Letters to the Editor

# Evaluation of Online Large Language Models for Patient Education Regarding Chronic Low Back Pain

## To the Editor:

We read with great interest the article by Tong et al "Chronic Low Back Pain and Sleep Disturbance in Adults in the US: The NHANES 2009-2010 Study."published in *Pain Physician* (1). The contribution of this article to the research on chronic lower back pain (CLBP) is commendable. By leveraging data from the 2009-2010 NHANES, it thoroughly investigates the critical role of sleep disorders not only as symptoms but also as exacerbating factors of CLBP. The detailed exploration of the severity and prevalence of CLBP provides valuable insights, paving the way for the development of effective interventions. Moreover, this study inspires further research into health education for CLBP patients.

Chronic lower back pain remains a significant global health challenge, affecting millions worldwide (2). As a primary cause of disability and chronic pain, managing CLBP necessitates continuous self-management and sustained patient support. In the digital era, Large Language Models (LLMs)-advanced AI systems trained on extensive textual data capable of generating human-like text responses- have demonstrated potential as effective tools for enhancing patient education, particularly through facilitating natural language interactions (3,4).

Our comparative assessment study, conducted from April 6 to April 16, 2024, formulated 26 questions derived from guideline-recommended topics and our clinical expertise. These questions covered common concerns related to chronic lower back pain (CLBP), including risk factors, diagnosis, treatment, follow-up, and prognosis (5). We independently posed these questions to four large language models (LLMs): Chat-GPT-3.5, ChatGPT-4, Claude3, and Gemini1.5pro, each providing 26 responses. To ensure the freshness of each session and to prevent any carryover effects from previous interactions, we utilized the new chat feature for all queries. Subsequently, we collated all responses and removed any identifying details, such as the disclaimer "I am not a doctor" from ChatGPT. Two experienced physicians independently evaluated the responses across 5 rounds within a week, ensuring an interval of at least 12 hours between each round to minimize memory bias.

As illustrated in Table 1, our analysis indicates that ChatGPT-4 was the most effective in providing appropriate responses to queries related to CLBP, achieving an appropriateness rate of 96.1%. This rate demonstrates a high level of accuracy and reliability. Gemini1.5pro followed with an appropriateness rate of 92.3%. In contrast, both ChatGPT-3.5 and Claude3 recorded appropriateness rates of 80.7%. A common issue observed was the tendency to conflate general pain management strategies with those specifically tailored for CLBP, resulting in responses that were overly generic. For instance, in response to the query "How is chronic low back pain typically treated? ChatGPT-3.5 incorrectly generalized treatments that were not specific to CLBP, underscoring the need for more precise differentiation in LLM responses.

The superiority of ChatGPT-4 is likely attributable to its extensive parameter set and the vast amount of feedback from both users and experts, which are integral to its training and inference mechanisms (6).While LLMs have demonstrated superior performance in defining and diagnosing conditions, their effectiveness in suggesting treatments and preventive measures for CLBP is less consistent. This discrepancy may stem from the reliance on training datasets that do not include the most current research and guidelines, underscoring the importance of continual updates to the models. Our study benefits from a rigorous design that includes appropriate randomization, washout periods, and a consensus rating process. However, it has limitations; the queries used, although patient-centered and adherent to guidelines, represent just a fraction of the practical applications. Given the rapid evolution of large language models, ongoing evaluations and updates are crucial to ensure they remain aligned with the latest advancements in CLBP management.

Table 1. *Performance of large language models in responding to queries on chronic low back pain.*

| | GPT-3.5 | GPT-4 | Gemini1.5pro | Claude3 |
|---|---|---|---|---|
| **Appropriateness, n (%)** | **24 (80.7)** | **29 (96.1)** | **27 (92.3)** | **24 (80.7)** |
| **1. Risk factors** | | | | |
| Who is more likely to develop chronic low back pain? | √ | √ | √ | √ |
| What lifestyle changes can help manage chronic low back pain effectively? | √ | √ | √ | √ |
| How does posture affect chronic low back pain, and what are the best practices to improve posture? | √ | √ | | √ |
| What type and amount of physical activity is recommended for someone with chronic low back pain? | | | √ | |
| Despite being physically active, how can someone develop chronic low back pain? | √ | √ | √ | √ |
| **2. Diagnosis** | | | | |
| What are the early signs and symptoms of chronic low back pain that I should be aware of? | √ | √ | √ | √ |
| How is chronic low back pain diagnosed? | √ | √ | √ | √ |
| Are there different types of chronic low back pain, and how do they differ? | √ | √ | √ | √ |
| Are there stages of severity in chronic low back pain, and how do they differ? | √ | √ | √ | √ |
| When is imaging such as MRI or CT recommended for chronic low back pain? | √ | √ | √ | √ |
| How should I interpret my functional and pain assessment scores? | | √ | | |
| **3. Treatment** | | | | |
| How is chronic low back pain typically treated? | | √ | √ | |
| Are there any medications specifically prescribed for chronic low back pain? | √ | √ | √ | √ |
| How should I use medication to minimize side effects, especially gastrointestinal or dependency issues? | √ | √ | √ | √ |
| What lifestyle modifications are most effective for managing chronic low back pain? | √ | √ | √ | |
| In severe cases, are there surgical options available for treating chronic low back pain? | √ | √ | √ | √ |
| **4. Continuous Health Assessment** | | | | |
| How often should someone with chronic low back pain have follow-up visits? | √ | √ | √ | √ |
| What tests or evaluations are typically performed during follow-up appointments for chronic low back pain? | | √ | √ | |
| What signs or symptoms of chronic low back pain should prompt immediate medical attention? | √ | √ | √ | √ |
| **5. Comorbidities and Prognosis** | | | | |
| What other health conditions are commonly associated with chronic low back pain? | √ | √ | √ | √ |
| What is the typical prognosis for someone with chronic low back pain? | √ | √ | √ | √ |
| Is there an increased risk of other musculoskeletal problems when living with chronic low back pain? | √ | √ | √ | √ |
| How does chronic low back pain affect mental health, and vice versa? | √ | √ | √ | √ |
| Does having chronic low back pain increase my risk for other chronic conditions such as arthritis or spinal issues? | | √ | √ | √ |
| Can chronic low back pain be reversed or significantly improved? | √ | √ | √ | √ |

Xiang Liu, MD
Department of Rehabilitation, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China

Pengfei Li, MD
Department of General Medicine, the Affiliated Hospital of Qingdao University, Qingdao, Shandong, China

Xiaohua Fan, MD
Department of Rehabilitation, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China

Chao Wang, MD
Department of Rehabilitation, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China
E-mail: wangchao@sdfmu.edu.cn

## REFERENCES

1. Tong Y, Zhang X, Zhou H. Chronic low back pain and sleep disturbance in adults in the US: The NHANES 2009-2010 study. *Pain Physician* 2024; 27:E255-E262.

2. Scholich SL, Hallner D, Wittenberg RH, Hasenbring MI, Rusu AC. The relationship between pain, disability, quality of life and cognitive-behavioural factors in chronic back pain. *Disabil Rehabil* 2012; 34:1993-2000.

3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023; 29:1930-1940.

4. Varghese J, Chapiro J. ChatGPT: The transformative influence of generative AI on science and healthcare. *J Hepatol* 2024; 80:977-980.

5. George SZ, Fritz JM, Silfies SP, et al. interventions for the management of acute and chronic low back pain: Revision 2021. *J Orthop Sports Phys Ther* 2021; 51:CPG1-CPG60.

6. Gallifant J, Fiske A, Levites Strekalova YA et al. Peer review of GPT-4 technical report and systems card. PLOS digital health 2024; 3:e0000417.