Family Medicine and Community Health

Performance of generative pre-trained transformers (GPTs) in Certification Examination of the College of Family Physicians of Canada

Mehdi Mousavi ,[1] Shabnam Shafiee,[2] Jason M Harley,[3,4,5] Jackie Chi Kit Cheung,[6,7] Samira Abbasgholizadeh Rahimi [8,9,10,11]

To cite: Mousavi M, Shafiee S, Harley JM, *et al*. Performance of generative pre-trained transformers (GPTs) in Certification Examination of the College of Family Physicians of Canada. *Fam Med Com Health* 2024;**12**:e002626. doi:10.1136/fmch-2023-002626

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/fmch-2023-002626).

For numbered affiliations see end of article.

Correspondence to
Dr Samira Abbasgholizadeh Rahimi;
samira.rahimi@mcgill.ca

ABSTRACT

Introduction The application of large language models such as generative pre-trained transformers (GPTs) has been promising in medical education, and its performance has been tested for different medical exams. This study aims to assess the performance of GPTs in responding to a set of sample questions of short-answer management problems (SAMPs) from the certification exam of the College of Family Physicians of Canada (CFPC).

Method Between August 8th and 25th, 2023, we used GPT-3.5 and GPT-4 in five rounds to answer a sample of 77 SAMPs questions from the CFPC website. Two independent certified family physician reviewers scored AI-generated responses twice: first, according to the CFPC answer key (ie, CFPC score), and second, based on their knowledge and other references (ie, Reviews' score). An ordinal logistic generalised estimating equations (GEE) model was applied to analyse repeated measures across the five rounds.

Result According to the CFPC answer key, 607 (73.6%) lines of answers by GPT-3.5 and 691 (81%) by GPT-4 were deemed accurate. Reviewer's scoring suggested that about 84% of the lines of answers provided by GPT-3.5 and 93% of GPT-4 were correct. The GEE analysis confirmed that over five rounds, the likelihood of achieving a higher CFPC Score Percentage for GPT-4 was 2.31 times more than GPT-3.5 (OR: 2.31; 95% CI: 1.53 to 3.47; p<0.001). Similarly, the Reviewers' Score percentage for responses provided by GPT-4 over 5 rounds were 2.23 times more likely to exceed those of GPT-3.5 (OR: 2.23; 95% CI: 1.22 to 4.06; p=0.009). Running the GPTs after a one week interval, regeneration of the prompt or using or not using the prompt did not significantly change the CFPC score percentage.

Conclusion In our study, we used GPT-3.5 and GPT-4 to answer complex, open-ended sample questions of the CFPC exam and showed that more than 70% of the answers were accurate, and GPT-4 outperformed GPT-3.5 in responding to the questions. Large language models such as GPTs seem promising for assisting candidates of the CFPC exam by providing potential answers. However, their use for family medicine education and exam preparation needs further studies.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Prior to this study, there was an understanding of the general capabilities of artificial intelligence (AI) models like GPTs in various applications and some exams. However, ChatGPT is not specifically designed for medical purposes and there is no specific insights into their performance in open-ended, complex medical examinations like the Certification Examination of the College of Family Physicians of Canada (CFPC). The need for this study stemmed from the growing integration of AI in medical education and the potential of large language models such as GPTs in preparing for this complex exam.

WHAT THIS STUDY ADDS

⇒ This study demonstrates that the latest iteration of ChatGPT, particularly GPT-4, can accurately respond to a significant portion of CFPC examination questions. It reveals that GPT-4 notably outperforms its predecessor, GPT-3.5, in both the accuracy and efficiency of responses. Moreover, the study indicates that the timing and conditions under which ChatGPT is queried, along with the regeneration of answers and the strategic use of prompts for debriefing, might not significantly impact the accuracy and consistency of the responses.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The findings from this study could influence future research directions, focusing on incorporating advanced AI models such as GPTs in medical education and examination preparation. It suggests a new, innovative method for medical students and professionals to prepare for examinations. Policy-wise, it could open discussions on the role of AI in formal medical education and certification processes, potentially leading to the integration of AI as a standard tool in medical learning and assessment.

BACKGROUND

ChatGPT, released by Open AI (San Francisco, California, USA) in November 2022, is an advanced large language model (LLM) to

generate humane-like dialectic responses to text inquiries. ChatGPT, by default, uses the generative pre-trained transformer (GPT) 3.5 model, specifically designed for conversational application and is the freely accessible version. In contrast, ChatGPT Plus uses GPT-4.0, which is claimed to be a more accurate and efficient tool with improved and safer responses to complex problems.[1]

Several potential implications have been described for ChatGPT in medical education. These include the creation of clinical vignettes to help with the training and evaluation of healthcare professionals[2]; answering specific questions related to various medical encounters, including diagnoses or treatments[3]; generating exercises and quizzes for teaching purposes[3]; generating lists of differential diagnoses[4 5]; and facilitating self-directed learning[5 6] by creating helpful mnemonics.[5] However, while AI-based chatbots offer valuable contributions to medical learning, ethical concerns exist about their use in education and research. For instance, data privacy and security are essential considerations when employing chatbots.[6 7]

ChatGPT has demonstrated promising outcomes in reputable medical examinations, suggesting its potential utility in medical exam preparation.[6] In an official multiple-choice progress test, GPT-3.5's performance was comparable to that of family medicine residents from the University of Toronto, while GPT-4 outperformed both groups.[8] Moreover, ChatGPT has also been used to answer the different steps of the United States Medical Licensing Examination (eg, USMLE,[9–11] membership of the Royal College of General Practitioners Applied Knowledge Test (AKT),[12] ophthalmology,[13] neurology[14] and radiology[15] specialty exams. ChatGPT's performance has been acceptable not only in English-based medical exams but also in tests conducted in other languages; for instance, the Japanese medical licensing examination,[16] the Chinese National Medical Licensing Examination (NMLE)[17] and the Iranian Medical Residency Examination.[18]

The Certification Examination in Family Medicine conducted by the College of Family Physicians of Canada (CFPC) is a comprehensive assessment of broad clinical knowledge in the field of family medicine in Canada.[19] This exam consists of the oral component, the simulated office oral exam, and the written component, consisting of short-answer management problems (SAMPs). Typically, SAMPs include around 40 clinical scenarios, with two to seven questions for each scenario.

The rapid expansion and widespread accessibility of LLM-based AIs have increased their use in medical education and medical exam preparation.[3 6 8–11] Nevertheless, ChatGPT is not specifically designed for medical purposes and may not be accurate in this domain. Therefore, it is unclear if it could be employed to help candidates find potential answers to SAMP questions for CFPC exam preparation. Huang *et al* compared the performance of GPT-3.5 and GPT-4 with that of family medicine residents at the University of Toronto, employing an official multiple-choice medical knowledge test sourced from their university, designed for preparation for the SAMPs exam.[8] However, to our knowledge, no studies have assessed LLMs' capacity to assist candidates in preparing for the open-ended questions. Furthermore, it remains uncertain whether factors such as questioning ChatGPT at different times, regenerating answers or employing (different) prompts for debriefing could influence the accuracy of responses. Therefore, we conducted this study to assess the performance of both GPT-3.5 and GPT-4, in addressing a series of sample open-ended SAMPs questions. Additionally, we examined the consistency and accuracy of ChatGPT and ChatGPT Plus responses in various rounds with different contexts.

METHODS

Dataset

We conducted this study using all the questions from a sample set of SAMPs obtained from the official website of the CFPC.[19] This sample set comprises 19 clinical scenarios, accompanied by a total of 77 questions related to these scenarios. Each scenario has between two and seven associated questions designed to simulate the format of the actual computer-based examination. These questions require brief, concise responses and typically, answers should consist of no more than 10 words per line, with each question necessitating 1 to 5 lines of response. The clinical scenarios spanned various domains within family medicine, such as cardiology, neurology and emergency, except for dermatology (table 1).

Data collection

We employed GPT-3.5 (ChatGPT, August 3, Version 2023, OpenAI, San Francisco, California, USA) and GPT-4 (ChatGPT Plus, August 3, version 2023, OpenAI, San Francisco, California, USA) from 8 August to 24 August 2023, to respond to these sample SAMPs-CFPC questions.

Initially, we experimented with one or two scenarios and found that while ChatGPT's answers were highly informative and valuable for learning, they were also textually rich (average of 150 words per answer), while the amount of text that is required in the real test is the short answers. As a result, we introduced a prompt before each run to limit the answers to fewer than 10 words per line to simulate the actual real exam format. However, in the final round (ie, the fifth round), we included a session without a prompt for comparison purposes. In each instance, we presented ChatGPT with the scenario, followed by its related questions, without repeating the clinical scenario. To eliminate the potential impact of memory retention bias (ie, the tendency of GPT to remember the responses from the previous round of questions and answers), we copied all the responses into a Word document. Subsequently, we completely erased all the conversations of that round from the ChatGPT window before initiating a new session for the subsequent round. Table 2 summarises the various rounds in which both GPT-3.5 and GPT-4 were used for our study.

Table 1 Diverse topics within the spectrum of family medicine represented in the sample SAMP questions from the CFPC website[19]

Category	Number of cases	Topic
Cardiology	3	Atrial fibrillation, dizziness, hypertension
Endocrine	1	Osteoporosis
Emergency	1	Poisoning
Surgical skills	1	Laceration
Gastroenterology	3	Abdominal pain, abnormal liver function test, dyspepsia
Musculoskeletal disease	1	Low back pain
Neurology	1	Seizure
Women's health and obstetrics	2	Breast lump, fertility
Paediatrics and infectious disease	1	Fever in newborn
Psychiatry and mental health	2	Depression, eating disorder
Respirology	1	Chronic obstructive pulmonary disease
Urology	1	Prostate
Other topics	1	Fatigue

CFPC, College of Family Physicians of Canada; SAMP, short-answer management problem.

Scoring and review of responses

Two experienced CFPC-certified practicing family physicians independently reviewed and scored the AI-generated responses and explanations. Both reviewers, MM and SS, possess over 2 years of Canadian family medicine practice experience. However, they are International Medical Graduates with over 15 years of extensive professional backgrounds. MM has also a background of over 16 years of experience in medical education and currently serves as a faculty member in the Department of Family Medicine at a Canadian university.

First, the reviewers strictly adhered to the answer key provided on the CCFP website[19] for scoring, which we refer to as 'CFPC Scoring'. Initially, the two reviewing physicians scored the answers independently, blinded to each other. Responses that were entirely incorrect for each line received a score of zero during the evaluation process, while those deemed accurate were assigned a score of one per line. Following their initial evaluations, the two reviewers observed that 71 out of 77 CFPC Score Percentages (92.2%) were identical. Subsequently, after a collaborative discussion, they reached a consensus on the final score for all questions (100%). A fractional scoring system of 0.5 was employed in certain instances, deviating from the binary scale of zero or one for a line of answers. Subsequently, the total score for each question (comprising the total lines of correct answers) was divided by the maximum possible score for each question and then multiplied by 100 to derive the 'CFPC Score Percentage'. However, the reviewers noted that ChatGPT mainly produced accurate and acceptable answers based on their expertise, although absent in the official answer key. Consequently, they did a second round of scoring. In this second scoring, the reviewers jointly reassessed the responses simultaneously, using the latest version of UpToDate (August 2023),[20] and agreed completely on the 'Reviewer's Score'.

Additionally, to assess the consistency of the answers between rounds, we used the 'Percentage of Repeated Answers' for each question. To compute this percentage, we compared each round with a selected reference round to determine the extent of repetition of the same concepts within the answers for each question.

Finally, each question's difficulty level was evaluated based on the reviewers' judgments. The questions were classified as difficult if they were textually dense and complex questions that required the responder to judiciously weigh multiple clinical indicators while eliminating various potential answers based on the cues provided in the question. Conversely, questions that did not exhibit these characteristics and mostly needed one-word answers were classified as easy.

Table 2 Summary of the multiples runs using GPT-3.5 and GPT-4 between 8 August and 24 August 2023

Rounds	Date that GPT-3.5 was used	Date that GPT-4 was used
Round 1: we used 'Prompt 1*' for the first time	8, 10 August	11, 12 August
Round 2: we used 'Prompt-1*' for the second time	15, 16 August	19 August
Round 3: we regenerated the answers of the second run	15, 16 August	19 August
Round 4: we used 'Prompt 2†'	16, 17, 19 August	21 August
Round 5: we answered the set of questions without any prompt	19, 24, 25 August	23, 24 August

*Prompt 1: we started the ChatGPT run with the following prompt—'Hello ChatGPT, I am going to ask you questions from the CFPC exam. There is a clinical scenario with subsequent questions about it. Please limit your answers to less than 10 words per line. When asked to give several answers provide the best possible answers to the question'.
†Prompt 2: we removed the word CFPC exam from the prompt. We started the ChatGPT run with the following prompt—'Hello ChatGPT, I am going to ask you some questions. There is a clinical scenario with subsequent questions about it. Please limit your answers to less than 10 words per line. When asked to give several answers provide the best possible answers to the question'.

Data analysis

We conducted our statistical analyses using SPSS V.16.0 software (SPSS Inc, Chicago, Illinois, USA). We presented the results as median values with the (25th and 75th percentiles) for variables that did not follow a normal distribution and reported the mean (SD) only for comparative purposes. Categorical variables were reported as numbers (percentages). We examined differences in the scores assigned to each question by GPT-3.5 and GPT-4 using the Wilcoxon signed-rank non-parametric test. To compare the outcome of repeated measures across five rounds of GPT-4 and GPT-3.5 results, we used the ordinal logistic generalised estimating equation (GEE). The outcome variables were the CFPC score, or Reviewers' Score, categorised as 0, 33.3, 50, 66.67, 75, 80 and 100. The independent variable was the usage of GPT-3.5 vs GPT-4 to answer the questions across the five rounds. We employed an independent working correlation matrix structure in the GEE analysis with link function of cumulative logit. All the reported p values were two-sided, with a significance level of ≤0.05 considered statistically significant.

Ethical considerations

This study exclusively used and analysed publicly available data and did not involve human participants. Consequently, there was no requirement for approval from the Review Board of McGill University. The authors have no conflicts of interest to disclose.

Patient and public involvement

Patients and the public were not involved in the design, recruitment, conduct or any other stages of the research process in this study.

RESULTS

We evaluated 19 clinical scenarios, each with two to seven pertinent questions. These scenarios included 77 specific questions, generating 165 lines of answers. The possible responses to each question varied in length, ranging from 1 to 5 lines, with a median length of 2 (1, 3). The two reviewers categorised 28 questions (36.4%) as easy, and 49 (63.6%) as difficult. Both reviewers agreed that the answers given by ChatGPT in the fifth round without any prompts were very informative and valuable for education and better understanding.

Over five rounds, out of 852 lines of answers, 607 (73.6%) provided by GPT-3.5 and 691 (81%) offered by GPT-4 were deemed correct based on the CFPC answer key. The mean CFPC score percentage for all five rounds was 76.0 for GPT-3.5 and 85.2 for GPT-4. The mean Reviewers' Scores for GPT-3.5 and GPT-4 were 86.1 and 93.4, respectively. The GEE analysis revealed that the likelihood of achieving a higher CFPC score percentage was significantly greater for GPT-4 compared with GPT-3.5, with GPT-4 being 2.31 times more likely to score higher (OR: 2.31; 95% CI: 1.53 to 3.47; p<0.001). Similarly, over five rounds, the Reviewers' Score percentage for responses provided by GPT-4 were found to be significantly higher, being 2.23 times more likely to exceed those of GPT-3.5 (OR: 2.23; 95% CI: 1.22 to 4.06; p=0.009).

The results of five distinct rounds using GPT-3.5 and GPT-4 to respond to the sample CFPC questionnaire are presented in table 3. Comparing the results of GPT-3.5 and GPT-4 showed that CFPC scores were significantly higher for GPT-4 as opposed to GPT-3.5 for rounds 1, 3, 4 and 5, and we noted a trend towards an increase in

Table 3 Comparison of accuracy of answering GPT-3.5 and GPT-4 using a percentage of the score for each question across five rounds

	CFPC score percentage for GPT-3.5 Answers to Each Question (%)	CFPC score percentage for GPT-4 Answers to each question (%)	P value	Reviewers' Score percentage for GPT-3.5 Answers to Each Question (%)	Reviewers' Score percentage for GPT-4 Answers to each question (%)	P value
Round 1	100 (50,100)	100 (78,100)	**0.002**	100 (90, 100)	100 (100, 100)	**0.017**
	73.7 (33.9)	85.0 (27.8)		84.0 (31.0)	91.9 (23.3)	
Round 2	100 (50, 100)	100 (71, 100)	0.113	100 (100, 100)	100 (100, 100)	**0.015**
	73.9 (35.0)	81.2 (32.7)		85.6 (30.3)	94.4 (19.1)	
Round 3	100 (55,100)	100 (77, 100)	**0.005**	100 (100, 100)	100 (100, 100)	**0.037**
	79.0 (30.0)	86.4 (26.3)		88.8 (26.1)	93.7 (21.3)	
Round 4	100 (67, 100)	100 (80, 100)	**0.011**	100 (100, 100)	100 (100, 100)	**0.014**
	80.2 (29.0)	87.3 (25.5)		87.1 (27.5)	94.8 (19.4)	
Round 5	100 (50, 100)	100 (75, 100)	**0.003**	100 (87.5, 100)	100 (100, 100)	0.121
	73.1 (34.0)	86.2 (23.9)		85.2 (29.6)	92.1 (21.5)	

Data are presented as median (25 percentile, 75 percentiles). Mean (SD) is given for comparison. Wilcoxon signed-rank test was done to compare GPT-3.5 scores and GPT-4 scores. P values <0.05 (in bold) were considered as statistically significant.
CFPC Score Percentage: percentage of score given to the questions according to College of Family Physicians of Canada (CFPC) answers key; Reviewers' Score Percentage: percentage of scores given to the questions according to the reviewers' knowledge; Round: for definition of rounds see table 2.
CFPC, College of Family Physicians of Canada.

Table 4 Median (25th, 75th) and mean (SD) of the percentage of repeated answers in the comparison of GPT-3.5 and GPT-4 (Wilcoxon signed-rank non-parametric test for comparison), left and percentage of questions that did not show a change in score for CFPC and Reviewers' Score

	Percentage of repeated answers to each question		P value	Percentage of questions with no change to the CFPC score		Percentage of questions with no change to the Reviewer's Score	
	GPT-3.5	GPT-4		GPT-3.5	GPT-4	GPT-3.5	GPT-4
Round 1 and 2 comparisons	100 (66.7, 100) 82.0 (27.9)	100 (80, 100) 88.7 (19.6)	**0.025**	61 (79.2%)	65 (84.4%)	66 (85.7%)	70 (90.9%)
Round 1 and 4 comparisons	100 (63.3, 100) 79.8 (27.9)	100 (83.3, 100) 90.6 (17.2)	**0.002**	55 (71.4%)	70 (90.9%)	60 (77.9%)	72 (93.5%)
Round 1 and 5 comparisons	100 (50, 100) 76.6 (32.2)	100 (75, 100) 89.2 (17.4)	**<0.001**	52 (67.5%)	64 (83.1%)	60 (77.9%)	70 (90.9%)
Round 2 and 3 comparisons	100 (70.8, 100) 85.0 (26.3)	100 (80, 100) 89.5 (19.6)	0.167	65 (84.4%)	69 (89.6%)	69 (89.6%)	72 (93.5%)

The total number of questions was 77. P values <0.05 (in bold) were considered statistically significant.

rounds 2 (table 3). The right side of the table represents the 'Reviewers' Score Percentage' for GPT-3.5 and GPT-4 answers to each question. Similar to the CFPC Score Percentages, the Reviewers' Score Percentages assigned by GPT-4 tended to be higher in round 5 and were significantly higher in rounds 1, 2, 3 and 4 (table 3).

GPT-3.5 exhibited consistent repetition of the same concepts in the answers across all five rounds in 31 out of 77 questions (40.3%), whereas GPT-4 repeated the same concepts in 37 out of 77 questions (48.1%). Table 4 compares GPT-3.5 and GPT-4 regarding the percentage of repeated answers for each question on the left side and the percentage of questions with no change to the CFPC and Reviewers' Score on the two right columns, respectively.

When comparing the responses to each question in rounds 1 and 2 (with an approximate 1 week interval, as shown in table 2), there was no significant change in the 'CFPC Score Percentage' for both GPT-3.5 and GPT-4 (p=0.79 for GPT-3.5 and p=0.26 for GPT-4 respectively, Wilcoxon signed-rank test). Both GPT-3.5 and GPT-4 consistently demonstrated a high percentage of repeated answers for each question, approximately 80% (table 4), with mean percentages of 82.0 and 88.7, respectively (table 4). However, the percentage of repeated answers was higher for GPT-4 (p=0.025, table 4). Among the answers that differed between rounds 1 and 2, the CFPC or Reviewers' Scores predominantly remained unchanged for both GPT-3.5 and GPT-4 (table 4).

In round 4, we excluded the term 'CFPC exam' from 'Prompt 1', which was used in round 1 (table 2). The 'CFPC Score Percentage' was significantly higher for round 4 compared with round 1 (p=0.014) for GPT-3.5, but this trend was not significant for GPT-4 (p=0.089). The percentage of repeated answers was found to be higher for GPT-4 than for GPT-3.5 (p=0.002, table 4). Additionally, the scores remained largely unchanged, particularly for GPT-4 (table 4, last two columns on the right).

Comparing round 5 (without any prompt) and round 1 (with prompt 1) showed no significant difference in 'CFPC Score Percentage' for both GPT-3.5 and GPT-4 (p=0.83and p=0.72, respectively). However, GPT-4 showed a higher percentage of repeated answers than GPT-3.5 (p<0.001, table 4). Most of the scores remained unchanged, similar to previous comparisons (table 4, the two columns on the right side).

Lastly, round 3 was a regeneration of responses from round 2. When comparing these two rounds, the 'CFPC Score Percentage' tended to increase for GPT-3.5 and GPT-4 (p=0.058and p=0.098, respectively), while remaining unchanged for GPT-4. The percentages of repeated answers were not significantly different between GPT-3.5 and GPT-4 (table 4). Like other comparisons, most scores remained unchanged between these two rounds (table 4, the two columns on the right side).

Online supplemental appendix box 1 presents an illustrative CFPC sample question along with responses generated by GPT-3.5 and GPT-4 across multiple rounds.

DISCUSSION

In this study, we used GPTs to answer the Sample CFPC questions and responded satisfactorily to our complex sample questions. When the reviewers scored the questions using the fixed answer key provided by the CFPC website, the mean score for all five rounds was 76.0±27.7 for GPT-3.5 and 85.2±23.7 for GPT-4. Additionally, the authors found that most of the answers, although not explicitly stated in the answer key, were reasonable and acceptable, and only about 16% of the lines of answers provided by GPT-3.5 and 7% of the lines of answers provided by GPT-4 were deemed incorrect in the Reviewers' scoring.

Although ChatGPT has been used to respond to medical examination questions,[6 9–18] only one study has evaluated its efficacy in preparing for the Canadian family medicine exam.[8] In this study, Huang and colleagues demonstrated that GPT-4 significantly outperformed the other test takers, achieving an impressive accuracy rate of 82.4%, whereas GPT-3.5 achieved 57.4% accuracy, and

family medicine residents scored 56.9% correctly.[8] In our study, the mean CFPC score across five rounds was 85.2 for GPT-4, which closely resembled their score, while GPT-3.5 scored lower at 76.0. However, it is important to note that Huang and his team's questionnaire comprised multiple-choice questions, differing from the open-ended format of the questions in the SAMPs exam. Furthermore, their questionnaire was sourced from their university, specifically designed to prepare their family medicine residents for the exam and may lack standardisation. In contrast, our study employed a comprehensive and standardised set of questions sourced directly from the CFPC website. These questions were open-ended, mirroring the SAMPs structure, and included official answer keys approved by CFPC, providing a more accurate representation of the CFPC exam format.

Thirunavukarasu and coworkers used GPT-3.5 to answer the AKT exam designed for Membership of the Royal College of General Practitioners in the UK. They achieved a performance level of 60.17%, which was lower than our score, and it fell short of the 70.45% passing threshold in this primary care examination.[12] Nevertheless, like the University of Toronto study,[8] this study employed a multiple-choice questionnaire and was not specific to a Canadian family medicine exam. Other studies have reported similar scores for GPT-3.5 on various medical examinations at the undergraduate level. Kung and colleagues reported that ChatGPT achieved near-passing accuracy levels of around 60% for Step 1, Step 2 of CK and Step 3 of the USMLE.[9] Similarly, Gilson and colleagues observed an accuracy range of 44% to 64.4% for sample USMLE Step one and Step two questions.[10] ChatGPT's performance on the Chinese NMLE stayed behind that of medical students and was below the passing threshold.[17]

Similar to our study, scores were higher when GPT-4 was used instead of GPT-3.5 in other studies. For instance, while GPT-3.5 fell short of the passing criteria for the Japanese medical licensing examination, GPT-4 met the threshold criteria.[16] Nori *et al* used GPT-4 and observed a passing score on USMLE by over 20 points.[11] Finally, GPT-4 accurately answered 81.3% of the questions on the Iranian Medical Residency Examination.[18]

The combined analysis of five rounds using the GEE model revealed that the CFPC Score Percentages were significantly higher for GPT-4 than GPT-3.5 ($p<0.001$). Likewise, on re-evaluating the responses using their medical expertise, the Reviewers' Score percentages for GPT-4 over five rounds were significantly higher for GPT-4 compared with GPT-3.5 ($p=0.009$). This finding is probably because GPT-4 is able to perform more efficiently under challenging questions from complex situations.[3 4] This trend has been previously shown through assessments of ChatGPT (GPT-3.5) and ChatGPT Plus (GPT-4) on various exams, including a sample of multiple choice progress tests from the University of Toronto,[8] two sets of official practice materials for the USMLE exam from the National Board of Medical Examiners,[11] the Japanese Medical Licensing Examination,[16] the StatPearls ophthalmology Question Bank[13] and the 2022 SCE neurology examination.[14] However, other studies primarily involved multiple-choice questions,[8 11] were related to the undergraduate level,[11] were conducted in different languages[16] or focused on other specialties.[13 14] Our study focused on the complex task of open-ended Canadian family medicine questions and demonstrated that GPT-4 can provide more accurate answers to complex Canadian SAMPs exam questions than GPT-3.5 (the free version).

In the fifth round of our study, when AI was not specifically instructed to offer brief responses, it consistently provided informative justifications and reasoning. These responses were highly instructive and aligned well with our educational objectives (see online supplemental appendix box 1). Therefore, our study demonstrated that GPT-3.5 and GPT-4 can be used to guess the answers to complex tasks such as those outlined in the study, making it a potential help for CFPC exam preparation. However, using these technologies to learn family medicine and prepare for exams needs further study.

Despite several benefits and potential roles of LLMs in medical education and research, they have several pitfalls. These pitfalls include the absence of up-to-date sources of literature[1] (the current versions of ChatGPT are trained in September 2021), inaccurate data,[13 14] inability to distinguish between fake and reliable information,[21] generating incorrect answers known as hallucinations,[6 7 21–24] which is potentially misleading or dangerous in a healthcare context.[7 24] ChatGPT is still in an experimental phase and is not intended for medical application.[7] Therefore, using ChatGPT in preparation for exams should serve as a prompt to reinforce existing knowledge derived from reliable sources. Responses generated by ChatGPT should undergo rigorous fact-checking by human experts before being considered a primary knowledge resource.

Our testing comprised several rounds, including repeating identical prompts at intervals, modifying the prompts by eliminating the reference to 'CFPC exam' from the prompts, regenerating responses and removing prompts to evaluate outcomes. When comparing rounds 1 and 2 with a similar 'Prompt 1' but with an approximately 1 week interval, both GPT-3.5 and GPT-4 demonstrated high consistency and accuracy. This observation suggests that the passage of time does not significantly impact the chatbot's performance. Instead, future improvements may arise through the AI's learning curve and the introduction of newer versions of LLMs trained on updated material, warranting further investigation.

Removing the phrase 'CFPC exam' in round 4 led to an unexpected outcome. The accuracy, indicated by 'CFPC Score Percentage', noticeably increased for GPT-3.5 and showed an upward GPT-4 trend contrary to our initial hypothesis. We speculated that omitting the exam's name might limit GPT's access to the source questions, potentially reducing scores. However, the observed increase may be accidental or suggest other underlying factors, necessitating further investigation to understand these results.

The comparison between rounds 1 and 5 aimed to determine whether prompting influenced responses and resulted in consistently accurate outcomes. The absence of significant change for 'CFPC Score Percentage' for both GPT-3.5 and GPT-4 may suggest that prompting did not significantly alter the accuracy of the responses. Also, in most of the questions, the CFPC score remained unchanged (67.5% for GPT-3.5 and 83.1% for GPT-4). This result suggests that running ChatGPT without any prompt could lead to detailed responses with justifications with similar accuracy, which could be valuable for candidates preparing for the CFPC exam.

Finally, the regeneration of responses from round 2 in round 3 was conducted to assess whether response regeneration could enhance accuracy. We removed the output from each round except for the third run, a repetition of the second run, to minimise potential learning curve effects on the GPT's performance. As a result of this approach, the 'CFPC Score Percentage' tended to increase for GPT-3.5, while remaining unchanged for GPT-4. This finding may further emphasise that regeneration of the responses may improve the results for GPT-3.5 but not GPT-4.

In summary, GPT-4 showed considerable consistency in our comparisons. This consistency was more impressive when the reviewers realised that changing the answer choices by GPT would not impact the scores (table 4). In most cases, GPT-4 repeated answers more frequently than GPT-3.5 or at least showed a trend of higher repetition. In a related study, Thirunavukarasu *et al* conducted two independent sessions of the AKT exam using ChatGPT for 10 days and observed consistent performance.[12]

Study limitation

It is important to acknowledge that there is no established cut-off score for passing the SAMPs part of the CFPC exam. Instead, the minimal passing score is set based on the performance of a reference group of first-time test-takers who graduate from Canadian family medicine residency programmes in each exam.[19] Consequently, whether ChatGPT's current performance would be sufficient to pass the exam remains inconclusive. Additionally, we lack access to the scores of candidates, making it impossible to compare ChatGPT's performance with that of human candidates. Comparing ChatGPT's performance in answering a sample question with that of candidates could potentially reveal whether ChatGPT outperforms or is not inferior to human candidates. It is necessary to emphasise that ChatGPT is not designed to practice family medicine or pass the related exam. Instead, we may propose that it could be used to assist candidates with exam preparation by helping them determine correct responses.

A significant component of learning in family medicine involves the interpretation of images, such as ECGs, X-rays and skin conditions—capabilities that text-based models like ChatGPT lack. In our study, we encountered this limitation when one question included an ECG image, which we had to exclude the image. Interestingly, our two reviewers found that the absence of this image did not impact the accuracy or relevance of ChatGPT's answers to the associated clinical scenario question.

In this study, we used GPT-3.5 and GPT-4 from OpenAI, which were trained in September 2021 and were not specialised for medical purposes.[1] It's important to note that other LLMs may use more recent sources of information, potentially yielding different results and warranting further investigation. Furthermore, even within the same version of OpenAI, the GPT's performance can be influenced by the repetition of questions and the feedback provided over time, meaning that the performance of ChatGPT may evolve over time. To avoid the possibility of learning curve effects and memory retention bias impacting the AI's performance, we took the precaution of erasing the results of each round from the ChatGPT window before initiating a new session for the subsequent round.

In an actual exam setting, residents typically read the clinical scenario once and then respond to each two to seven related questions and the scenario is not reaped before each question. We adopted a similar approach and did not reiterate the clinical scenario before each related question. Nevertheless, ChatGPT's responses might differ if the clinical scenario were repeated before each question. Confirming this hypothesis would necessitate further investigation.

In this study, we examined a sample of SAMP questions provided by CFPC, which is very similar to the actual exam. These question sets comprised only 19 clinical scenarios and 77 questions. Expanding the number of questions examined could enhance the study's reliability. However, it's important to note that many of the available sample questions from other sources on the market may not represent the actual examination, or their answer keys may be reliable.

CONCLUSION

Given the high accuracy and consistency of the answers generated by ChatGPT—particularly GPT-4 —our study suggests that these GPTs are promising as supplementary learning tools for candidates preparing for the CFPC exam. Future studies need to assess the long-term efficacy and reliability of these models in educational settings, especially in preparing candidates for exams like the CFPC. This would involve tracking performance over multiple years and across various curriculum updates and study how the use of these AI-enabled tools influences learning behaviours, including understanding of complex concepts, and critical thinking skills.

Author affiliations
[1]Department of Family Medicine, Faculty of Medicine, University of Saskatchewan, Nipawin, Saskatchewan, Canada
[2]Department of Family Medicine, Saskatchewan Health Authority, Riverside Health Complex, Turtleford, Saskatchewan, Canada
[3]Department of Surgery, Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada
[4]Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
[5]Institute for Health Sciences Education, Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada
[6]McGill University School of Computer Science, Montreal, Quebec, Canada

[7]CIFAR AI Chair, Mila-Quebec AI Institute, Montreal, Quebec, Canada
[8]Department of Family Medicine, McGill University, Montreal, Quebec, Canada
[9]Mila Quebec AI-Institute, Montreal, Quebec, Canada
[10]Faculty of Dentistry Medicine and Oral Health Sciences, McGill University, Montreal, Quebec, Canada
[11]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada

Acknowledgements SAR is Canada Research Chair (Tier II) in Advanced Digital Primary Health Care, received salary support from a Research Scholar Junior 1 Career Development Award from the Fonds de Recherche du Québec-Santé (FRQS) during a portion of this study, and her research program is supported by the Natural Sciences Research Council (NSERC) Discovery (grant 2020-05246).

Contributors MM: software, formal analysis, investigation, data curation, writing–original draft, visualization; SS: formal analysis, writing-revision and edits; JH: conceptualization, writing-revision and edits; JCKC: conceptualization, writing-revision and edits; SAR: conceptualization, methodology, data analysis, supervision, project administration, writing-revision and edits and acting as guarantor. In this research, we have used ChatGPT (GPT-3.5 and GPT-4) to answer a sample question from the College of Family Physicians of Canada questionnaire. We evaluated the AI-generated responses and used Grammarly Premium to edit our manuscript draft for grammar and punctuation.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iDs
Mehdi Mousavi http://orcid.org/0000-0003-3644-5741
Samira Abbasgholizadeh Rahimi http://orcid.org/0000-0003-3781-1360

REFERENCES

1 OpenAI. Models: OpenAI. 2023. Available: https://beta.openai.com/docs/models
2 Benoit JRA. ChatGPT for clinical vignette generation, revision, and evaluation. *medRxiv* 2023;2023.
3 Khan RA, Jawaid M, Khan AR, *et al*. ChatGPT - reshaping medical education and clinical management. *Pak J Med Sci* 2023;39:605–7.
4 Hirosawa T, Harada Y, Yokose M, *et al*. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 Chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health* 2023;20:3378.
5 Wang L-P, Paidisetty PS, Cano AM. The next paradigm shift? ChatGPT, artificial intelligence, and medical education. *Medical Teacher* 2023;45:925.
6 Sallam M. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023;11:887.
7 Li J, Dada A, Puladi B, *et al*. Chatgpt in healthcare: a taxonomy and systematic. *Comput Methods Programs Biomed* 2024;245:108013.
8 Huang RS, Lu KJQ, Meaney C, *et al*. Assessment of resident and AI Chatbot performance on the University of Toronto family medicine residency progress test: comparative study. *JMIR Med Educ* 2023;9:e50514.
9 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:e0000198.
10 Gilson A, Safranek CW, Huang T, *et al*. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.
11 Nori H, King N, McKinney SM, *et al*. Capabilities of Gpt-4 on medical challenge problems. *arXiv Preprint arXiv* 2023;230313375.
12 Thirunavukarasu AJ, Hassan R, Mahmood S, *et al*. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med Educ* 2023;9:e46599.
13 Moshirfar M, Altaf AW, Stoakes IM, *et al*. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering Statpearls questions. *Cureus* 2023;15:e40822.
14 Giannos P. Evaluating the limits of AI in medical Specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. *BMJ Neurol Open* 2023;5:e000451.
15 Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307:e230582.
16 Takagi S, Watari T, Erabi A, *et al*. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023;9:e48002.
17 Wang X, Gong Z, Wang G, *et al*. Chatgpt performs on the Chinese national medical licensing examination. *In Review* [Preprint].
18 Khorshidi H, Mohammadi A, Yousem DM, *et al*. Application of ChatGPT in multilingual medical education: how does ChatGPT fare in 2023's Iranian residency entrance examination. *Informatics in Medicine Unlocked* 2023;41:101314.
19 The College of Family Physicians of Canada. Preparing for the certification examination in family medicine. 2023. Available: https://www.cfpc.ca/en/education-professional-development/examinations-and-certification/certification-examination-in-family-medicine/preparing-for-the-certification-examination-in-fam
20 UpToDate. 2023. Available: https://www.uptodate.com/contents/search
21 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;6:1169595.
22 Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 2023;13:410.
23 Abbasgholizadeh Rahimi S, Légaré F, Sharma G, *et al*. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. *J Med Internet Res* 2021;23:e29839.
24 Akinci D'Antonoli T, Stanzione A, Bluethgen C, *et al*. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol* 2024;30:80–90.