

ChatGPT and Neck Pain: Accuracy as a Learning Tool

Vinay R. Bijoor, BA,¹ Eli Berglas, BA,¹ Aaron Lavi, BA, BBA,¹ Rachel Berglas, BA,²
Isaac Inoyatov, BA,¹ David Musheyev, BA,¹ Michael Mesa, BS,¹ Mitchell Ng, MD³

¹State University of New York Downstate College of Medicine, Brooklyn, NY, USA

²Albert Einstein College of Medicine, Bronx, NY, USA

³Maimonides Medical Center, Department of Orthopaedic Surgery, Brooklyn, NY, USA

Disclosures

Presenting Author: Vinay R. Bijoor	Nothing to disclose
-------------------------------------------	---------------------

Co-author: Eli Berglas	Nothing to disclose
-------------------------------	---------------------

Co-author: Aaron Lavi	Nothing to disclose
------------------------------	---------------------

Co-author: Rachel Berglas	Nothing to disclose
----------------------------------	---------------------

Co-author: Isaac Inoyatov	Nothing to disclose
----------------------------------	---------------------

Co-author: David Musheyev	Nothing to disclose
----------------------------------	---------------------

Co-author: Michael Mesa	Nothing to disclose
--------------------------------	---------------------

Co-author: Mitchell Ng	Nothing to disclose
-------------------------------	---------------------

Introduction



Neck pain is increasingly prevalent; projections indicate 269 million cases by 2050 (33% rise from 2020)



Artificial intelligence (AI), particularly ChatGPT, has gained attention for its potential to provide reliable health information



Validating the use of free online resources would help to provide students and patients with more accessible and user-friendly learning modalities



Previous ChatGPT studies showed promising results for low back pain and degenerative spinal issues




Large audience consisting of millions of users creates large potential to positively impact societal health.




Evidence of evolution and improving quality in the orthopedic space

Goals of Present Study

No previous study has evaluated ChatGPT v3.5 or v4.0 in response to general neck pain inquiries

A large, light blue downward-pointing arrow is centered between the first and second text boxes, indicating a logical flow or consequence.

Assess 1. accuracy, 2. readability, 3. quality, 4. understandability, and 5. actionability relative to established guidelines

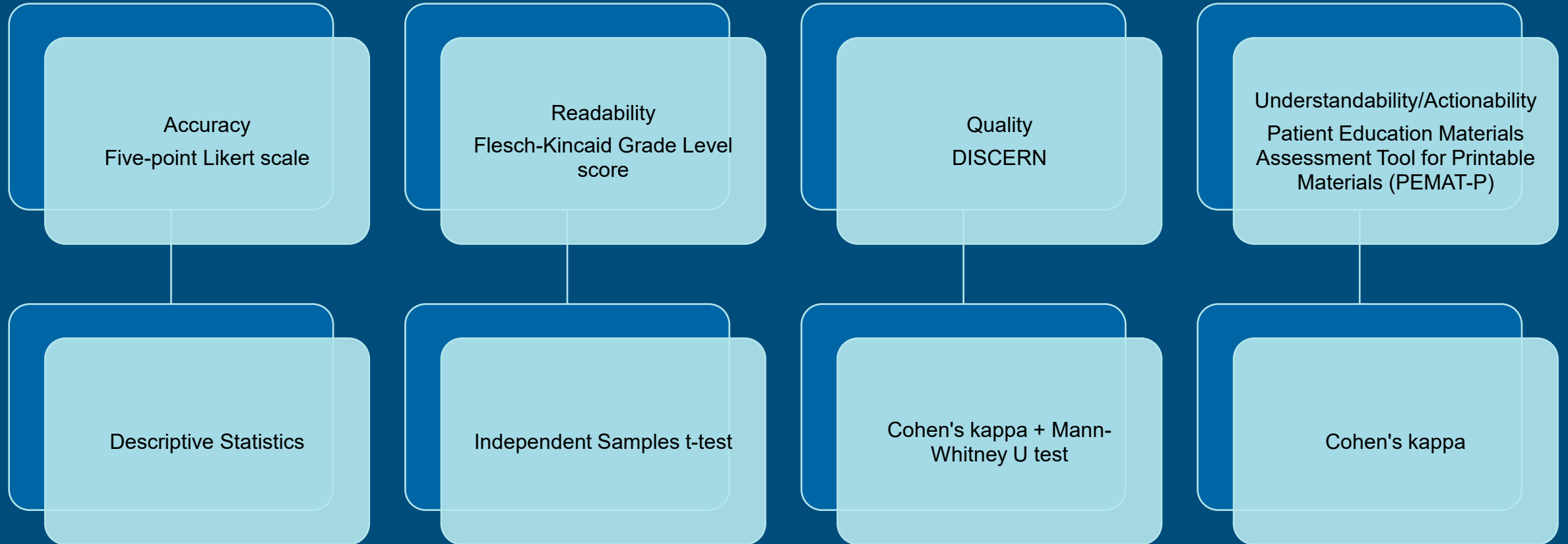
A large, light blue downward-pointing arrow is centered between the second and third text boxes, indicating a logical flow or consequence.

Determine if ChatGPT v4.0 shows quantifiable improvements compared to its predecessor

Methods

- 21 queries were derived from APTA guidelines by rephrasing headings into questions
- Both Chat GPT v3.5 and v4.0 were prompted with these queries
- Responses were assessed for accuracy, readability, quality, understandability, and actionability
- Memory of each chatbot was cleared before each input to ensure unbiased evaluation

Validated Grading Instruments



Results

- Both versions of ChatGPT demonstrated no misinformation
 - Median Likert score = 1
- No significant difference in readability (Flesch-Kincaid Grade Level) between ChatGPT v3.5 and v4.0
 - Both above 12th grade reading level
- ChatGPT v4.0 produced significantly longer responses (318.8 words) compared to v3.5 (229.3 words)

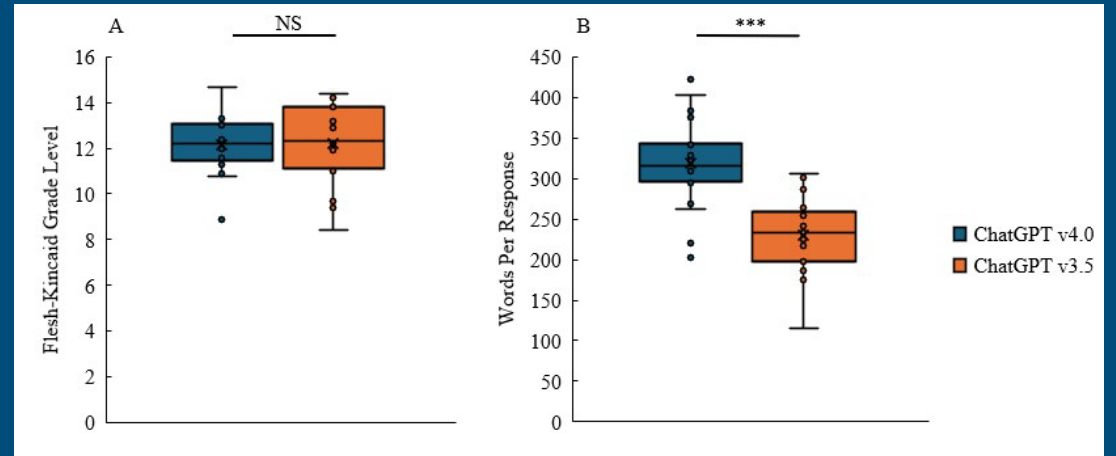


Figure 1. Distribution of chatbot responses to 21 inputs regarding neck pain on **A**. Flesch-Kincaid Grade Level and **B**. Total words per response. *** indicates a significant difference between chatbots with $p < 0.0001$. NS indicates no significant difference ($p > 0.05$)

Results

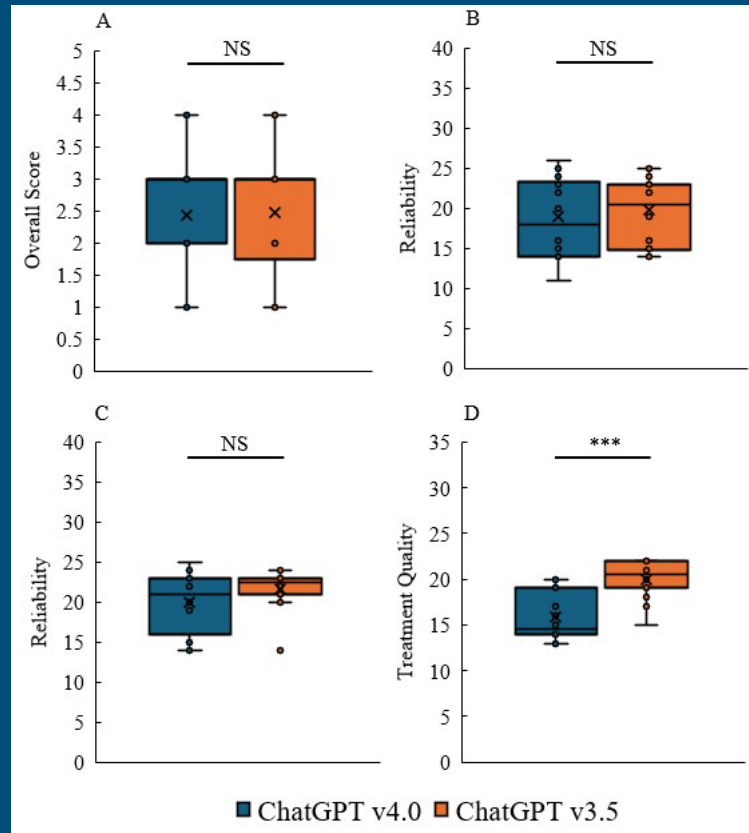


Figure 2. DISCERN scores for chatbots in response to A. Overall score (1-5) B. Reliability of responses to non-intervention related questions C. Reliability of responses to intervention-related questions D. Quality of treatment information provided. *** indicates a significant difference between chatbots with $p < 0.0001$. NS indicates no significant difference ($p > 0.05$)

- Similar overall, intervention-related, and nonintervention-related DISCERN scores
- Exclusively, quality of treatment information was significantly difference

Results

Table 2. PEMAT understandability and actionability scores categorized by query type.

	Non-Intervention Related Query			Intervention Related Query		
	ChatGPT v4.0	ChatGPT v3.5	<i>p</i>	ChatGPT v4.0	ChatGPT v3.5	<i>p</i>
Understandability (%)	78.7	77.0	0.66	78.0	76.4	0.66
Actionability (%)	10.9	10.0	0.83	24.0	21.0	0.61

Conclusions

- Both ChatGPT versions deliver accurate and readable responses of moderate quality regarding neck pain
- ChatGPT v3.5 provides higher-quality treatment information despite being the older version.
- This study highlights the potential of AI tools like ChatGPT in enhancing medical education.



Thank You!