Journal of the Canadian Association of Gastroenterology, 2025, **8**, 58–62
https://doi.org/10.1093/jcag/gwae055
Advance access publication 22 February 2025
Original Article

Artificial intelligence in hepatology: a comparative analysis of ChatGPT-4, Bing, and Bard at answering clinical questions

Sama Anvari[1], Yung Lee[2,3], David Shiqiang Jin[4], Sarah Malone[5], Matthew Collins[*,1]

[1]Division of Gastroenterology, Department of Medicine, McMaster University, Hamilton, ON, L8S 4L8, Canada,
[2]Division of General Surgery, McMaster University, Hamilton, ON, L8S 4L8, Canada,
[3]Center for Minimal Access Surgery (CMAS), St. Joseph's Healthcare Hamilton, McMaster University, Hamilton, ON, L8N 4A6, Canada,
[4]Faculty of Health Sciences, McMaster University, Hamilton, ON, L8S 4L8, Canada,
[5]Faculty of Sciences, Queens University, Kingston, ON, K7L 3N6, Canada

[*]Corresponding author: Matthew Collins, MD, FRCPC, Division of Gastroenterology, Department of Medicine, McMaster University, Hamilton, ON, L8S 4L8, Canada, (collim4@mcmaster.ca).

Abstract

Background and Aims: The role of artificial intelligence (AI) in hepatology is rapidly expanding. However, the ability of AI chat models such as ChatGPT to accurately answer clinical questions remains unclear. This study aims to determine the ability of large language models (LLMs) to answer questions in hepatology, as well as compare the accuracy and quality of responses provided by different LLMs.

Methods: Hepatology questions from the Digestive Diseases Self-Education Platform were entered into three LLMs (OpenAI's ChatGPT-4, Microsoft's Bing, and Google's Bard) between September 7 and 13, 2023. Questions were posed with and without multiple-choice answers. Generated responses were assessed based on accuracy and number of correct answers. Statistical analysis was performed to determine the number of correct responses per LLM per category.

Results: A total of 144 questions were used to query the AI models. ChatGPT-4's accuracy was 62.3%, Bing's accuracy was 53.5%, and Bard's accuracy was 38.2% ($P < .001$) for multiple-choice questions. For open-ended questions, ChatGPT-4's accuracy was 44.4%, Bing's was 28.5%, and Bard's was 21.4% ($P < .001$). ChatGPT-4 and Bing attempted to answer 100% of the questions, whereas Bard was unable to answer 11.8% of the questions. All 3 LLMs provided a rationale in addition to an answer, as well as counselling where appropriate.

Conclusions: LLMs demonstrate variable accuracy when answering clinical questions related to hepatology, though show comparable efficacy when presented with questions in an open-ended versus multiple choice (MCQ) format. Further research is required to investigate the optimal use of LLMs in clinical and educational contexts.

Key words: large language models; hepatology; artificial intelligence; clinical; medical education.

Introduction

The role of artificial intelligence (AI) in medicine is a rapidly evolving area of study. Within the domain of hepatology, AI models have been used to enhance the analysis of liver radiology, blood work, and histopathology; diagnose; predict of recurrence of hepatocellular carcinoma; and interpret non-invasive tests.[1]

One area of recent interest has been the feasibility of using AI chat models, also known as large language models (LLMs), to answer clinical questions. The ability of LLMs such as ChatGPT to accurately answer medical and subspecialty board examination questions has been studied previously, with variable results.[2–4] ChatGPT has also demonstrated a promising capacity to answer questions related to cirrhosis, hepatocellular carcinoma,[5] and liver transplantation.[6] However, to our knowledge, the use of AI chat models to answer questions in other domains of hepatology, such as viral hepatitis and other liver diseases, has not been studied. Furthermore, there is a paucity of research comparing the ability of different AI chat models to accurately respond to clinical questions in hepatology. Given the broader use of AI chat models by patients, trainees, and healthcare providers, understanding LLMs' ability to provide accurate responses has both clinical and educational implications. Therefore, the purpose of this study was to assess and compare the ability of different AI chat models to respond to hepatology questions as well as provide insight into their strengths and limitations.

Methods

Question selection and use of AI-based LLMs

Multiple choice questions taken from the Digestive Diseases Self-Education Platform (DDSEP+), created by the American Gastroenterological Association, were used as prompts in this study.[7–9] This resource was chosen as it is published by an established medical society, widely recognized by gastroenterology providers, and its questions are regularly updated to reflect recent guidelines. Questions were taken from the

end-of-chapter review questions for hepatology-focused chapters (5-7): "viral hepatitis," "metabolic, hereditary, inflammatory, and vascular diseases of the liver," and "cirrhosis and liver transplantation." All published questions were included; repeat questions were removed. When lab values were presented in a table format, they were transcribed to sentence form, with reference values provided as seen in the original question.

Prompts were queried into three AI-based chat models: OpenAI's ChatGPT-4 (San Francisco, California), Microsoft's Bing (Redmond, Washington), and Google's Bard (Mountain View, California). Each question was posed as a new chat conversation, and queries were input between September 7, 2023, and September 13, 2023. Questions were input twice: first with the prompt "Answer this question," omitting the multiple-choice answers, and then again with the prompt "Answer this multiple-choice question" and the multiple-choice answers listed. For multiple-choice questions, if the LLM did not provide an answer in line with one of the options listed, it was prompted to choose a single answer for a maximum of 3 attempts. Questions and responses were organized using an Excel spreadsheet. Questions were organized into columns related to the chapter and question number, question text, multiple choice question options, and the answer provided by DDSEP+. Questions were sorted into the following categories: case scenario (ie, a specific patient scenario was presented), epidemiology/risk factors (ie, a question concerned specific risk factors or disease statistics), and guideline/policy (ie, a question asked for management based on specific guidelines or policies without much case detail); each question was also classified by question type (pathophysiology, investigation, diagnosis, management, surveillance, epidemiology, pharmacology). Responses were organized by type of LLM. The inclusion of laboratory values, imaging findings, and pathology findings within the question stem were noted.

Outcomes

Responses generated by LLMs were compared to the responses provided by DDSEP+. Answers were marked as correct, incorrect, or unable to answer. Questions that the LLM was unable to answer were counted as incorrect. If the LLMs provided multiple potential courses of action, including the correct answer, this was classified as a separate category ("multiple answers given") but counted as incorrect in final analyses. Similarly, answers that required re-prompting were classified as incorrect in order to provide a more conservative estimate of LLM capability, though whether or not the next answer given was correct was recorded. The number of correct answers based on subject matter and question type, as well as the accuracy of correct answers, were assessed by medical trainees under the guidance of a hepatologist. Questions were separated into the following subgroups: correct, incorrect, and unable to answer. Subgroup analyses were conducted based on the question type, with questions being classified as epidemiology/risk factors (ie, a question concerning specific risk factors or natural history of a condition), pathophysiology (ie, regarding a specific disease process/characteristic), investigation (ie, question about what diagnostic test is most appropriate in a given scenario), diagnosis (ie, the question asked about a most likely diagnosis/features of a diagnosis), management (ie, the question concerning most appropriate treatment or medication), surveillance (ie, the question asked about surveillance for a particular condition), and pharmacology (ie, the question asked about a medication interaction or side effect).

Statistical analysis

All statistical analyses were performed using STATA (StatCorp version 17, College Station, TX). Data were reported as the proportion of correct answers per LLM and per question category in percentages (%). A 2 × 3 chi-square test was used to compare dichotomous variables. Statistical tests were all 2-sided, and the threshold for significance was set at P <.05. 95% confidence intervals (CIs) were provided where applicable.

Results

A total of 144 questions were used to query the LLMs. These included case scenarios (n = 126), epidemiology/risk factors (n = 3), and guideline/policy-based question (n = 15); question subgroups included epidemiology/risk factors (n = 15), pathophysiology (n = 8), investigation (n = 22), diagnosis (n = 15), management (n = 69), surveillance (n = 6), and pharmacology (n = 9). Seventy-four % of questions included laboratory values, 34% included imaging findings (including fibrosis index), and 6% included pathology results to interpret.

Of the 144 multiple-choice questions used, ChatGPT-4 correctly answered 90 (62.3%), Bard correctly answered 55 (38.3%), and Bing correctly answered 77 (53.5%) (Table 1). When prompts were input into the LLMs in an open-ended fashion, ChatGPT-4 correctly answered 64 (44.4%), Bard correctly answered 31 (21.5%), and Bing correctly answered 41 (28.5%) (Table 2). Both ChatGPT-4 and Bing attempted to answer all provided questions (Table 2); Bard was unable to answer 17 (11.8%) multiple choice questions and 18 (12.5%) open-ended questions; of these, 1 concerned diagnosis, 1 was related to epidemiology, 5 were investigation-based questions, 9 were related to management, and 1 was related to pharmacology. Re-querying Bard did not lead to an answer in this setting, with the LLM answering "I am unable to answer" after 3 attempts. No rationale for this was provided. ChatGPT-4 had the highest number of correct answers across question subjects (57.1% of viral hepatitis, 68.9% of cirrhosis and liver transplant, 62.0% of metabolic, hereditary, inflammatory, and vascular disease of the liver), while Bard had the lowest number of correct responses (Table 3). We also assessed instances where multiple answers were provided in response to a multiple choice question: this occurred for 2 questions answered by ChatGPT, 16 questions answered by Bing, and 0 questions answered by Bard. When the LLMs

Table 1. Accuracy of provided answers for multiple choice questions by different large language models.

AI model	Correct n (%)	Incorrect n (%)
Chat-GPT4	90 (62.5%)	54 (37.5%)
Bard	55 (38.2%)	89 (61.8%)
Bing	77 (53.5 %)	67 (46.5%)
P-value	<.001	<.001

Abbreviation: AI, artificial intelligence.

were re-prompted to choose, ChatGPT correctly answered 1/2 (50%) and Bing correctly answered 9/16 (56%) of questions.

The results of subgroup analysis by question type are summarized in Table 4. ChatGPT answered significantly more questions correctly in the diagnosis category (80% correct) compared to Bing (66.7% correct) and Bard (25% correct) (P = .002). In the surveillance category, Bing correctly answered significantly more questions (100%) compared to ChatGPT (83.3%) and Bard (33.3%). There were no significant differences in accuracy noted between the LLMs when answering questions related to epidemiology/risk factors, investigations, pathophysiology, or pharmacology.

Discussion

The present study sought to evaluate the ability of LLMs to answer clinical questions in hepatology and it is the first to compare the quality and reliability of provided responses. We found a statistically significant difference in the overall accuracy between the three different LLMs, with ChatGPT-4 consistently providing the highest number of correct answers both with and without multiple choice prompts, followed by Bing, then Bard. ChatGPT-4 and Bing attempted to answer all questions provided, while Bard did not attempt to answer a proportion of the provided questions. Interestingly, all 3 models provided a rationale for their responses, as well as counselling, where appropriate. These findings not only highlight an emerging role for the use of AI technology in clinical and educational settings but also demonstrate that not all LLMs are created equal.

Given the increasing ubiquity of AI technology, its integration into medical practice and education has been an area of recent study. LLMs are a form of AI technology that predicts the likelihood of a given word sequence based on the context of a prompt and the previous patterns on which it has been trained. LLMs such as ChatGPT have gained popularity due to their capabilities for answering questions in plain language and performing a wide range of clinical tasks. Our study showed that, while LLMs show a promising ability to draw conclusions from complex clinical problems, generated answers should be interpreted with caution. A recent study showed that ChatGPT was unable to pass the American College of Gastroenterology's self-assessment examination, highlighting the limitations of LLMs as an educational tool.[4] Similarly, Yeo et al[5] found that despite showing extensive knowledge of cirrhosis and HCC, clinical advice provided by ChatGPT lacked specific knowledge, such as decision-making cut-offs, treatment durations, and regional guideline variations. One identified pitfall of LLMs is their ability to generate specific, fluent answers that are factually incorrect, also known as stochastic parroting.[10] Providers should proceed with caution when utilizing these modalities to provide definitive answers to clinical questions. Interestingly, the LLMs used in our study performed better when there were multiple-choice options provided compared to open-ended questions, suggesting that with appropriate training, an LLM could serve as a useful clinical adjunct with close supervision. We also found significant variability in the accuracy and quality of provided answers between LLMs based on the type of question provided. Two out of three models demonstrated high accuracy when given questions concerning surveillance (eg, for hepatocellular carcinoma) and diagnosis, topics that are arguably clearer cut and algorithmic in nature, suggesting LLMs may be best suited for answering questions in situations with well-established protocols. However, further studies are required in order to better understand the limitations of this technology.

Interestingly, while all 3 models alluded to the importance of consulting recent guidelines or provider expertise

Table 2. Accuracy of provided answers for open-ended questions by different large language models.

AI model	Correct *n* (%)	Incorrect *n* (%)	Multiple answers given with correct answer included *n* (%)
Chat-GPT4	64 (44.4)	59 (41.0)	21 (14.6)
Bard	31 (21.5)	104 (72.2)	9 (6.3)
Bing	41(28.5)	76 (52.8)	27 (18.8)
P-value	<.001	<.001	.006

Abbreviation: AI, artificial intelligence.

Table 3. Accuracy of provided answers by question subject.

AI model	Viral hepatitis, *n* (%) *n* = 49	Cirrhosis and liver transplant, *n* (%) *n* = 45	Metabolic, hereditary, inflammatory, and vascular diseases of the liver *n* (%) *n* = 50
Chat-GPT4	28 (57.14)	31 (68.88)	31 (62.00)
Bard	15 (30.61)	23 (51.11)	24 (47.85)
Bing	27 (55.10)	23 (51.11)	25 (50.35)
P-value	.014	.144	.929

Abbreviation: AI, artificial intelligence.

Table 4. Accuracy of provided answers by question type by different large language models.

AI model	Diagnosis *n* correct (%) Total *n* = 15	Epidemiology/risk factors *n* correct (%) Total *n* = 15	Investigations *n* correct (%) Total *n* = 22	Management *n* correct (%) Total *n* = 69	Pathophysiology *n* correct (%) Total *n* = 8	Surveillance *n* correct (%) Total *n* = 6	Pharmacology *n* correct (%) Total *n* = 9
Chat-GPT4	12 (80.0)	11 (73.3)	12 (54.5)	39 (56.5)	5 (62.5)	5 (83.3)	6 (66.7)
Bard	3 (25.0)	5 (33.3)	8 (36.4)	29 (42.0)	3 (37.5)	2 (33.3)	5 (55.6)
Bing	10 (66.7)	8 (53.3)	12 (54.5)	32 (46.4)	7 (87.5)	6 (100)	3 (33.3)
P-value	.002	.090	.379	.217	.118	.027	.354

Abbreviation: AI, artificial intelligence.

while answering questions ("If you are concerned that you or someone you know may have hepatic encephalopathy, it is important to see a doctor for evaluation"; "This is a medical question that requires professional advice"), this tendency was more regularly observed in responses provided by Bing and Bard compared to ChatGPT. Furthermore, all 3 models showed some capacity to provide non-pharmacologic counselling, a finding that has been demonstrated previously.[5] While ChatGPT-4 showed the highest overall accuracy, it is worth noting that Bing and Bard have not been well studied in these contexts and may have other strengths or advantages that were not explored here. It is also important to note that all three LLMs used were created for general use, and as such were not specifically designed for or trained in the context of healthcare; as such, the data used to train them was likely sourced from readily available information. Given that many questions used in this exercise require specialized information only available through paid subscriptions to journals or databases, the LLMs' performance in this study likely does not reflect the potential of LLMs in general. Thus, future studies comparing LLMs' ability to perform specific tasks (such as patient counselling or medical education) are warranted to determine which model—if any—is best suited for use in a particular clinical context, as well as whether priming LLMs with specific medical information makes a difference in their performance. We also found that re-querying LLMs did not appear to consistently improve performance, and further studies may seek to investigate the capacity of LLMs to be trained in order to provide accurate responses.

This study is not without limitations. First, LLMs are a dynamic, adaptable resource whose ability to perform a specific task is constantly evolving. Though we endeavoured to use the most up-to-date versions of each software available at the time of our study, these models will continue to be updated and likely will improve their accuracy over time.[11] Furthermore, ChatGPT-4 and Bing (which uses ChatGPT-4 as its base model) previously utilized databases last updated in September 2021, which may have led to responses based on outdated information, though this limitation has since been removed. Next, our study only used a single question bank as a reference, with answers typically based on North American guidelines, limiting generalizability to other contexts. Previous studies have shown that LLMs are unable to independently decide which guidelines to follow unless explicitly specified,[5] which may have led to responses that were graded as incorrect. Although we used the most up-to-date version of the DDSEP+ question bank, it is also possible that the provided textbook responses were out of date by the time of this study. However, we did review all control questions to ensure their accuracy to ensure that this was not a significant issue. This question bank also did not include any questions including photographs or images, limiting any assessment of the LLMs' ability to interpret these data. Finally, the question stems used were designed to be the level of a gastroenterology provider. As such, the findings of this study may not reflect LLMs' ability to answer questions at the level of a layperson or healthcare provider without specialized knowledge in hepatology.

Conclusion

Large language models demonstrate some ability to answer clinical questions in hepatology, with comparable efficacy when presented with questions in an open-ended versus MCQ format. Of available modalities, ChatGPT-4 provided the highest proportion of accurate responses. However, in their present form, all three studied LLMs showed significant limitations. As such, at this time, LLMs should not be used to guide clinical decision-making by healthcare providers, medical trainees, or patients, nor they should not be used in place of standard clinical or educational modalities. While they provide an exciting potential avenue for growth in the medical field that may help with managing the ever-growing discrepancy between healthcare needs and available resources, further research is required to investigate the optimal use of LLMs in hepatology, as well as potentially refine their use in clinical and educational contexts.

Author Contributions

SA and YL conceptualized the project. SA and MC confirmed the methods. SA, DJ, SM, and MC extracted and reviewed the data. SA and YL performed data analysis. SA wrote the initial manuscript. SA, DJ, SM, YL, and MC reviewed and edited the final manuscript.

Funding

None declared.

Conflicts of interest

The authors do not have any relevant conflicts of interest to disclose. Conflict of interest disclosure forms (ICMJE) have been collected for all co-authors and can be accessed as supplementary material here.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Khalifa A, Obeid JS, Erno J, Rockey DC. The role of artificial intelligence in hepatology research and practice. *Curr Opin Gastroenterol.* 2023;39(3):175–180. https://doi.org/10.1097/MOG.0000000000000926
2. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9(1):e45312. https://doi.org/10.2196/45312
3. 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;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
4. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American college of gastroenterology self-assessment test. *Am J Gastroenterol.* 2023;118(12):2280–2282. https://doi.org/10.14309/ajg.0000000000002320
5. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol.* 2023;29(3):721–732. https://doi.org/10.3350/cmh.2023.0089
6. Endo Y, Sasaki K, Moazzam Z, et al. Quality of ChatGPT responses to questions related to liver transplantation. *J Gastrointest Surg.* 2023;27(8):1716–1719. https://doi.org/10.1007/s11605-023-05714-9

7. Lindenmeyer C, Sundaram V. Viral hepatitis. In: Lindenmeyer C, Sundaram V, eds. *Digestive Diseases Self-Education Platform+*. American Gastroenterological Association Institute; 2023.
8. Anderson K, Berg CL. Metabolic, hereditary, inflammatory and vascular diseases of the liver. In: Anderson K, Berg CL, eds. *Digestive Diseases Self-Education Platform+*. American Gastroenterological Association Institute; 2023.
9. Garcia-Tsao G, Berg CL. Cirrhosis and liver transplantation. In: Garcia-Tsao G, Berg CL, eds. *Digestive Diseases Self-Education Platform+*. American Gastroenterological Association Institute; 2023.
10. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. *Hepatol Commun*. 2023;7(4):e0097. https://doi.org/10.1097/HC9.0000000000000097
11. Isenberg M. ChatGPT isn't stuck in 2021 anymore, can browse web for recent answers. CNET September 2023. Accessed April 22, 2024. https://www.cnet.com/tech/chatgpt-isnt-stuck-in-2021-anymore-can-browse-web-for-recent-answers/