?	https://chatgpt.com/share/694d0e67-a090-8011-9a2d-566b6da94b5c	https://gemini.google.com/share/12d80184f2f4	https://chat.deepseek.com/share/zcshoxt68cba6dkzkf
13.As an older adult, which flexibility exercises should I perform?	https://chatgpt.com/share/694d2014-1dd8-8011-838d-40e170aee3b5	https://gemini.google.com/share/256869f6a1b0	https://chat.deepseek.com/share/rfwfyebmsvtgy59w60
14.As an older adult, could walking slowly be an indicator of a health problem?	https://chatgpt.com/share/694d0e9a-4fe8-8011-9770-cdbac3acc76a	https://gemini.google.com/share/c7342a04bbce	https://chat.deepseek.com/share/rpuxt0snu9t6u8wg0i
15. As a geriatric individual, can posture-supporting braces or assistive devices be beneficial for me?	https://chatgpt.com/share/694d0ec4-7b50-8011-9e8d-183e8666c365	https://gemini.google.com/share/18c2b4e0aa0f	https://chat.deepseek.com/share/36s69svq6pqfy9rbtu
16.As an older adult, can vision problems increase my risk of falling?	https://chatgpt.com/share/694d0ee6-37c4-8011-aecc-c79e1172aa36	https://gemini.google.com/share/37510f1b1495	https://chat.deepseek.com/share/o10tfsh39zthhrclrc
17.As a geriatric individual, which breathing exercises should I perform and how often?	https://chatgpt.com/share/694d0f15-3b04-8011-ade1-174702473c6b	https://gemini.google.com/share/d18541656ff6	https://chat.deepseek.com/share/mhho882bjik7g5h3xe

Table 1. Continued...

Questions	ChatGPT 5.2	Gemini 3	DeepSeek V3.2
18.As an older adult, is it normal to feel short of breath during exercise?	https://chatgpt.com/share/694d0f36-649c-8011-8a84-807179fbf3c4	https://gemini.google.com/share/e75695493926	https://chat.deepseek.com/share/luk2sth3x31b8347sc
19.I am an older adult with a respiratory condition; can I still exercise?	https://chatgpt.com/share/694d0f7b-ffa0-8011-87f2-27d94ca0e7fc	https://gemini.google.com/share/7b2d29f7524b	https://chat.deepseek.com/share/d6jqzlsekvoy5unilt
20.As an older adult, should I consult a physician before starting an exercise program?	https://chatgpt.com/share/694d0f9b-7e5c-8011-9cb8-75f038ccc15c	https://gemini.google.com/share/03da70b4031c	https://chat.deepseek.com/share/7psmolk2d8eyghfya2
21.As a geriatric individual, which symptoms occurring during exercise require consulting a physician?	https://chatgpt.com/share/694d0fbe-ef98-8011-bd77-0c43b8cd1da1	https://gemini.google.com/share/405dac24e5fb	https://chat.deepseek.com/share/cidgwf8fgxdkspkwd6
22.As an older adult living at home, what modifications can I make at home to reduce the risk of falling besides exercising?	https://chatgpt.com/share/694d1002-c020-8011-b5dd-71d52abea9c9	https://gemini.google.com/share/3c93878d9a93	https://chat.deepseek.com/share/bm6rjgr42n7ahr2u0q
23.As an older adult, I am afraid of falling; what can I do to cope with this fear?	https://chatgpt.com/share/694d102c-040c-8011-a7d7-67b3136c6fd7	https://gemini.google.com/share/4fbffad8fd1e	https://chat.deepseek.com/share/nz3x53177do4w9nk1g
24.As an older adult, which aerobic exercises are most beneficial for my overall health?	https://chatgpt.com/share/694d1066-7640-8011-b083-0520ce176a28	https://gemini.google.com/share/4d4b0aad6a7a	https://chat.deepseek.com/share/se1frv81ftdye5xe4p

Medicine and Rehabilitation, Physiotherapy and Rehabilitation, and geriatric rehabilitation. Each item was evaluated in terms of relevance, clarity, and representativeness using a 4-point scale, and the questionnaire showed strong content validity (overall S-CVI/Ave = 0.992; S-CVI/UA = 0.933). Based on the expert feedback, minor revisions were made to improve the geriatric specificity of the instrument. In particular, the wording was refined to better reflect the older adult's clinical condition, comorbidities, and potential risk factors. The final questionnaire consisted of five items scored on a 3-point scale, with 0 indicating incorrect or absent information, 1 indicating partially correct or incomplete information, and 2 indicating correct and adequate information. Higher total scores indicated greater clinical accuracy (Table 2).

Quality: The overall quality level of the responses was evaluated using the 5-point Likert-type Global Quality Scale (GQS), which indicates the general quality, flow, and user value of health-related information. Overall quality dimensions, such as the organization and clarity of the information, as well as its clinical usefulness and appropriateness for the target audience, are considered in this scale. Higher GQS scores indicate that the presented content is of better overall quality and more understandable and useful for older adults (Table 2) (16).

Usefulness: Information usefulness was assessed using the usefulness score, which indicates the clarity, comprehensibility, and applicability of the generated responses to daily life from a patient-centered perspective, particularly for older adults (14, 17). The usefulness score is rated on a scale from 1 to 7, with higher scores indicating that the information is more useful, practical, and applicable for older adults (Table 2).

Readability: Readability for older adults was assessed using the Flesch Reading Ease (FRE) formula, which provides a quantitative measure of this parameter on the basis of average sentence

Table 2. Scales for Reliability, Clinical Accuracy, Information Quality, Usefulness, and Flesch Reading Ease–Based Readability Levels

Scale/Score	Level/ Description	Interpretation /Details
mDISCERN	Item 1	Are the aims clearly stated and achieved?
	Item 2	Are reliable sources of information used?
	Item 3	Is the information provided balanced and unbiased?
	Item 4	Are additional sources listed for users?
	Item 5	Does the response address areas of uncertainty?
Clinical Accuracy Questionnaire	Item 1	The response contains clinically accurate information.
	Item 2	The response is consistent with current guidelines and academic knowledge.
	Item 3	The response is appropriate to the older adult's clinical condition and risk factors.
	Item 4	The response presents exercise or rehabilitation recommendations in a safe, applicable, and, when necessary, professionally directed manner.
	Item 5	The response does not include harmful or overly definitive recommendations; it clearly states necessary safety warnings and the limits of medical advice.
GQS	1	Low quality, poor flow, most information is missing, not useful for patients at all.
	2	Generally low quality and poor flow; some information is present, but many important topics are missing, providing very limited benefit for patients.
	3	Moderate quality with suboptimal flow; some important information is adequately addressed while others are insufficient, partially useful for patients
	4	Good quality with generally good flow; most relevant information is included, although some topics may be missing, useful for patients.
	5	Excellent quality and flow; highly useful, comprehensive, and very beneficial for patients.
Usefulness Score	1	Not useful at all: Unintelligible language, contradictory or missing important information.
	2	Very little useful: Partly clear language; some important information is missing or incorrect.
	3	Reasonably useful: Clear language; most important information missing, some is mentioned.
	4	Partly useful: Clear language; some important information is missing, most is addressed.
	5	Moderately useful: Clear language; most important information is covered but some incomplete/incorrect.
	6	Very useful: Clear language; all major information mentioned; some minor info missing.
	7	Extremely useful: Clear language and all key information included; additional helpful resources provided.
Flesch Reading Ease	91–100	Very easy - US Grade 5
	81–90	Easy - US Grade 6
	71–80	Fairly easy - US Grade 7
	61–70	Standard - US Grade 8-9
	51–60	Fairly difficult - US Grade 10-12
	31–50	Difficult - US Grade 13-16
	0–30	Very difficult - College graduate

length and the number of syllables per word. FRE scores range from 0 to 100, with lower scores indicating more complex and difficult-to-read texts and higher scores indicating simpler and more easily understandable content (Table 2) (14, 18).

Ethics Approval

This study was based on the analysis of textual responses generated by LLMs and patient-centered questions derived from geriatric rehabilitation content available on YouTube. No direct intervention



involving human participants was involved, and no individual patient data were used within the scope of the study. All materials analyzed were obtained from publicly available sources and did not contain personal data or identifiable information. Accordingly, in line with the decision of the Hatay Mustafa Kemal University Clinical Research Ethics Committee (Decision No. 44), ethical approval was not required for this study.

Post Hoc Power Analysis

A post hoc power analysis was conducted in G*Power (version 3.1.9.4) using a chi-square approximation based on the observed Friedman test statistic. Clinical accuracy was defined as the primary outcome. The analysis indicated high statistical power ($1-\beta = 0.99$) to detect the observed difference among the models in clinical accuracy scores ($\chi^2 = 29.183$; Kendall's W = 0.608; Effect size w = 1.103; n = 24; $\alpha = 0.05$).

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 27 (IBM Corp., Armonk, NY, USA). Data distribution was assessed using histograms, Q–Q plots, and the Kolmogorov–Smirnov and Shapiro–Wilk tests. As the assumptions of normality were not met, continuous variables were reported as medians (interquartile ranges). Comparisons among ChatGPT-5.2, DeepSeek V3.2, and Gemini 3 were performed using the Friedman test. When the overall Friedman test result was significant, pairwise post hoc comparisons were conducted using the Wilcoxon signed-rank test with Bonferroni correction. Interrater reliability was evaluated using intraclass correlation coefficients (ICCs) based on a two-way random-effects model with absolute agreement, calculated for each assessment category and total scores. ICC values were interpreted as poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), and excellent (>0.90) agreement (19). Differences with a p-value of less than 0.05 were considered statistically significant.

RESULTS

Comparative analysis of the three models revealed no statistically significant difference in mDISCERN scores (p = 0.097). However, statistically significant differences were observed in clinical accuracy, GQS, and usefulness scores (all p = 0.001). For clinical accuracy, ChatGPT-5.2 and DeepSeek V3.2 achieved similarly high median scores (10 [9–10] and 10 [9.25–10], respectively), whereas Gemini 3 had a significantly lower median score (9 [7–9]). For GQS and usefulness, ChatGPT-5.2 had the highest median scores, followed by DeepSeek V3.2 and Gemini 3, with pairwise post hoc comparisons indicating significant differences between all three models for both outcomes (Table 3).

With regard to the GQS distribution, most of the ChatGPT-5.2 responses were rated as Excellent (n = 15), whereas the Gemini 3 and DeepSeek V3.2 responses were mainly classified as Good (n = 18 and 14, respectively). With regard to usefulness, the ChatGPT-5.2 responses were predominantly rated as extremely useful (n = 14), whereas the Gemini 3 and DeepSeek V3.2 responses were mostly rated as moderately useful (n = 14 and 17, respectively).

The models differed significantly in terms of FRE scores (p = 0.001). ChatGPT-5.2 had the highest median FRE score (55.8 [47–65.2]), followed by Gemini 3 (53 [50.1–59.9]), while DeepSeek V3.2 had the lowest score (46.4). Post hoc comparisons showed that DeepSeek V3.2 differed significantly from both ChatGPT-5.2 and Gemini 3, whereas no significant difference was found between ChatGPT-5.2 and Gemini 3. Significant differences were also observed in word count, sentence count, average words per sentence, and average syllables per word (all p = 0.001). DeepSeek V3.2 produced the longest texts, whereas ChatGPT-5.2 generated the shortest. (Table 3).

The ICC values indicating interrater agreement are presented in Table 4. For mDISCERN scores, DeepSeek V3.2 (ICC: 0.753) and ChatGPT-5.2 (ICC: 0.742) showed good agreement, whereas

Gemini 3 (ICC: 0.638) demonstrated moderate agreement. For clinical accuracy, ChatGPT-5.2 showed good agreement (ICC: 0.777), while DeepSeek V3.2 (ICC: 0.676) and Gemini 3 (ICC: 0.606) showed moderate agreement. GQS scores indicated moderate agreement across all models, with ChatGPT-5.2 yielding the highest ICC (0.644), followed by DeepSeek V3.2 (0.610) and Gemini 3 (0.580). Similarly, usefulness scores showed moderate agreement for all models, with DeepSeek V3.2 showing the highest ICC (0.693), followed by Gemini 3 (0.602) and ChatGPT-5.2 (0.567).

DISCUSSION

The findings of this study suggest that all three evaluated LLMs are capable of providing generally reliable responses to information needs in geriatric rehabilitation. In addition, ChatGPT-5.2 and DeepSeek V3.2 appeared to offer more clinically accurate information than Gemini 3. ChatGPT-5.2 also demonstrated relatively stronger performance in terms of overall quality and usefulness. Furthermore, both ChatGPT-5.2 and Gemini 3 produced more readable responses than DeepSeek V3.2, which may indicate greater accessibility for older adults.

When the mDISCERN scores were compared, all three models were found to provide generally reliable and credible information. Similarly, Sari et al. (14) reported that ChatGPT-4 and DeepSeek V3 generated reliable information, as reflected by their mDISCERN scores. In their study, although DeepSeek V3 exhibited higher reliability and usefulness scores in certain domains, no consistent

Table 3. Comparison of Reliability, Clinical Accuracy, Information Quality, Usefulness, Readability, and Text Characteristics Across Large Language Models

	ChatGPT 5.2 Median (IQR)	Gemini 3 Median (IQR)	DeepSeek V3.2 Median (IQR)	p
mDISCERN score	5 (4–5)	4 (4–5)	4.5 (4–5)	0.097
Clinical accuracy questionnaire	10 (9-10)[a]	9 (7-9)[b]	10 (9.25-10)[a]	0.001
GQS score	5 (4–5)[a]	4 (3-4)[b]	4 (4–4.75)[c]	0.001
Usefulness score	7 (6-7)[a]	5 (4.5–5)[b]	5 (5-6)[c]	0.001
FRE score	55.8 (47–65.2)[a]	53 (50.1–59.9)[a]	46.4 (43.5–57.6)[b]	0.001
Word count	330 (290–375)[a]	445 (412–478)[b]	663 (605–767)[c]	0.001
Sentence count	50 (39.5-60.8)[a]	42.5 (36.3-46)[b]	70.5 (59.5-80.7)[c]	0.001
Average words per sentence	6.75 (6.1–7.7)[a]	11.1 (10.1–12)[b]	9.2 (8.3–10.5)[c]	0.001
Average syllables per word	1.7 (1.6–1.8)[a,b]	1.7 (1.6–1.7)[a]	1.8 (1.6–1.8)[b]	0.001

a, b, c: Different superscript letters indicate statistically significant pairwise differences between groups based on Bonferroni-corrected post hoc comparisons.

Table 4. ICCs for Inter-rater Reliability of LLM-Generated Responses

Scores	ChatGPT 5.2 (ICC, 95% CI)	Gemini 3 (ICC, 95% CI)	DeepSeek V3.2 (ICC, 95% CI)
mDISCERN	0.742 (0.490–0.879)	0.638 (0.323–0.825)	0.753 (0.508–0.885)
Clinical accuracy	0.777 (0.526-0.900)	0.606 (0.180-0.823)	0.676 (0.299–0.857)
GQS	0.644 (0.333–0.829)	0.580 (0.133–0.772)	0.610 (0.282–0.810)
Usefulness	0.567 (0.221–0.787)	0.602 (0.270–0.806)	0.693 (0.410–0.855)



superiority in reliability across all domains was identified among the models (14). In contrast, Zhou et al. (20) reported that the DISCERN scores of surgical information generated by different models remained at a moderate level, primarily as a result of insufficient source citations. In another study, although both ChatGPT-3.5 and -4 were shown to provide reliable and high-quality information related to anterior cruciate ligament conditions, the readability of the responses was limited because of their excessive complexity (21). Our findings are consistent with these observations, indicating that source transparency remains a significant limitation despite a high perceived level of text reliability. Accordingly, although LLMs demonstrate considerable potential for information delivery in geriatric rehabilitation, further improvements are required, particularly with respect to source citation and content verifiability.

Although the reliability levels were comparable across the models in our study, significant differences were observed in terms of clinical accuracy, response quality, and usefulness. From the perspective of clinical accuracy, ChatGPT-5.2 tended to provide simpler, more cautious, and better calibrated responses, particularly for questions requiring risk communication, clear safety boundaries, and clinically balanced explanations. DeepSeek V3.2, in contrast, appeared to generate more detailed and clinically stronger responses in areas requiring broader content coverage, such as exercise types, balance, strength, posture, flexibility, and fall prevention in the home environment. Although Gemini 3 provided practical and structured recommendations in some topics, it generally remained behind the other two models in terms of clinical balance and emphasis on safety. Taken together, these findings suggest that DeepSeek V3.2 may offer an advantage in content scope and depth, whereas ChatGPT-5.2 appears to demonstrate a stronger profile in terms of clinical balance and safe guidance.

A similar pattern was also observed in the quality and usefulness assessments. The highest GQS and usefulness scores were obtained by ChatGPT-5.2, followed by DeepSeek V3.2 and then Gemini 3. This suggests that ChatGPT-5.2 is able to generate more comprehensive and comprehensible responses to common rehabilitation-related questions. A possible explanation may be the ability of ChatGPT to access a broader knowledge base and its greater capacity to adapt information to user needs (22). In contrast, despite being a powerful model, DeepSeek V3.2 tended to generate longer and more complex responses, a pattern that has also been reported in previous comparative studies on rehabilitation (14). These results suggest that DeepSeek may provide more effective responses to specialized questions, whereas ChatGPT delivers more balanced answers to general patient inquiries. However, although the responses may be informative, excessively long and complex outputs can reduce perceived usefulness among older adults, thereby limiting the comprehension and practical applicability of the information.

Readability is a fundamental component in the evaluation of health information provided to older adults, as clear and simplified educational materials substantially facilitate the comprehension and practical application of information (23). We found that the responses generated by ChatGPT-5.2 and Gemini 3 were written at an easier reading level than those produced by DeepSeek V3.2, as reflected by their higher FRE scores (median ≈55–53 vs. 46 for DeepSeek). From a practical perspective, the responses from ChatGPT-5.2 and Gemini 3 corresponded to an approximately high-school reading level ("fairly difficult"), whereas the DeepSeek V3.2 responses reflected university-level readability ("difficult"). Nevertheless, despite offering relatively better readability, neither ChatGPT-5.2 nor Gemini 3 produced patient-friendly, easy-to-read texts, and all model outputs remained above the recommended eighth-grade reading

level. This finding is consistent with a growing body of evidence indicating that health-related responses generated by artificial intelligence are often overly complex for the average patient (14, 21, 24). For example, Fahy et al. (21) reported that responses generated by ChatGPT-3.5 and ChatGPT-4 regarding anterior cruciate ligament injuries were written at approximately twelfth-grade to college-graduate reading levels. Similarly, Kolac et al. (24) found that LLM-generated responses related to rotator cuff injuries exhibited low readability. This persistent gap between the readability of medical artificial intelligence outputs and the literacy level of patients may limit the effective use of even high-quality information. Such limitations become more pronounced among older individuals in whom cognitive, sensory, and educational barriers are more prevalent.

The length and structural organization of responses also appear to contribute to these readability challenges. Responses generated by DeepSeek V3.2 were not only more difficult to read but also substantially longer. Conversely, ChatGPT-5.2 structured information in a more comprehensible manner, using clearer segmentation, shorter sentences, and illustrative exercise images, which enhanced perceived readability. Sari et al. (14) reported a similar pattern for rehabilitation-focused LLM comparisons in the context of ankylosing spondylitis. Moreover, previous studies have noted that LLMs may tend toward "over-answering" with encyclopedic detail, which can be counterproductive for patient education (25). Therefore, LLMs intended for use in geriatric rehabilitation should be optimized to generate simpler responses tailored to patient literacy levels, with clearly stated and transparent sources.

Although all three models produced responses with acceptable reliability, patient-oriented recommendations in geriatric rehabilitation involve clinically sensitive issues with direct implications for patient safety. These include exercise prescription, fall prevention, pain or dyspnea during activity, and situations in which medical evaluation is required. In this context, responses that appear fluent and convincing but are incomplete, overly assertive, or inaccurate may delay appropriate healthcare seeking and encourage unsafe self-management. Moreover, the use of such responses in clinical decision-making raises important medicolegal concerns, particularly regarding responsibility in cases of misinformation, the boundaries of professional obligations, and potential legal consequences. Therefore, our findings suggest that LLMs should be regarded only as supportive educational tools in geriatric rehabilitation and should not replace healthcare professionals, especially in safety-sensitive situations involving older adults.

Limitations

This study has several limitations. First, because of the rapid evolution of LLM technology, model updates may alter performance characteristics over time. Second, the question set was developed specifically for this study and was not pilot-tested, which may limit its validity and the sensitivity to detect subtle differences between models. Third, each question was submitted only once to each model; therefore, the findings may not fully reflect the variability of LLM-generated responses across repeated interactions. Fourth, although a multi-step expert evaluation process was used, the inclusion of a larger and more diverse evaluator panel might have provided a broader clinical perspective. Fifth, LLMs may generate hallucinated, incomplete, or contextually inappropriate responses, and this potential limitation should be considered when interpreting the findings. Finally, the study was restricted to English-language outputs, which may limit the generalizability of the findings to older adults who do not speak English.

CONCLUSION

All evaluated LLMs generally produced responses with acceptable reliability, while ChatGPT-5.2 and DeepSeek V3.2 appeared to demonstrate relatively stronger performance in terms of clinical accuracy. In terms of quality and perceived usefulness, ChatGPT-5.2 outperformed the other models. With regard to readability, ChatGPT-5.2 and Gemini 3 generated more accessible texts than DeepSeek V3.2; however, the outputs of all models remained above the recommended patient-centered reading level, suggesting that readability was still suboptimal. Therefore, although these tools may serve as supportive resources in geriatric rehabilitation, they should not be regarded as substitutes for healthcare professionals and should preferably be used under professional supervision. Future studies should incorporate repeated response generation for each question, larger and more diverse multidisciplinary evaluator panels, and updated LLM versions and multilingual outputs in real-world geriatric rehabilitation settings.

Acknowledgement: During the preparation of this work, the authors used Gemini (Google) to improve the clarity and language of the manuscript. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the content of the published article.

Clinical trial number: Not applicable.

Author contributions: US and SMG designed the study. US, SMG, and SDT developed the question set. GÖ conducted the response generation process. US and SMG independently evaluated the responses, SDT resolved disagreements, and GÖ performed the readability assessments. All authors contributed to data interpretation, critically reviewed the manuscript, approved the final version, and accept accountability for all aspects of the work.

Funding: No funding was received for this research.

Data availability: The data supporting the findings of this study are available from the corresponding author upon request.

Consent to Participate declaration: Not applicable.

Ethics approval and consent to participate: This study was based on the analysis of textual responses generated by LLMs and patient-centered questions derived from geriatric rehabilitation content available on YouTube. No direct intervention entailing human participants was involved, and no individual patient data were used within the scope of the study. All materials analyzed were obtained from publicly available sources and did not contain personal data or identifiable information. Accordingly, in line with the decision of the Hatay Mustafa Kemal University Clinical Research Ethics Committee (Decision No. 44), ethical approval was not required for this study.

Consent for publication: Not applicable.

Conflict of Interest: All authors declare no conflict of interest.

REFERENCES

1. Lim JY. Core principles and structures of geriatric rehabilitation: a narrative review. Ewha Med J. 2025 Oct;48(4):e57. doi: 10.12771/emj.2025.00829.
2. Lubbe AL, van Rijn M, Groen WG, et al. The quality of geriatric rehabilitation from the patients' perspective: a scoping review. Age Ageing. 2023 Mar 1;52(3):afad032. doi: 10.1093/ageing/afad032.
3. Kraaijkamp JJM, Persoon A, Chavannes NH, et al. eHealth in geriatric rehabilitation: an international consensus study. Eur Geriatr Med. 2025 Jun;16(3):985-997. doi: 10.1007/s41999-025-01170-7.
4. Sozlu U, Gunay SM. Educational quality, reliability, and comprehensiveness of YouTube content on fall prevention exercises. Educ Gerontol. 2026;52(2):101–116. doi:10.1080/03601277.2025.2515557.
5. Park Y, Kim SR, Seo HJ, Cho J. Health Literacy in Fall-Prevention Strategy: A Scoping Review. Asian Nurs Res (Korean Soc Nurs Sci). 2024 Dec;18(5):532-544. doi: 10.1016/j.anr.2024.10.011.

6. 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 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198.

7. Koç M. Artificial intelligence in geriatrics. Turkish J Geriatr. 2023;26(4):352–360. doi: 10.29400/tjgeri.2023.362

8. Kutsal Y. Geriatric Rehabilitation in an Ageing World: Challenges, Opportunities, and the Future of Person-Centered Care. Turkish J Geriatr. 2025;28(4):433–438. doi: 10.29400/tjgeri.2025.460.

9. Cheng HY. ChatGPT's Attitude, Knowledge, and Clinical Application in Geriatrics Practice and Education: Exploratory Observational Study. JMIR Form Res. 2025 Jan 3;9:e63494. doi: 10.2196/63494.

10. Hatem R, Simmons B, Thornton JE. A Call to Address AI "Hallucinations" and How Healthcare Professionals Can Mitigate Their Risks. Cureus. 2023 Sep 5;15(9):e44720. doi: 10.7759/cureus.44720.

11. Bélisle-Pipon JC. Why we need to be careful with LLMs in medicine. Front Med. 2024 Dec 4;11:1495582. doi: 10.3389/fmed.2024.1495582.

12. Del Fiol G, Weber AI, Brunker CP, Weir CR. Clinical questions raised by providers in the care of older adults: a prospective observational study. BMJ Open. 2014 Jul 4;4(7):e005315. doi: 10.1136/bmjopen-2014-005315.

13. Grund S, Gordon AL, van Balen R, et al. European consensus on core principles and future priorities for geriatric rehabilitation: consensus statement. Eur Geriatr Med. 2020 Apr;11(2):233-238. doi: 10.1007/s41999-019-00274-1.

14. Sari F, Çelik Z, Mirza Y. ChatGPT-4 vs. DeepSeek-V3: a comparative study of response quality, reliability, usefulness, and readability for exercise and rehabilitation strategies in patients with ankylosing spondylitis. Clin Rheumatol. 2026 Jan;45(1):187-195. doi: 10.1007/s10067-025-07789-y.

15. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999 Feb;53(2):105-111. doi: 10.1136/jech.53.2.105.

16. Bernard A, Langille M, Hughes S, et. al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007 Sep;102(9):2070-2077. doi: 10.1111/j.1572-0241.2007.01325.x.

17. Uz C, Umay E. "Dr ChatGPT": Is it a reliable and useful source for common rheumatic diseases? Int J Rheum Dis. 2023 Jul;26(7):1343-1349. doi: 10.1111/1756-185X.14749.

18. Flesch R. A new readability yardstick. J Appl Psychol. 1948 Jun;32(3):221-233. doi: 10.1037/h0057532.

19. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. doi: 10.1016/j.jcm.2016.02.012.

20. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. Int J Med Inform. 2025 Jun;198:105871. doi: 10.1016/j.ijmedinf.2025.105871.

21. Fahy S, Oehme S, Milinkovic D, Jung T, Bartek B. Assessment of Quality and Readability of Information Provided by ChatGPT in Relation to Anterior Cruciate Ligament Injury. J Pers Med. 2024 Jan 18;14(1):104. doi: 10.3390/jpm14010104.

22. Xu Y, Liu Q, Pang J, et al. Assessment of Personalized Exercise Prescriptions Issued by ChatGPT 4.0 and Intelligent Health Promotion Systems for Patients with Hypertension Comorbidities Based on the Transtheoretical Model: A Comparative Analysis. J Multidiscip Healthc. 2024 Nov 9;17:5063-5078. doi: 10.2147/JMDH.S477452.

23. Okuhara T, Furukawa E, Okada H, Yokota R, Kiuchi T. Readability of written information for patients across 30 years: A systematic review of systematic reviews. Patient Educ Couns. 2025 Jun;135:108656. doi: 10.1016/j.pec.2025.108656.

24. Kolac UC, Karademir OM, Ayik G, et al. Can popular AI large language models provide reliable answers to frequently asked questions about rotator cuff tears? JSES Int. 2024 Nov 29;9(2):390-397. doi: 10.1016/j.jseint.2024.11.012.

25. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8.