Cross-sectional Study

INTERNATIONAL JOURNAL OF SURGERY

OPEN

The performance of large language model-powered chatbots compared to oncology physicians on colorectal cancer queries

Shan Zhou, MD, PhD[b,*], Xiao Luo, MD, MS[a], Chan Chen, MD, MS[c], Hong Jiang, MD, PhD[d,e], Chun Yang, MD, MS[a], Guanghui Ran, MD, MS[a], Juan Yu, MD, PhD[a,*], Chengliang Yin, MD, PhD[e,*]

Background: Large language model (LLM)-powered chatbots have become increasingly prevalent in healthcare, while their capacity in oncology remains largely unknown. To evaluate the performance of LLM-powered chatbots compared to oncology physicians in addressing colorectal cancer queries.

Methods: This study was conducted between August 13, 2023, and January 5, 2024. A total of 150 questions were designed, and each question was submitted three times to eight chatbots: ChatGPT-3.5, ChatGPT-4, ChatGPT-4 Turbo, Doctor GPT, Llama-2-70B, Mixtral-8x7B, Bard, and Claude 2.1. No feedback was provided to these chatbots. The questions were also answered by nine oncology physicians, including three residents, three fellows, and three attendings. Each answer was scored based on its consistency with guidelines, with a score of 1 for consistent answers and 0 for inconsistent answers. The total score for each question was based on the number of corrected answers, ranging from 0 to 3. The accuracy and scores of the chatbots were compared to those of the physicians.

Results: Claude 2.1 demonstrated the highest accuracy, with an average accuracy of 82.67%, followed by Doctor GPT at 80.45%, ChatGPT-4 Turbo at 78.44%, ChatGPT-4 at 78%, Mixtral-8x7B at 73.33%, Bard at 70%, ChatGPT-3.5 at 64.89%, and Llama-2-70B at 61.78%. Claude 2.1 outperformed residents, fellows, and attendings. Doctor GPT outperformed residents and fellows. Additionally, Mixtral-8x7B outperformed residents. In terms of scores, Claude 2.1 outperformed residents and fellows. Doctor GPT, ChatGPT-4 Turbo, and ChatGPT-4 outperformed residents.

Conclusions: This study shows that LLM-powered chatbots can provide more accurate medical information compared to oncology physicians.

Keywords: ChatGPT, Claude 2.1, colorectal cancer, google bard, HuggingChat, physician

[a]Department of Radiology, The First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, Shenzhen, China, [b]Florida Research and Innovation Center, Cleveland Clinic, Port St. Lucie, FL, USA, [c]Department of Clinical Laboratory, Shenzhen Baoan Hospital, The Second Affiliated Hospital of Shenzhen University, Shenzhen, [d]Statistical Office, Zhuhai People's Hospital, Zhuhai Clinical Medical College of Jinan University, Zhuhai and [e]Faculty of Medicine, Macau University of Science and Technology, Macau, China

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding authors. Address: Florida Research and Innovation Center, Cleveland Clinic, Port St. Lucie, FL 34987, USA. Tel.: +183 287 07733. E-mail: zhoushanjnu@hotmail.com (S. Zhou); Department of Radiology, The First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen Second People's Hospital, Shenzhen 518035, China. Tel.: +86 185 766 858 49. E-mail: yujuan0072@qq.com (J. Yu); Faculty of Medicine, Macau University of Science and Technology, Macau 999078, China. Tel.: +861 821 108 6033. E-mail: chengliangyin@163.com (C. Yin).

International Journal of Surgery (2024) 110:6509–6517

Received 23 January 2024; Accepted 6 June 2024

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww.com/international-journal-of-surgery.

Published online 27 June 2024

http://dx.doi.org/10.1097/JS9.0000000000001850

Introduction

With the rapid advancements in generative artificial intelligence (AI), foundational large language models (LLMs) have been developed by companies using the transformer architecture, such as OpenAI's Generative Pre-trained Transformer (ChatGPT), Google Gemini, Large Language Model Meta AI (LLaMA), and Anthropic Claude[1]. These LLMs have emerged as promising tools in medicine, particularly ChatGPT[2]. The public release of ChatGPT by OpenAI on November 30, 2022[3], has triggered immediate global attention. ChatGPT acquired one million users just 5 days after launch and reached 100 million monthly active users after 2 months. ChatGPT, an LLM with billions of parameters, enables users to converse with it in natural language. Its application has gained rapid interest over the past 2 years, especially in exploring its clinical utility[4–7].

ChatGPT showed near-pass performance in the United States Certified Public Accountant Examination (USMLE) and the German State Examination in Medicine[8,9]. However, it failed to pass the ophthalmology board examination and Royal College of Surgeons (Trauma & Orthopaedics) examination[10,11]. With a growing research focus on using ChatGPT for medical information, ChatGPT has shown promise in assisting healthcare professionals, including clinical decision support, clinical training, and education[12]. ChatGPT demonstrates encouraging results as a support tool for breast tumor board decision-making, selecting

imaging examinations, and generating radiology referrals in the emergency department[13–15]. ChatGPT has the potential to aid in clinical training and education; ChatGPT can simulate standardized patients, and their responses are colloquial, vivid, and accurate[16]. ChatGPT is a promising tool for healthcare providers, and emerging research is exploring the use of ChatGPT in answering queries concerning cancer care[17,18]. ChatGPT-3.5's performance has been assessed in various cancers, and it has shown moderate to high accuracy compared to other tools[19–24]. ChatGPT's performance is still controversial despite its high accuracy in various cancers. Cao *et al.*[25] evaluated the performance of ChatGPT-3.5 on questions regarding liver cancer surveillance and diagnosis and reported that ChatGPT-3.5 posted poor performance and had 48% accuracy. Yeo *et al.*[26] assessed the performance of ChatGPT-3.5 in answering questions related to knowledge, management, and emotional support for cirrhosis and hepatocellular carcinoma (HCC). They demonstrated that ChatGPT-3.5 had 74.0% accuracy in knowledge of HCC and correctly answered 76.9% of the questions but failed to specify decision-making cut-offs and treatment durations[26]. However, Benary *et al.*[27] reported that ChatGPT could not provide treatment options for precision oncology comparable to human experts. ChatGPT could transform travel medicine by offering travelers and healthcare professionals precise, personalized, and up-to-date health information[28]. Besides, ChatGPT holds promise for enhancing veterinary anatomy education through visual representation, interactive learning, accessible reference materials, comparative anatomy, case-based learning, and the reinforcement of key concepts[29].

As LLMs continue evolving at an extremely rapid pace. A new version of ChatGPT, ChatGPT-4, was released on March 14, 2023, and can support an estimated 1.8 trillion parameters[30]. Recent reports demonstrated that ChatGPT-4 performed better than ChatGPT-3.5 in USMLE and neurology board-style examinations[31,32]. ChatGPT-4 passed the China National Medical Licensing Examination in Chinese (CNMLE) with an accuracy of more than 80%, while ChatGPT-3.5 failed to pass CNMLE[33]. ChatGPT-4 had superior accuracy for clinical decision support in radiology and myopia care[13,15,34]. ChatGPT-4 achieved a 60% agreement with the guideline recommendations in 25 answering questions concerning a combination of five benign and malignant hepato-pancreatic-biliary-related conditions. In contrast, the agreement varied between conditions; pancreatitis received the highest total score, while HCC received the lowest total score[35]. ChatGPT-4 Turbo, a new model trained with data dating back to April 2023, has been available through an API preview for all paying users since November 6, 2023[36]. Doctor GPT is a specialized version of the ChatGPT model that has been trained with data dating back to April 2023; it was designed to help users diagnose their illness based on their symptoms[37].

Bard is a conversational AI chatbot released by Google on March 21, 2023[38]. Furthermore, Google released a new version of Bard powered by Gemini Pro on December 6, 2023[39,40]. Llama 2, a new language model released by Meta with up to 70 billion parameters and two trillion pretraining tokens, has been available through Huggingface since July 18, 2023[41,42]. Moreover, Mixtral-8x7b, an LLM released by Mistral, has been available through Huggingface since December 11, 2023[43]. On November 21, 2023, Entropic released Claude 2.1, currently available in the USA and UK, and powered their generative AI chatbot[44]. A comparative analysis of these LLMs addressing colorectal cancer (CRC) queries has not been fully explored.

HIGHLIGHTS

- Claude 2.1 significantly outperformed residents, fellows, and attendings.
- Doctor GPT significantly outperformed residents and fellows.
- ChatGPT-4 Turbo, ChatGPT-4, and Mixtral-8x7B significantly outperformed residents.

In this study, our objective was to evaluate the performance of publicly accessible state-of-the-art LLM-powered chatbots compared to oncology physicians in addressing CRC-related questions. We assessed the accuracy and scores of ChatGPT-3.5, ChatGPT-4, ChatGPT-4 Turbo, Doctor GPT, LLaMa-2-70B, Mixtral-8x7B, Bard (Gemini Pro), and Claude 2.1 according to the National Comprehensive Cancer Network (NCCN) guidelines. We compared their performances to those of oncology residents, fellows, and attendings.

Methods

Study design

Our study was conducted between August 13, 2023, and January 5, 2024. According to the NCCN guidelines for colon and rectal cancer, we generated a set of 150 questions, being categorized into six domains. These domains encompassed the principles of imaging, principles of pathology and molecular review, principles of surgery, treatment of nonmetastatic colon cancer, treatment of nonmetastatic rectal cancer, and management of metastatic CRC (Supplementary eTable 1, Supplemental Digital Content 1, http://links.lww.com/JS9/C923).

To ensure that chatbots perform effectively on these questions without additional prompting, we established a baseline for the responses of LLMs by using close-ended questions such as 'Is it appropriate to…,' 'Should… be performed?,' and 'Can… be given to…?' Each prompt was posed three times to ChatGPT-3.5 (8/13/2023 to 8/15/2023), ChatGPT-4 (8/14/2023 to 8/16/2023), ChatGPT-4 Turbo (11/19/2023 to 11/21/2023), Doctor GPT (12/1/2023 to 12/4/2023), Bard (12/7/2023 to 12/8/2023), HuggingChat (Llama-2-70B: 12/18/2023 to 12/19/2023; Mixtral-8x7B: 1/5/2024), and Claude 2.1 (12/31/2023 to 1/2/2024). No feedback was provided to any of these chatbots during the assessment process. Additionally, we recruited nine oncology physicians (three residents, three fellows, and three attendings) from three hospitals in China. We provided Chinese doctors with questions in both English and Chinese and recorded their answers. The answers provided by eight LLM-powered chatbots and nine physicians were evaluated by the authors. Answers that were consistent or inconsistent with the NCCN guidelines were scored as 1 or 0, respectively. The total score for each question was determined by the number of answers aligned with the guidelines, ranging from 0 to 3. The overall accuracy was determined by calculating the number of correct answers out of 150 questions. Any instances of divergent opinions were meticulously addressed and resolved through discussion. The work has been reported in line with the STROCSS criteria[45] (Supplemental Digital Content 5, http://links.lww.com/JS9/C927).

Statistical analysis

Statistical analyses were conducted using GraphPad Prism 9 (GraphPad Software Inc., San Diego, CA, USA). The accuracy of LLM-powered chatbots and physicians was compared using one-way analysis of variance and Tukey's multiple comparisons post-hoc test. The score of LLM-powered chatbots and physicians was compared using the Kruskal–Wallis rank sum test and Dunn's multiple comparison post-hoc test. *P* value <0.05 was considered statistically significant.

Results

The accuracy of LLM-powered chatbots compared to physicians

Figure 1A illustrates the average accuracy of LLM-powered chatbots and physicians' responses to CRC-related questions. Claude 2.1 achieved the highest accuracy, with 82.67 ± 1.767%, followed by Doctor GPT with 80.45 ± 2.037%, ChatGPT-4 Turbo with 78.44 ± 3.417%, ChatGPT-4 with 78.00 ± 2.406%, attendings with 73.56 ± 3.795%, Mixtral-8x7B with 73.33 ± 0.665%, fellows with 72.89 ± 1.018%, Bard with 70.00 ± 1.763%, residents with 65.78 ± 3.149%, ChatGPT-3.5 with 64.89 ± 1.389%, and Llama-2-70B with 61.78 ± 3.006%. Claude 2.1 displayed superior accuracy than ChatGPT-3.5, Llama-2-70B, Mixtral-8x7B, Bard, residents, fellows, and attendings. Supplementary eTable 2 (Supplemental Digital Content 2, http://links.lww.com/JS9/C924) details the comparative analysis of accuracy of LLM-based chatbots and physicians.

We employed close-ended questions in this study and found that these chatbots were able to use confident responses, such as 'yes' and 'no.' Figure 1B depicts the percentage of confident responses out of the 150 questions. Over 95% of Claude 2.1's responses are confident, which is significantly higher than other chatbots (all $P < 0.0001$). Llama-2-70B showed a significantly higher proportion of confident answers than ChatGPT-3.5 ($P = 0.021$), ChatGPT-4 Turbo ($P = 0.0009$), and Doctor GPT ($P = 0.0005$). However, as illustrated in Figure 1C, the accuracy of each chatbot increases for confident answers, except for Claude 2.1.

The scores of LLM-powered chatbots compared to physicians

To investigate the repeatability of these chatbots, we asked each question three times. Furthermore, we recruited nine oncology physicians, with three for each resident, fellow, and attending to answer these questions. Supplementary eTable 3 (Supplemental Digital Content 3, http://links.lww.com/JS9/C925) presents details of the scores of the 150 questions provided by LLM-powered chatbots and physicians. Table 1 summarizes the distribution of scores across six domains. As shown in Figure 2A, both Claude 2.1 and Doctor GPT achieved a higher percentage of questions scoring 3 out of 3, followed by ChatGPT-4, ChatGPT-4 Turbo, Mixtral-8x7B, Bard, ChatGPT-3.5, attendings, Llama-2-70B, fellows, and residents. Residents had the highest percentage of questions scoring 1 or 2 out of 3, followed by fellows, attendings, Bard, Mixtral-8x7B, ChatGPT-3.5, Mixtral-8x7B, ChatGPT-4 Turbo, Claude 2.1, ChatGPT-4, and Doctor GPT. Llama-2-70B received the highest percentage of questions scoring 0 out of 3, followed by ChatGPT-3.5, Bard, Mixtral-8x7B, ChatGPT-4, and Doctor GPT, ChatGPT-4 Turbo, Claude 2.1, attendings, residents, and fellows.

Table 2 summarizes the medians and interquartile ranges of LLM-powered chatbots and physicians across six domains. As illustrated in Figure 2B, Claude 2.1 surpassed ChatGPT-3.5 ($P = 0.0017$), Llama-2-70B ($P < 0.0001$), residents ($P < 0.0001$), and fellows ($P = 0.0367$); Doctor GPT surpassed ChatGPT-3.5 ($P = 0.0074$), Llama-2-70B ($P = 0.0005$), and residents ($P < 0.0001$); ChatGPT-4 Turbo surpassed Llama-2-70B ($P = 0.0157$) and residents ($P = 0.0023$); ChatGPT-4 surpassed Llama-2-70B ($P = 0.0065$) and residents ($P = 0.0009$). To further explore the performance of these chatbots on different fields of questions related to CRC, the scores were further analyzed based on domains. In the domain of principles of pathology and

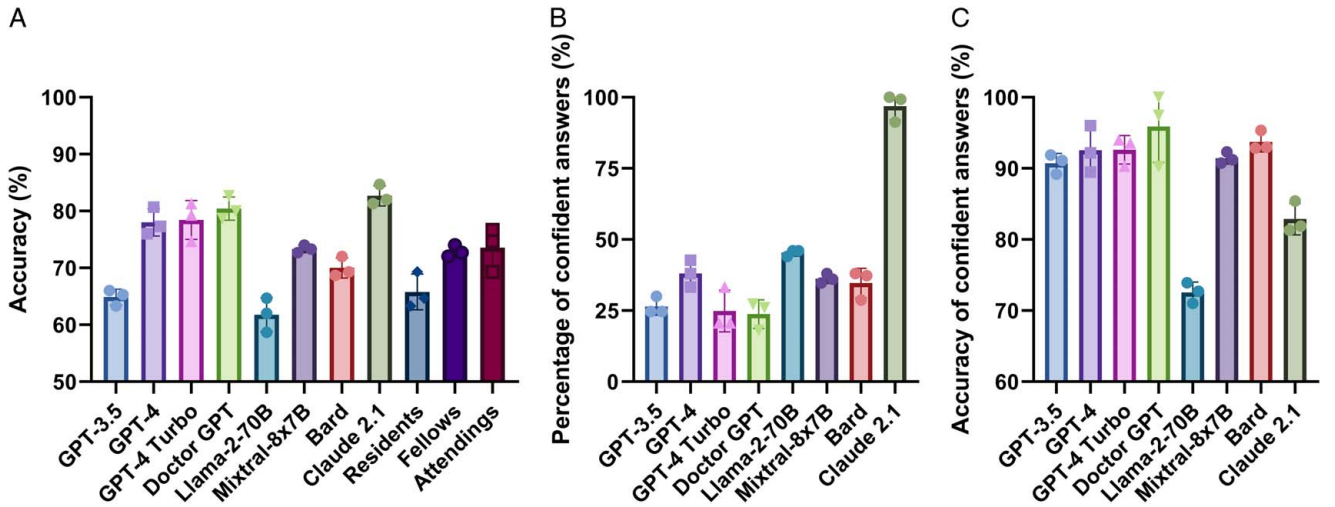

Figure 1. The accuracy of LLM-powered chatbots compared to physicians. (A) The accuracy of LLM-powered chatbots and physicians' responses to questions related to colorectal cancer (CRC). (B) The percentage of confident answers of LLM-powered chatbots. (C) The accuracy of confident answers. LLM, large language model.

Table 1

Comparison of scores between large language model-powered chatbots and physicians across six domains.

	Principles of imaging	Principles of pathology and molecular review	Principles of surgery	Treatment of nonmetastatic colon cancer	Treatment of nonmetastatic rectal cancer	Management of mCRC	Total
No. of questions	19	26	28	13	22	42	150
GPT-3.5, *n* (%)							
Incorrect	2 (10.53)	2 (7.69)	10 (35.71)	2 (15.38)	6 (27.27)	14 (33.33)	36 (24.00)
Partially	8 (42.11)	4 (15.38)	2 (7.14)	1 (7.69)	10 (45.45)	8 (19.05)	33 (22.00)
Correct	9 (47.37)	20 (76.92)	16 (57.14)	10 (76.92)	6 (27.27)	20 (47.62)	81 (54.00)
GPT-4, *n* (%)							
Incorrect	2 (10.53)	2 (7.69)	5 (17.86)	2 (15.38)	2 (9.09)	11 (26.19)	24 (16.00)
Partially	2 (10.53)	4 (15.38)	3 (10.71)	1 (7.69)	1 (4.55)	6 (14.29)	17 (11.33)
Correct	15 (78.95)	20 (76.92)	20 (71.43)	10 (76.92)	19 (86.36)	25 (59.52)	109 (72.67)
GPT-4 Turbo, *n* (%)							
Incorrect	2 (10.53)	1 (3.85)	3 (10.71)	2 (15.38)	2 (9.09)	12 (28.57)	22 (14.67)
Partially	3 (15.79)	5 (19.23)	7 (25.00)	0 (0.00)	2 (9.09)	7 (16.67)	24 (16.00)
Correct	14 (73.68)	20 (76.92)	18 (64.29)	11 (84.62)	18 (81.82)	23 (54.76)	104 (69.33)
Doctor GPT, *n* (%)							
Incorrect	2 (10.53)	1 (3.85)	4 (14.29	1 (7.69)	2 (9.09)	14 (33.33)	24 (16.00)
Partially	4 (21.05)	4 (15.38)	1 (3.57)	2 (15.38)	0 (0)	1 (2.38)	12 (8.00)
Correct	13 (68.42)	21 (80.77)	23 (82.14)	10 (76.92)	20 (90.91)	27 (64.29)	114 (76.00)
Llama-2-70B, *n* (%)							
Incorrect	2 (10.53)	8 (30.77)	10 (35.71)	2 (15.38)	4 (18.18)	13 (30.95)	39 (26.00)
Partially	5 (26.32)	3 (11.54)	3 (10.71)	2 (15.38)	9 (40.91)	12 (28.57)	34 (22.67)
Correct	12 (63.16)	15 (57.69)	15 (53.57)	9 (69.23)	9 (40.91)	17 (40.48)	77 (51.33)
Mixtral-8x7B, *n* (%)							
Incorrect	5 (26.32)	0 (0)	3 (10.71)	1 (7.69)	3 (13.64)	14 (33.33)	26 (17.33)
Partially	4 (21.05)	3 (11.54)	7 (25.00)	5 (38.46)	3 (13.64)	8 (19.05)	30 (20.00)
Correct	10 (52.63)	23 (88.46)	18 (64.29)	7 (53.85)	16 (72.73)	20 (47.62)	94 (62.67)
Bard, *n* (%)							
Incorrect	3 (15.79)	6 (23.08)	6 (21.43)	1 (7.69)	8 (36.36)	4 (9.52)	28 (18.67)
Partially	3 (15.79)	2 (7.69)	10 (35.71)	5 (38.46)	5 (22.73)	12 (28.57)	37(24.67)
Correct	13 (68.42)	18 (69.23)	12 (42.86)	7 (53.85)	9 (40.91)	26 (61.90)	85 (56.67)
Claude 2.1, *n* (%)							
Incorrect	2 (10.53)	5 (19.23)	1 (3.57)	0 (0)	2 (9.09)	7 (16.67)	17 (11.33)
Partially	1(5.26)	1 (3.85)	4 (14.29)	1 (7.69)	0 (0)	11 (26.19)	18 (12.00)
Correct	16 (84.21)	20 (76.92)	23 (82.14)	12 (92.31)	20 (90.91)	24 (57.14)	115 (76.67)
Residents, *n* (%)							
Incorrect	2 (10.53)	2 (7.69)	2 (7.14)	0 (0)	2 (9.09)	5 (11.90)	13 (8.67)
Partially	10 (52.63)	13 (50)	13 (46.43)	6 (46.15)	12 (54.55)	24 (57.14)	78 (52.00)
Correct	7 (36.84)	11 (42.31)	13 (46.43)	7 (53.85)	8 (36.36)	13 (30.95)	59 (39.33)
Fellows, *n* (%)							
Incorrect	2 (10.53)	2 (7.69)	2 (7.14)	1 (7.69)	1 (4.55)	3 (7.14)	11 (7.33)
Partially	8 (42.11)	8 (30.77)	15 (53.57)	3 (23.08)	10 (45.45)	19 (45.24)	63 (42.00)
Correct	9 (47.37)	16 (61.54)	11 (39.29)	9 (69.23)	11 (50)	20 (47.62)	76 (50.67)
Attendings, *n* (%)							
Incorrect	1 (5.26)	5 (19.23)	1 (3.57)	0 (0)	1 (4.55)	7 (16.67)	15 (10.00)
Partially	9 (47.37)	7 (26.92)	9 (32.14)	9 (69.23)	8 (36.36)	14 (33.33)	56 (37.33)
Correct	9 (47.37)	14 (53.85)	18 (64.29)	4 (30.77)	13 (59.09)	21 (50)	79 (52.67)

Incorrect: questions scored 0; partially: questions scored 1 or 2; correct: questions scored 3.

molecular review, Mixtral-8x7B outperformed residents ($P = 0.0289$) (Fig. 2C). In the domain of treatment of nonmetastatic rectal cancer, Claude 2.1 outperformed ChatGPT-3.5 ($P = 0.0013$) and Bard ($P = 0.0295$); Doctor GPT surpassed ChatGPT-3.5 ($P = 0.0013$) and Bard ($P = 0.0295$); both ChatGPT-4 Turbo ($P = 0.0074$) and ChatGPT-4 ($P = 0.0029$) outperformed ChatGPT-3.5 (Fig. 2D). However, there were no significant differences between chatbots and physicians in other domains (Supplementary eFigure 1A-D, Supplemental Digital Content 4, http://links.lww.com/JS9/C926).

Discussion

Transformer-based LLMs are a powerful form of AI that has the potential to change healthcare fundamentally, while their application in oncology is still largely unknown. To comprehensively evaluate the performance of state-of-the-art LLM-powered chatbots in answering CRC-related questions, we assessed the accuracy and scores of eight LLM-powered chatbots, including ChatGPT-3.5, ChatGPT-4, ChatGPT-4 Turbo, Doctor GPT, Llama-2-70B, Mixtral-8x7B, Bard (Gemini Pro), and Claude 2.1. We compared them to those of resident, fellow, and attending

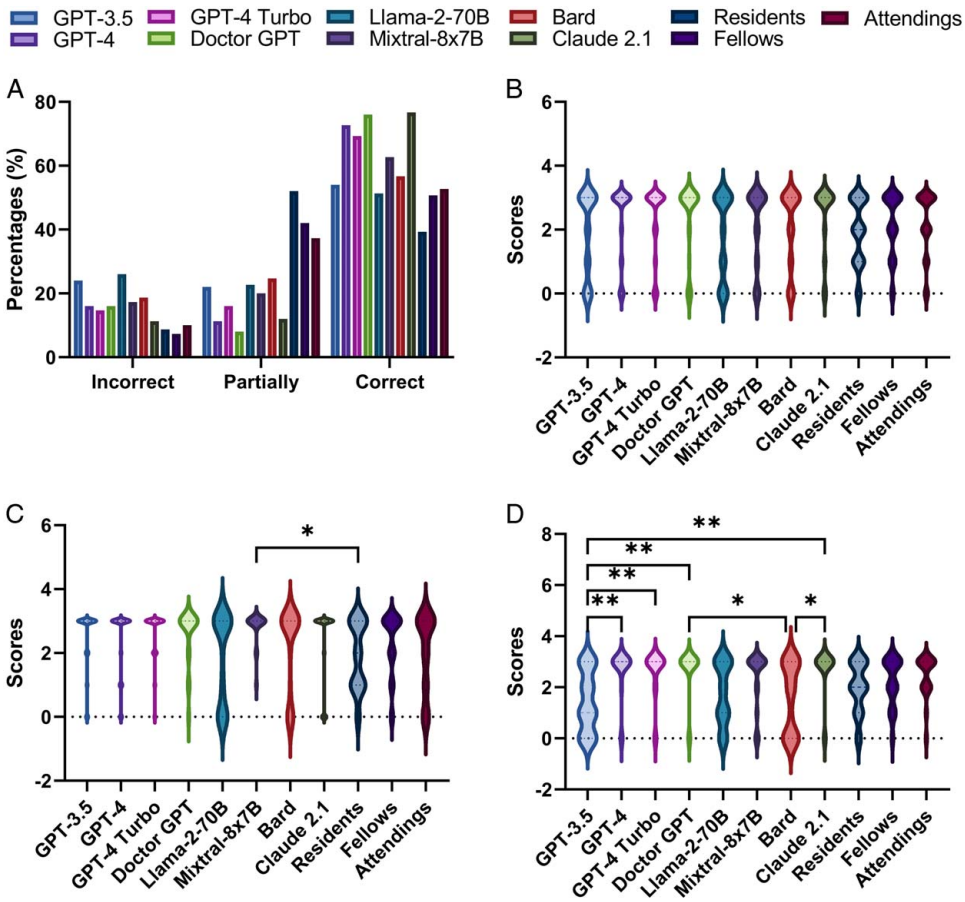

Figure 2. The scores of LLM-powered chatbots compared to physicians. (A) The distribution of scores in LLM-powered chatbots and physicians. Incorrect: questions scored 0; partially: questions scored 1 or 2; correct: questions scored 3. (B) The scores of LLM-powered chatbots compared to physicians across six domains. (C) The scores of LLM-powered chatbots compared to physicians in the domain of principles of pathology and molecular review. (D) The scores of LLM-powered chatbots compared to physicians in the domain of treatment of nonmetastatic rectal cancer. *P < 0.05, **P < 0.01. LLM, large language model.

physicians. Our findings suggest that all chatbots demonstrated commendable performance, and their answers were largely concordant with the NCCN guidelines for colon and rectal cancer. Notably, Claude 2.1 achieved a superior performance overall, whereas the performance of Llama-2-70B and ChatGPT-3.5 appear inferior compared to other chatbots. Regarding accuracy, Claude 2.1, Doctor GPT, ChatGPT-4 Turbo, and ChatGPT-4 showed comparable performance. In addition, Doctor GPT significantly outperformed fellows, and Claude 2.1 significantly outperformed both fellows and attendings. Furthermore, chatbots, except for Llama-2-70B, showed an increased number of questions scored 3 out of 3 compared to physicians, particularly for Claude 2.1 and Doctor GPT. However, physicians achieved a lower percentage of questions scored 0 out of 3. In terms of scores, both Claude 2.1 and Doctor GPT achieved superior performance; Claude 2.1 outperformed fellows, while Doctor GPT did not. Moreover, ChatGPT-4 Turbo and ChatGPT-4 showed comparable performance, outperforming Llama-2-70B and residents.

Consistent with previous work by Liu *et al.*[46], our findings suggest that ChatGPT-4 achieved superior performance, surpassing ChatGPT-3.5 and residents; ChatGPT-3.5 showed comparable performance with residents; ChatGPT-4 showed comparable performance with attendings. Recently, Emile *et al.*[47] studied the appropriateness and consistency of ChatGPT in addressing questions related to the prevention, diagnosis, and management of colon cancer. They reported that 87% of questions provided by ChatGPT-3.5 were deemed appropriate and consistent by at least two independent experts[47]. However, we demonstrated that ChatGPT-4 had 78% of responses consistent with the NCCN guidelines, while ChatGPT-3.5 had 64.89% answering CRC-related questions. Notably, this study assessed accuracy according to the NCCN guidelines, which is different from the previous report and highlights the importance of reference guidelines. The landscape of therapies in CRC is rapidly evolving, followed by a rapidly growing group of approved drugs[48–50]. The clinical practice guidelines are constantly evolving and are updated annually or more frequently. Furthermore, real-world doctors follow different guidelines across global regions. The frequency of LLM updates should adhere to the guidelines to ensure accurate and latest medical information.

Previous reports showed that more than one-third of Americans use the Internet to self-diagnose[51]. LLM-powered chatbots have emerged as a new method for patients to seek medical advice online. Colonoscopy is a commonly used tool for CRC screening. The number of yearly colonoscopies has reached

Table 2

Comparison of median scores between large language model-powered chatbots and physicians across six domains.

	Principles of imaging (median, IQR)	Principles of pathology and molecular review (median, IQR)	Principles of surgery (median, IQR)	Treatment of nonmetastatic colon cancer (median, IQR)	Treatment of nonmetastatic rectal cancer (median, IQR)	Management of metastatic CRC (median, IQR)	Total (median, IQR)
GPT-3.5	2, 1–3	3, 2.75–3	3, 0–3	3, 2.5–3	1, 0–3	2, 0–3	3, 1–3
GPT-4	3, 3–3	3, 2.75–3	3, 1.25–3	3, 2.5–3	3, 3–3	3, 0–3	3, 2–3
GPT-4 Turbo	3, 2–3	3, 2.75–3	3, 2–3	3, 3–3	3, 3–3	3, 0–3	3, 2–3
Doctor GPT	3, 2–3	3, 3–3	3, 3–3	3, 2.5–3	3, 3–3	3, 0–3	3, 3–3
Llama-2-70B	3, 2–3	3, 0–3	3, 0–3	3, 1–3	2, 1–3	1, 0–3	3, 0–3
Mixtral-8x7B	3, 0–3	3, 3–3	3, 2–3	3, 2–3	3, 2–3	2, 0–3	3, 1–3
Bard	3, 2–3	3, 0.75–3	2, 1–3	3, 1.5–3	2, 0–3	3, 2–3	3, 1–3
Claude 2.1	3, 3–3	3, 2.75–3	3, 3–3	3, 3–3	3, 3–3	3, 1–3	3, 3–3
Residents	2, 1–3	2, 1–3	2, 1–3	3, 2–3	2, 1–3	2, 1–3	2, 1–3
Fellows	2, 1–3	3, 2–3	2, 1–3	3, 2–3	2.5, 1.75–3	2, 1.75–3	3, 1.75–3
Attendings	2, 1–3	3, 1–3	3, 2–3	2, 2–3	3, 2–3	2.5, 1–3	3, 2–3

IQR, interquartile range.

70 million, with a 20% increase[52]. Lee *et al.*[53] reported that ChatGPT-3.5 received only 48% accuracy in answering eight common patient questions about colonoscopy. Their four gastroenterologist raters had an accuracy of 33.33% in senior 1 and fellow 1, 44.44% in senior 2, and 80.56% in fellow 2[53]. Kerbage and colleagues demonstrated that ChatGPT-4 had 33% of answers 100% accurate, 53% partially inaccurate, 13% accurate with missing information, and no answers being 100% inaccurate in answering 15 questions on colonoscopy and CRC screening; ChatGPT-4 achieved 40% of answers with 100% accurate, 33% partially inaccurate, 7% accurate with missing information, and 20% completely inaccurate for 15 physician-oriented questions on CRC screening and surveillance[54]. Consistent with previous reports, our results suggest that ChatGPT-3.5 and ChatGPT-4 had a moderate to high accuracy according to the NCCN guidelines in answering 150 CRC-related questions across six domains. Our results showed 54% of questions scoring 3 out of 3, 22% of questions scoring 1 or 2 out of 3, 24% of questions scoring 0 out of 3; while ChatGPT-4 had 72.67% of questions scoring 3 out of 3, 11.33% of questions scoring 1 or 2 out of 3, and 16.00% of questions scoring 0 out of 3 (Table 1).

The performance of Bard is controversial, with an increasing number of reports demonstrating that both ChatGPT-4 and ChatGPT-3.5 outperformed Bard, such as in simplifying radiology reports and answering queries related to myopia, anesthesia, neurosurgery oral boards preparation question bank[34,55–57]. In contrast, Bard achieved a superior accuracy in response to questions related to lung cancer with 95%, compared to ChatGPT-4 with 79% and ChatGPT-3.5 with 84%[20]. Bard performed better START triage by correctly identifying and assigning 60% of patients to the right level of care, compared to ChatGPT-3.5, with an accuracy of 26.67%[58]. Koga and colleagues demonstrated that ChatGPT-3.5 and Bard had comparable diagnostic accuracy in neurodegenerative disorders; ChatGPT-4, ChatGPT-3.5, and Bard had predicted 52, 32, and 40% of cases, correct diagnoses were included in 84% of cases for ChatGPT-4, and 76% for both ChatGPT-3.5 and Bard[59]. Contrary to previous studies, we assessed the latest version of Bard powered by Gemini Pro. Our findings suggest that ChatGPT-4 achieved a higher accuracy, surpassing Bard in addressing queries concerning CRC, while ChatGPT-3.5 showed comparable accuracy with Bard. In addition, we found that Bard outperformed Llama-2-70B.

Anthropic, a public benefit corporation, released two versions of Claude (Claude and Claude Instant) on March 14, 2023[60]. Song *et al.*[61] demonstrated Claude scored the highest accuracy and overall performance in answering urolithiasis-related questions, followed by ChatGPT-4, Bing, and Bard; both Claude and ChatGPT-4 possess a high capacity for analyzing clinical cases of urolithiasis. Wilhelm *et al.*[62] demonstrated that Claude-instant-v1.0 significantly outperformed ChatGPT-3.5-Turbo in simulating treatment recommendation requests on 60 arbitrarily chosen diseases. However, several studies reported that ChatGPT-4 and ChatGPT-3.5 performed better than Claude. Recent studies demonstrated that ChatGPT-4 answered 73.3% of nephrology self-assessment program multiple-choice questions correctly, surpassing Claude 2, who answered 54.4% correctly, which was released on July 11, 2023[63,64]. Moreover, ChatGPT-4 had 96.7% accuracy on questions related to steatotic liver disease, compared to Bard and Llama 2, who had 90% accuracy, and ChatGPT-3.5 and Claude 2, who had 80% accuracy[65]. This study assessed the latest model, Claude 2.1, released on November 21, 2023[44]. Our findings suggest that Claude 2.1, Doctor GPT, ChatGPT-4 Turbo, and ChatGPT-4 have comparable accuracy in answering CRC-related queries, with an average accuracy of 82.67, 80.45, 78.44, and 78%, respectively. Furthermore, our findings showed that Claude 2.1 achieved superior accuracy, outperforming the remaining four chatbots, residents, fellows, and attendings; Doctor GPT outperformed ChatGPT-3.5, Llama-2-70B, Mixtral-8x7B, Bard, residents, and fellows; both ChatGPT-4 and ChatGPT-4 Turbo outperformed ChatGPT-3.5, Llama-2-70B, and residents; Mixtral-8x7B outperformed ChatGPT-3.5, Llama-2-70B, and residents.

The performance of LLM-based chatbots varies across different types of cancer. Chen *et al.*[66] conducted a study to evaluate the performance of ChatGPT-3.5-Turbo in providing treatment recommendations for breast, prostate, and lung cancer that align with NCCN guidelines. They found the percentage of recommended treatments that were in accordance with NCCN guidelines ranged from 41.7 to 85%[66]. Furthermore, the performance of LLM-based chatbots varies across different studies for a

specific cancer, such as liver and prostate cancer. For live cancer, Cao *et al.*[25] reported that ChatGPT-3.5 accuracy was 48% when asked three times for 20 questions regarding liver cancer surveillance and diagnosis. While Yeo *et al.*[26] found that ChatGPT-3.5 was able to provide accurate and comprehensive responses to 74% of the 73 frequently asked questions (FAQs) regarding HCC knowledge and management, which were collected from posts in patient support groups on Facebook. Notably, both studies used the American Association for the Study of Liver Diseases guidelines to grade the responses. Cao and colleagues created questions that address fundamental concepts in liver cancer surveillance and diagnosis, while Yeo and colleagues identified questions from FAQs of professional societies and institutions and FAQs asked in social media. Since LLMs' performance primarily relies on the quality and representativeness of the training data, it is possible that FAQs from public platforms were included in the training data of LLMs and contributed to better performance. This phenomenon also occurs when assessing the performance in addressing questions related to prostate cancer. To evaluate the performance of ChatGPT-3.5, Lombardo *et al.*[67] prepared 195 questions according to the recommendations gathered in the prostate cancer section of the European Urology Association's (EAU) 2023 Guideline-3.5. They found that only 26% of the answers were completely correct. In contrast, Caglar *et al.*[68] prepared 86 questions from FAQs on the websites of urology associations, hospitals, and social media about prostate cancer, as well as the questions prepared according to the strong recommendations of the EAU guideline. They found that the answers to all prostate cancer-related questions were 94.2% completely correct, while the answers to questions prepared according to the strong recommendations of the EAU guideline were 76.2% completely correct[68].

Interestingly, unlike humans, chatbots cannot confidently answer all close-ended questions with 'yes' or 'no.' We demonstrated that Claude 2.1 generated more than 95% confident answers, while other chatbots generated less than 50%. Furthermore, ChatGPT-3.5, ChatGPT-4 Turbo, Doctor GPT, Mixtral-8x7B, and Bard all achieved more than 90% accuracy in their confident answers, outperforming both Llama-2-70B and Claude 2.1; nonetheless, Claude 2.1 outperformed Llama-2-70B. Consistent with previous reports, despite being incorrect, ChatGPT-3.5 and ChatGPT-4 displayed confident language when answering questions about neurology board-like written examinations[34]. These findings indicate that these LLMs are able to confidently provide inaccurate responses, which may be due to the hallucinations. In the field of AI, 'hallucination' is a term used to describe situations in which a model produces fabricated or divergent content that does not align with reality[69,70]. Furthermore, all of the LLM-based chatbots included in this study displayed a confident tone when answering questions, even if their responses were incorrect. For example, 95% of answers provided by Claude 2.1 used confident language, while the accuracy did not improve compared to the overall accuracy of all answers. This study highlights the potential risk of LLM-powered chatbots in the clinical provision of misinformation. The degree of hallucination can have significant implications for patients' well-being in clinical settings[71]. Both patients and doctors seeking medical information should be aware of the limitations of chatbots, which should not be relied upon and used without human expert reviews. Hallucinations are the main obstacle hindering the advancement of AI. However, there is currently no single perfect solution to eliminate LLM hallucinations.

Since its initial launch, ChatGPT has sparked a growing interest in the research field, including medical research. In this study, we evaluated the effectiveness of LLM-powered chatbots in responding to inquiries about CRC across six different domains. Our findings indicate that these chatbots were able to provide accurate medical information about CRC, including the principles of imaging, principles of pathology and molecular review, principles of surgery, treatment of nonmetastatic colon cancer, treatment of nonmetastatic rectal cancer, and management of metastatic CRC. These indicate that LLM-powered chatbots may play a role in tumor boards. The decision-making process at tumor boards is a complex and challenging task in the healthcare field. Recently, there has been a growing number of studies investigating the performance of LLM in tumor boards. Sorin *et al.*[14] conducted a study to evaluate ChatGPT-3.5 as a support tool for decision-making in breast tumor boards. They found that in seven (70%) out of 10 cases, ChatGPT's recommendations were similar to those made by the tumor board[14]. However, Benary and colleagues conducted a study to assess the performance and define the role of four LLMs (ChatGPT, Galactica, Perplexity, and BioMedLM) as support tools for precision oncology. They found that treatment options for LLMs in precision oncology did not reach the quality and credibility of human experts. In the near future, LLMs may already be utilized in healthcare with the rapid evolution of AI. AI is constantly reshaping the world, and the emergence of AI avatars marks the next phase of AI advancement. Al avatars are digital characters powered by cutting-edge algorithms and generative AI technology. Their interaction with users is incredibly realistic, incorporating real-time facial expressions, advanced natural language processing, and text-to-speech capabilities. AI avatars based on LLMs have the potential to assist doctors in diagnosing patients and discussing medical issues with users. AI avatars may be appealing to both doctors and patients.

Limitations

This study has several limitations. First, we recruited a relatively small sample of nine physicians from three hospitals, which may limit the generalizability of our findings. In future studies, it would be beneficial to include a larger and more diverse cohort of physicians from multiple centers. Second, our prompts were designed as close-ended questions that enable LLM-powered chatbots and physicians to respond with either 'yes' or 'no.' This approach may not fully capture the complexities of real-world clinical conditions. Third, our scoring system may not be the most effective way to assess accuracy. Finally, all tested chatbots, except for Doctor GPT, were created for a general audience and not specifically trained for medical professionals, which could result in higher hallucinations.

Conclusion

This study revealed that all LLM-powered chatbots exhibited a notably higher proportion of accurate answers in addressing CRC-related questions according to the NCCN guidelines. In particular, Claude 2.1 achieved superior performance, outperforming attendings, followed by Doctor GPT outperforming fellows. Moreover, the performance of Mixtral-8x7B, ChatGPT-4, and

ChatGPT-4 Turbo surpassed that of the residents. However, there are still some instances of hallucinations in these LLM-powered chatbots. These findings highlight the potential of AI chatbots to deliver reliable medical information. While the results are promising, further improvements and investigations are required for their implementation in clinical practice.

Ethical approval

This study does not include any individual-level data and thus does not require any ethical approval.

Consent

This study does not include any individual-level data and thus does not require it.

Source of funding

Yu is supported by grants from the Shenzhen Science and Technology Innovation Commission (grant number JCYJ20220530150416036) and Shenzhen Second People's Hospital (grant number 2023yjlcyj019). The funders had no involvement in the study design, data collection, analysis, interpretation, writing of the report, or the decision to submit the article for publication. Open access funding was provided by Shenzhen Second People's Hospital. No funding was received for this study.

Author contribution

S.Z.: conceptualization, methodology, data collection, formal analysis, writing – original draft, and writing – review and editing, supervision; X.L., C.C., H.J., C.Y., G.R.: data collection; J.Y.: data collection, writing – review and editing, supervision; C.Y.: conceptualization, methodology, formal analysis, writing – review and editing, supervision.

Conflicts of interest disclosure

The authors declare no conflicts of interest.

Research registration unique identifying number (UIN)

This study does not include any individual-level data and thus does not require it.

Guarantor

This study does not include any individual-level data and thus does not require it.

Data availability statement

The data that support the findings of this study are available on request from the corresponding authors.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Presentation

None.

Acknowledgments

The authors thank Dr Lin Lin, Dr Meixiang Li, Dr Chaocheng He, Dr Cong Hu, Dr Yanqun Chen, Dr Siping Luo, Dr Xinying Wei, Dr Ruibin Wang, and Dr Ji Cui for their valuable contribution in answering questions related to colorectal cancer.

References

[1] Winkler C, Hammoda B, Noyes E, *et al*. Entrepreneurship education at the dawn of generative artificial intelligence. Entrepreneurs Educ Pedag 2023;6:579–89.
[2] Varghese J, Chapiro J. ChatGPT: the transformative influence of generative AI on science and healthcare. J Hepatol 2024;80:977–80.
[3] Introducing ChatGPT. OpenAI Updated November 30, 2022. Accessed 13 November 2023. https://openai.com/blog/chatgpt
[4] Darkhabani M, Alrifaai MA, Elsalti A, *et al*. ChatGPT and autoimmunity – a new weapon in the battlefield of knowledge. Autoimmun Rev 2023; 22:103360.
[5] Uprety D, Zhu D, West HJ. ChatGPT-A promising generative AI tool and its implications for cancer care. Cancer 2023;129:2284–9.
[6] Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? Clin Infect Dis 2023;77:1322–8.
[7] Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023;330: 315–6.
[8] 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:e45312.
[9] Jung LB, Gudera JA, Wiegand TLT, *et al*. ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch Arztebl Int 2023;120:373–4.
[10] Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023;141:589–97.
[11] Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J 2023;99:1110–4.
[12] Zack T, Lehman E, Suzgun M, *et al*. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6:e12–22.
[13] Rao A, Kim J, Kamineni M, *et al*. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 Versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023;20:990–7.
[14] Sorin V, Klang E, Sklair-Levy M, *et al*. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023;9:44.
[15] Barash Y, Klang E, Konen E, *et al*. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol 2023;20:998–1003.
[16] Liu X, Wu C, Lai R, *et al*. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J Transl Med 2023;21:447.
[17] Blum J, Menta AK, Zhao X, *et al*. Pearls and pitfalls of ChatGPT in medical oncology. Trends Cancer 2023;9:788–90.
[18] Thirunavukarasu AJ, Ting DSJ, Elangovan K, *et al*. Large language models in medicine. Nat Med 2023;29:1930–40.
[19] Rahsepar AA, Tavakoli N, Kim GHJ, *et al*. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307: e230922.
[20] Haver HL, Lin CT, Sirajuddin A, *et al*. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions

about lung cancer and lung cancer screening. Am J Roentgenol 2023;221:701–4.
[21] Musheyev D, Pan A, Loeb S, *et al*. How well do artificial intelligence Chatbots respond to the top search queries about urological malignancies? Eur Urol 2024;85:13–6.
[22] Young JN, Ross OH, Poplausky D, *et al*. The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. J Am Acad Dermatol 2023;89:602–4.
[23] Haver HL, Ambinder EB, Bahl M, *et al*. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 2023;307:e230424.
[24] Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023;21:269.
[25] Cao JJ, Kwon DH, Ghaziani TT, *et al*. Accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis. Am J Roentgenol 2023;221:556–9.
[26] 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;29:721–32.
[27] Benary M, Wang XD, Schmidt M, *et al*. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6:e2343689.
[28] Choudhary OP, Priyanka. ChatGPT in travel medicine: a friend or foe? Travel Med Infect Dis 2023;54:102615.
[29] Choudhary OP, Saini J, Challana A, *et al*. ChatGPT for veterinary anatomy education: an overview of the prospects and drawbacks. Int J Morphol 2023;41:1198–202.
[30] GPT-4. OpenAI. Updated March 14, 2023. Accessed 13 November 2023. https://openai.com/research/gpt-4
[31] Brin D, Sorin V, Vaid A, *et al*. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13:16492.
[32] Schubert MC, Wick W, Venkataramani V. Performance of Large Language Models on a Neurology Board-Style Examination. JAMA Netw Open 2023;6:e2346721.
[33] Wang H, Wu W, Dou Z, *et al*. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. Int J Med Inform 2023;177:105173.
[34] Lim ZW, Pushpanathan K, Yew SME, *et al*. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770.
[35] Walker HL, Ghani S, Kuemmerli C, *et al*. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res 2023;25:e47479.
[36] New models and developer products announced at DevDay. OpenAI. Updated November 6, 2023. Accessed 13 November 2023. https://openai.com/blog/new-models-and-developer-products-announced-at-devday
[37] Doctor GPT. OpenAI. Accessed 1 December 2023. https://chat.openai.com/g/g-EiuGnRrIt-doctor-gpt
[38] Try Bard and share your feedback. Google. Updated March 21, 2023. Accessed 13 November 2023. https://blog.google/technology/ai/try-bard/
[39] Team G, Anil R, Borgeaud S, *et al*. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv 2023;231211805. doi:10.48550/arXiv.2312.11805
[40] Bard gets its biggest upgrade yet with Gemini. Google. Updated December 6, 2023. Accessed 31 December 2023. https://blog.google/products/bard/google-bard-try-gemini-ai/
[41] Touvron H, Martin L, Stone K, *et al*. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv 2023;230709288. doi:10.48550/arXiv.2307.09288
[42] Llama 2 is here - get it on Hugging Face. Hugging Face. Updated July 18, 2023. Accessed 13 November 2023. https://huggingface.co/blog/llama2
[43] Welcome Mixtral - a SOTA Mixture of Experts on Hugging Face. Hugging Face. Updated December 11, 2023. Accessed 31 December 2023. https://huggingface.co/blog/mixtral
[44] Introducing Claude 2.1. ANTHROP\C. Updated November 21, 2023. Accessed 31 December 2023. https://www.anthropic.com/index/claude-2-1
[45] Mathew G, Agha R. STROCSS 2021: Strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. IJS Short Reports. 2021;6:e35.
[46] Liu J, Zheng J, Cai X, *et al*. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience 2023;26:107590.
[47] Emile SH, Horesh N, Freund M, *et al*. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? Surgery 2023;174:1273–5.
[48] Bando H, Ohtsu A, Yoshino T. Therapeutic landscape and future direction of metastatic colorectal cancer. Nat Rev Gastroenterol Hepatol 2023;20:306–22.
[49] Ciardiello F, Ciardiello D, Martini G, *et al*. Clinical management of metastatic colorectal cancer in the era of precision medicine. CA Cancer J Clin 2022;72:372–401.
[50] Ciombor KK, Strickler JH, Bekaii-Saab TS, *et al*. BRAF-mutated advanced colorectal cancer: a rapidly changing therapeutic landscape. J Clin Oncol 2022;40:2706–15.
[51] Kuehn BM. More than one-third of US individuals use the Internet to self-diagnose. JAMA 2013;309:756–7.
[52] Ladabaum U, Mannalithara A, Meester RGS, *et al*. Cost-effectiveness and national effects of initiating colorectal cancer screening for average-risk persons at age 45 years instead of 50 years. Gastroenterology 2019;157:137–48.
[53] Lee TC, Staller K, Botoman V, *et al*. ChatGPT answers common patient questions about colonoscopy. Gastroenterology 2023;165:509–511 e7.
[54] Kerbage A, Kassab J, El Dahdah J, *et al*. Accuracy of ChatGPT in common gastrointestinal diseases: impact for patients and providers. Clin Gastroenterol Hepatol 2024;22:1323–25.e3.
[55] Patnaik SS, Hoffmann U. Quantitative evaluation of ChatGPT versus Bard responses to anaesthesia-related queries. Br J Anaesth 2024;132:169–71.
[56] Ali R, Tang OY, Connolly ID, *et al*. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery 2023;93:1090–8.
[57] Amin KS, Davis MA, Doshi R, *et al*. Accuracy of ChatGPT, Google Bard, and Microsoft bing for simplifying radiology reports. Radiology 2023;309:e232561.
[58] Gan RK, Ogbodo JC, Wee YZ, *et al*. Performance of Google bard and ChatGPT in mass casualty incidents triage. Am J Emerg Med 2024;75:72–8.
[59] Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol 2023;34:e13207.
[60] Introducing Claude. ANTHROP\C. Updated March 14, 2023. Accessed 13 November 2023. https://www.anthropic.com/index/introducing-claude
[61] Song H, Xia Y, Luo Z, *et al*. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. J Med Syst 2023;47:125.
[62] Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res 2023;25:e49324.
[63] Wu S, Koo M, Blum L, *et al*. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. arXiv preprint arXiv 2023;230804709. doi:10.1038/s41391-024-00789-0
[64] Claude 2. ANTHROP\C. Updated July 11, 2023. Accessed 13 November 2023. https://www.anthropic.com/index/claude-2
[65] Zhang Y, Wu L, Mu Z, *et al*. Letter 2 regarding "Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma". Clin Mol Hepatol 2024;30:113–7.
[66] Chen S, Kann BH, Foote MB, *et al*. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023;9:1459–62.
[67] Lombardo R, Gallo G, Stira J, *et al*. Quality of information and appropriateness of Open AI outputs for prostate cancer. Prostate Cancer Prostatic Dis 2024. [Online ahead of print]. doi:10.1038/s41391-024-00789-0
[68] Caglar U, Yildiz O, Meric A, *et al*. Evaluating the performance of ChatGPT in answering questions related to benign prostate hyperplasia and prostate cancer. Minerva Urol Nephrol 2023;75:729–33.
[69] Rawte V, Sheth A, Das A. A survey of hallucination in large foundation modelsarXiv preprint*arXiv:230905922*. 2023.
[70] Tonmoy S, Zaman S, Jain V, *et al*. A comprehensive survey of hallucination mitigation techniques in large language modelsarXiv preprint*arXiv:240101313*. 2024.
[71] Giuffre M, You K, Shung DL. Evaluating ChatGPT in Medical Contexts: The Imperative to Guard Against Hallucinations and Partial Accuracies. Clin Gastroenterol Hepatol 2024;22:1145–6.