Journal of Clinical and Translational Hepatology **2025**
DOI: 10.14218/JCTH.2025.00238

Research Letter

Hepatitis C Patient Education: Large Language Models Show Promise in Disseminating Guidelines

Jinyan Chen[1#], Ruijie Zhao[1#], Chiyu He[1], Huigang Li[1], Yajie You[2], Zuyuan Lin[1], Ze Xiang[1], Jianyong Zhuo[1], Wei Shen[1], Zhihang Hu[1], Shusen Zheng[1], Xiao Xu[2,3*] and Di Lu[2*]

[1]Zhejiang University School of Medicine, Hangzhou, Zhejiang, China; [2]Department of Hepatobiliary & Pancreatic Surgery and Minimally Invasive Surgery, Zhejiang Provincial People's Hospital (Affiliated People's Hospital), School of Clinical Medicine, Hangzhou Medical College, Hangzhou, Zhejiang, China; [3]Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China

Received: May 30, 2025 | **Revised:** September 14, 2025 | **Accepted:** November 20, 2025 | **Published online:** December 18, 2025

Citation of this article: Chen J, Zhao R, He C, Li H, You Y, Lin Z, *et al*. Hepatitis C Patient Education: Large Language Models Show Promise in Disseminating Guidelines. J Clin Transl Hepatol 2025;15(12). doi: 10.14218/JCTH.2025.00238.

This study evaluated the accuracy, completeness, and comprehensibility of responses from mainstream large language models (LLMs) to hepatitis C virus (HCV)-related questions, aiming to assess their performance in addressing patient queries about disease and lifestyle behaviors. The models selected were ChatGPT-4o, Gemini 2.0 Pro, Claude 3.5 Sonnet, and DeepSeek V3, with 12 questions chosen by two HCV experts from the domains of prevention, diagnosis, and treatment. After five hepatologists evaluated their satisfaction with the model responses across the dimensions of accuracy, completeness, and comprehensibility using a 6-point Likert scale (1 = Completely incorrect, 6 = Correct), the results show that the average score of all LLMs is close to 5 points or even higher than 5 points, indicating that there are only a few errors. This suggests that the experts somewhat agreed with the content of the model responses, with Gemini 2.0 Pro potentially outperforming the other models, as its content was agreed upon more extensively by the experts. All models achieved mean scores of 2 for completeness (Adequate, provides the minimal information to be considered complete) and comprehensibility (partly difficult to understand), indicating a certain level of overall usability, though notable deficiencies remain in certain dimensions. Although some pairwise comparisons showed statistically significant differences, the absolute differences in mean scores were small, and overall performance across models was broadly comparable. Although none were rated as "Very Poor," their performance in completeness and comprehensibility was generally similar. This study has revealed the significant potential of artificial intelligence in transforming patient education within conventional clinical practices, as the large models have already approached the level of experts in most responses. However, there are still certain limitations, and further development is required before it can be widely applied in clinical education for HCV.

HCV is a global health concern, with approximately 58 million people infected worldwide in 2019.[1] The World Health Organization has set a target to eliminate hepatitis C by 2030.[2] However, global progress is uneven, with 78.6% of HCV infections remaining undiagnosed and only 21% of diagnosed HCV patients receiving DAA (direct-acting antiviral agents) treatment.[3,4] These limitations are due to constraints in educational and medical resources, necessitating urgent strategic adjustments to achieve the 2030 hepatitis C elimination goal.[1,5,6]

LLMs have profoundly transformed people's lives in recent years and can provide effective personalized support and education for patients in the medical field, potentially helping to supplement healthcare resources.[7,8] Recent research has evaluated the performance of ChatGPT and Gemini in answering viral hepatitis-related questions, revealing both strengths and weaknesses in terms of accuracy, completeness, and comprehensibility.[9,10] In the area of public health, particularly in answering Centers for Disease Control and Prevention and social media-related questions, both models performed similarly and provided high-quality responses. However, when it came to addressing treatment-related questions, both models demonstrated significant shortcomings, particularly in providing accurate and complete information about drug interactions and therapy monitoring. These findings suggest that while LLMs hold promise for HCV education, they still fall short in delivering the practical information and emotional support that patients need. This study assessed the accuracy, completeness, and comprehensibility of responses from mainstream large models to HCV-related questions, aiming to evaluate their performance in addressing patient queries about disease and lifestyle behaviors, and to help select the most suitable general-purpose LLM for this task.

The selection and formulation of the 12 HCV-related questions were conducted through a structured process to ensure clinical relevance and comprehensiveness. Two expert physicians first identified key domains in HCV management—prevention, diagnosis, and treatment—based on current guide-

[#]Contributed equally to this work.

[*]**Correspondence to:** Di Lu and Xiao Xu, Department of Hepatobiliary & Pancreatic Surgery and Minimally Invasive Surgery, Zhejiang Provincial People's Hospital (Affiliated People's Hospital), School of Clinical Medicine, Hangzhou Medical College, Hangzhou, Zhejiang 310014, China. ORCID: https://orcid.org/0000-0002-8724-3739 (DL) and https://orcid.org/0000-0002-2761-2811 (XX). Tel: +86-571-87692654, Fax: +86-571-87692709, E-mail: zjuludi@zju.edu.cn (DL) and zjxu@zju.edu.cn (XX).

Table 1. The question list used in this study

Question list
1. What are the primary routes of HCV transmission, and how can individuals reduce their risk of infection?
2. Are there vaccines available for preventing HCV infection? If not, what preventive measures are recommended?
3. What precautions should healthcare workers take to prevent HCV transmission in clinical settings?
4. What screening recommendations exist for high-risk populations to prevent HCV transmission and progression?
5. What are the recommended diagnostic tests for HCV infection, and in what order should they be performed?
6. How should liver fibrosis be assessed in patients diagnosed with chronic HCV infection?
7. What are the clinical manifestations of acute versus chronic HCV infection? How can they be differentiated?
8. Which patient populations should be prioritized for HCV screening according to EASL guidelines?
9. What are the first-line DAA regimens recommended for treatment-naïve patients with chronic HCV infection?
10. How should treatment response to DAA therapy be monitored during and after completion of therapy?
11. What are the considerations for treating HCV in special populations (e.g., patients with decompensated cirrhosis, post-transplant, or HIV co-infection)?
12. What are the potential drug-drug interactions with DAA therapies that clinicians and patients should be aware of?

HCV, hepatitis C virus; DAA, direct-acting antiviral agents; HIV, human immunodeficiency virus.

lines and common patient inquiries. Within each domain, they collaboratively developed specific questions to address critical and frequently encountered clinical scenarios. For prevention (Q1–4), this included personal prevention methods and screening recommendations for high-risk groups. For diagnosis (Q5–8), questions focused on testing protocols, staging of liver fibrosis, and clinical evaluation. For treatment (Q9–12), prompts covered DAA regimen selection, therapy monitoring, management in special populations, and drug interactions. This methodology ensured that the question list was both systematically derived and representative of core challenges in HCV care. The finalized questions are detailed in Table 1.

Five liver disease experts with over 10 years of experience used the Likert scale to evaluate the accuracy, completeness, and comprehensibility of each model's responses, with a total of 720 ratings for assessment (12 questions × 4 models × 3 dimensions × 5 raters). In descriptive statistics, we described the experimental results using the mean and standard error of scores for each response. Furthermore, we conducted concordance tests separately for each question across the dimensions of accuracy, completeness, and comprehensibility using Kendall's coefficient. Kendall's coefficient of concordance, a non-parametric measure of agreement that considers both the magnitude and direction of differences between raters, was employed. A coefficient of 1 indicates complete agreement, while a coefficient of 0 indicates agreement no different from random judgment.

Accuracy was scored on a Likert scale ranging from 1 to 6 (The specific scoring criteria can be found in Table 2). The average accuracy scores of ChatGPT-4o, Gemini 2.0 Pro, Claude 3.5 Sonnet, and DeepSeek V3 were 4.57 ± 0.11, 5.17 ± 0.10, 4.25 ± 0.10, and 4.17 ± 0.11, respectively. Gemini 2.0 Pro achieved the highest score, and DeepSeek V3 achieved the lowest. Of all the responses, Gemini 2.0 Pro scored the highest on Questions 7 and 10, and Claude 3.5 Sonnet had the worst answer on Question 12. In terms of disease prevention, diagnosis, and treatment, Gemini 2.0 Pro scored the highest among the four models (5.15 ± 0.17, 5.30 ± 0.16, and 5.05 ± 0.18, respectively). ChatGPT-4o achieved the second-highest score in all three domains (5.15 ± 0.16, 5.30 ± 0.20, and 5.05 ± 0.21, respectively). In the fields of prevention and treatment, DeepSeek V3 had the lowest accuracy scores (4.05 ± 0.18 and 4.15 ± 0.18, respectively). In the field of diagnosis, Claude 3.5 Sonnet had the worst performance (4.15 ± 0.18) (Fig. 1A). The average value of Kendall's coefficients of concordance for 12 questions was 0.341, indicating consistency.

In the Likert scale for completeness, which ranged from 1 to 3, the average scores of ChatGPT-4o, Gemini 2.0 Pro, Claude 3.5 Sonnet, and DeepSeek V3 were 2.27 ± 0.08, 2.43 ± 0.07, 2.30 ± 0.08, and 1.92 ± 0.08, respectively. Among all responses, answers to Questions 3, 4, and 12 from Gemini 2.0 Pro and answers to Question 12 from ChatGPT-

Table 2. The Likert scale used in this study

	Accuracy rating Likert scale	Completeness rating Likert scale	Comprehensibility rating Likert scale
1	Completely incorrect	Incomplete. Some aspects of the question are tackled, but significant portions are missing or incomplete	Difficult to understand
2	More incorrect than correct	Adequate. Addresses all aspects of the question and provides the minimal information required to be considered complete	Partly difficult to understand
3	Approximately equally correct and incorrect	Comprehensive. Covers all aspects of the question and delivers additional information or context beyond expectations	Easy to understand
4	More correct than incorrect		
5	Nearly all correct		
6	Correct		

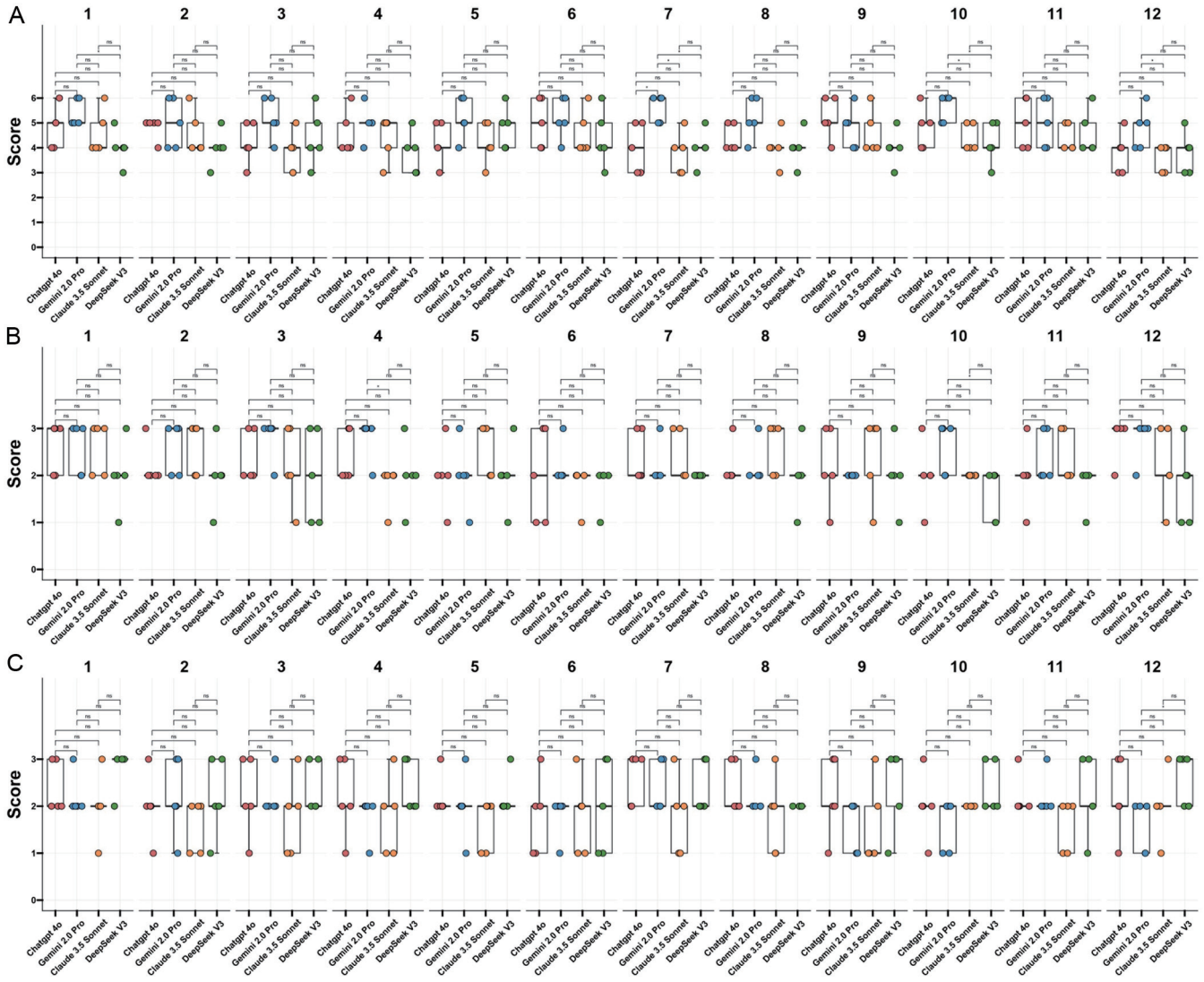

Fig. 1. Box plots illustrating the distribution of accuracy (A), completeness (B), and comprehensibility (C) scores across different large language models. *indicates that there is a significant difference between the two ($p < 0.05$); "ns" indicates that there is no significant difference between the two.

4o achieved the highest scores. In the fields of disease prevention and treatment, Gemini 2.0 Pro received the highest scores (2.70 ± 0.11 and 2.45 ± 0.11). In the field of disease diagnosis, Claude 3.5 Sonnet had the highest average score (2.35 ± 0.13). DeepSeek V3, however, had the lowest average scores in all three fields (Fig. 1B). Kendall's coefficient of concordance was 0.491, indicating a moderate level of agreement. Kendall's coefficients of concordance averaged 0.265, indicating consistency.

Additionally, many evaluating physicians provided suggestions for improving future medical consultation LLMs. They noted that when answering diagnosis-related questions, the models failed to understand patients' anxiety and instead devoted extensive text to explaining relevant principles. Conversely, testing procedures, timelines, accuracy rates, and other information that patients are more interested in were only briefly mentioned. The humanistic care aspect was also noted to be superficial and inadequate.

In terms of comprehensibility, DeepSeek V3 demonstrated a significant advantage with an average score of 2.3 ± 0.08, while Claude 3.5 Sonnet had the lowest average score (1.78 ± 0.09). Among all responses, DeepSeek V3 for Question 1 received the highest average score. Gemini 2.0 Pro received the lowest scores for Questions 9, 10, and 12, and Claude 3.5 Sonnet received the lowest scores for Questions 2, 5, and 9, indicating they were less comprehensible. In the field of disease prevention, DeepSeek V3 had the highest comprehensibility score (2.45 ± 0.14), and Claude 3.5 Sonnet had the lowest (1.80 ± 0.16). In the field of disease diagnosis, ChatGPT-4o was the most comprehensible (2.25 ± 0.14), and Claude 3.5 Sonnet was the least comprehensible (1.75 ± 0.16). In the field of treatment, DeepSeek V3 also achieved the highest score (2.40 ± 0.15), while Gemini 2.0 Pro received the lowest score (1.75 ± 0.12) (Fig. 1C). Kendall's coefficient of concordance was 0.460, indicating a moderate level of agreement. Although Fleiss' Kappa was calculated for consistency analysis among the five raters, it should be noted that this method is primarily designed for nominal categorical data. Given that our data were based on ordinal Likert scales (6-point for accuracy and 3-point for completeness and comprehensibility), Fleiss'

Kappa may underestimate the true agreement because it does not account for the ordinal nature of the ratings. Therefore, we additionally computed Kendall's coefficient of concordance (W), which is more appropriate for ordinal data, and obtained an average value of 0.341 across the 12 items, indicating a moderate level of inter-rater agreement.

All four evaluated LLMs demonstrated potential for supporting HCV patient education, consistently providing medically accurate information (scores > 4). This suggests current LLMs possess sufficient knowledge to address common patient queries about HCV. Among the four models, although DeepSeek had the lowest accuracy score, its score still indicates that it maintains a relatively low error rate. However, due to the persistence of hallucinations, it is still necessary to be cautious in identifying information when using LLMs for medical education. In the future, it may be necessary to fine-tune or conduct specialized training for LLMs in order to improve their accuracy and other aspects. However, important limitations were identified. Despite acceptable completeness scores, models often emphasized theoretical explanations over practical information relevant to patients (testing procedures, timelines, accuracy rates). This highlights a gap in prioritizing patient-centric information and understanding the emotional context of medical inquiries.[11] Comprehensibility varied significantly between models, with DeepSeek V3 excelling in prevention and treatment explanations, while ChatGPT-4o performed best for diagnostic information. Claude 3.5 Sonnet consistently scored lowest, suggesting potential communication issues for patient education. Emerging concerns include potential medical disputes arising from the gap between LLM recommendations and clinical practice. In clinical settings, physicians consider numerous patient-specific factors that standardized LLM responses may not address.[12] Current LLMs also lack the human judgment and empathetic communication that characterize effective physician–patient interactions.[13,14] To mitigate risks, a regulatory framework should govern LLM implementation in patient education, with guidelines for appropriate use, transparency about limitations, and regular quality evaluation. Developers should improve models' ability to understand patient anxiety, prioritize practical information, and communicate with empathy. Despite challenges, LLMs could potentially advance medical education and bridge knowledge gaps in regions with limited access to HCV specialists.[15,16] Future research should focus on specialized medical LLMs for patient education, balancing technical accuracy with emotional intelligence and communication clarity.

Funding

This research was funded by the National Key Research and Development Program of China (No. 2021YFA1100500), the National Natural Science Foundation of China (No. 82370662), and the Key Research & Development Plan of Zhejiang Province (No. 2024C03051).

Conflict of interest

The authors have no conflict of interests related to this publication.

Author contributions

Writing of the manuscript (JC, RZ, HL), graphic plotting (CH, JZ, ZL, YY, ZX), revision of the manuscript (WS, ZH), and literature review (DL, XX, SZ). All authors approved the manuscript.

Ethical statement

The study was deemed exempt from institutional review board approval and informed consent, as it did not include human subjects and therefore did not pose any risks.

Data sharing statement

All data are publicly available online.

References

[1] Fleurence RL, Alter HJ, Collins FS, Ward JW. Global Elimination of Hepatitis C Virus. Annu Rev Med 2025;76(1):29–41. doi:10.1146/annurev-med-050223-111239, PMID:39485830.
[2] WHO. Global health sector strategy on viral hepatitis 2016-2021. Towards ending viral hepatitis. Global health sector strategy on viral hepatitis 2016-2021. Geneva, Switzerland: WHO; 2016.
[3] Torre P, Festa M, Sarcina T, Masarone M, Persico M. Elimination of HCV Infection: Recent Epidemiological Findings, Barriers, and Strategies for the Coming Years. Viruses 2024;16(11):1792. doi:10.3390/v16111792, PMID:39599906.
[4] Hui Z, Yu W, Fuzhen W, Liping S, Guomin Z, Jianhua L, *et al*. New progress in HBV control and the cascade of health care for people living with HBV in China: evidence from the fourth national serological survey, 2020. Lancet Reg Health West Pac 2024;51:101193. doi:10.1016/j.lanwpc.2024.101193, PMID:39315090.
[5] Sallam M, Khalil R. Contemporary Insights into Hepatitis C Virus: A Comprehensive Review. Microorganisms 2024;12(6):1035. doi:10.3390/microorganisms12061035, PMID:38930417.
[6] Gragnani L, Monti M, De Giorgi I, Zignego AL. The Key Importance of Screening Underprivileged People in Order to Achieve Global Hepatitis Virus Elimination Targets. Viruses 2025;17(2):265. doi:10.3390/v17020265, PMID:40007020.
[7] Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023;388(13):1233–1239. doi:10.1056/NEJMsr2214184, PMID:36988602.
[8] Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. Ann Intern Med 2024;177(2):210–220. doi:10.7326/M23-2772, PMID:38285984.
[9] Sahin Ozdemir M, Ozdemir YE. Comparison of the performances between ChatGPT and Gemini in answering questions on viral hepatitis. Sci Rep 2025;15(1):1712. doi:10.1038/s41598-024-83575-1, PMID:39799203.
[10] Li Y, Huang CK, Hu Y, Zhou XD, He C, Zhong JW. Exploring the performance of large language models on hepatitis B infection-related questions: A comparative study. World J Gastroenterol 2025;31(3):101092. doi:10.3748/wjg.v31.i3.101092, PMID:39839898.
[11] Ma X, Zhu R, Wang Z, Xiong J, Chen Q, Tang H, *et al*. Enhancing Patient-Centric Communication: Leveraging LLMs to Simulate Patient Perspectives. arXiv 2025. doi:10.48550/arXiv.2501.06964.
[12] Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, *et al*. Sociodemographic biases in medical decision making by large language models. Nat Med 2025;31(6):1873–1881. doi:10.1038/s41591-025-03626-6, PMID:40195448.
[13] Jin Z, Levine S, Gonzalez Adauto F, Kamal O, Sap M, Sachan M, *et al*. When to make exceptions: Exploring language models as accounts of human moral judgment. Adv Neural Infor Proc Sys 2022;35:28458–28473.
[14] Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, *et al*. Large Language Models and Empathy: Systematic Review. J Med Internet Res 2024;26:e52597. doi:10.2196/52597, PMID:39661968.
[15] Rodriguez JA, Alsentzer E, Bates DW. Leveraging large language models to foster equity in healthcare. J Am Med Inform Assoc 2024;31(9):2147–2150. doi:10.1093/jamia/ocae055, PMID:38511501.
[16] Strika Z, Petkovic K, Likic R, Batenburg R. Bridging healthcare gaps: a scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts. Postgrad Med J 2024;101(1191):4–16. doi:10.1093/postmj/qgae122, PMID:39323384.