Use of Large Language Models for Family medicine education and examination

Mehdi Mousavi

Department of Family Medicine

McGill University, Montreal

December 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Family Medicine

TABLE OF CONTENTS

ABSTRACT

Background

The continuous emergence of new LLMs, specialized medical domain training, and access to online resources brings the potential for advantages of large language models (LLMs) for medical education and preparing for medical exams.

In this thesis, we have performed two studies to evaluate the usefulness of ten LLM-based chatbots on a sample questionnaire for the Canadian family medicine exam to evaluate the accuracy of the responses and to explore which aspects of these LLMs could be more helpful in this context.

Methods

We utilized a sample of the Short-Answer Management Problems (SAMPs) questionnaire from the College of Family Physicians of Canada (CFPC) official website. This sample set included 19 clinical scenarios covering various domains of family medicine, each comprising 2 to 7 questions, resulting in a total of 77 questions that required 165 lines of responses.

In the first study, we ran GPT-3.5 and GPT-4 (ChatGPT, OpenAI, San Francisco, CA) five times between August 8th and 25th, 2023, to respond to the questionnaire. We provided a prompt limiting response to 10 words per line for the first round. This procedure was repeated a week later in the second round. The third round involved immediately regenerating answers from the second run. The fourth round used a different prompt that did not reference CFPC exam questions. Finally, the fifth round was conducted without prompting.

Two CFPC-certified reviewers with an average experience of 15 years assessed the responses based on the CFPC answer key. Reviewers were blinded to each other's assessments but resolved any discrepancies through multiple meetings. We applied an ordinal logistic Generalized Estimating Equations (GEE) model to analyze the repeated measures across the five rounds.

In our second study, conducted in February and March 2024, we examined ten LLM-based chatbots as follows: (1) General Domain LLMs without online searching, including GPT-3.5 and

GPT-4 by Open AI, Gemini Pro 1.0 and Gemini Ultra 1.0 by Google and Personal Intelligence by Inflection AI (using Inflection 2.5); (2) LLM-based chatbots that include online search capabilities, such as Smart mode of YouChat (utilizing GPT-3), GPT-4 mode of YouChat, and Copilot by Microsoft (using GPT-4); (3) LLMs trained on medical knowledge: i.e., AI doctor from Docus AI (using GPT-4), and Chat Doctor created by meta-AI (using LLaMA).

Similar to the first study, three CFPC-certified reviewers scored the LLM-generated responses. We used Fleiss's Kappa to measure agreement among the reviewers' post-meetings.

Results

Following the discussion, both reviewers agreed on all the scores from the initial study. According to the CFPC answer key, GPT-3.5 provided accurate responses in 607 out of 825 lines of answers (73.6%), while GPT-4 achieved accuracy in 691 (81%) cases. GEE analysis revealed that GPT-4 was 2.31 times more likely to attain a higher CFPC score than GPT-3.5 (Odds Ratio: 2.31; 95% Confidence Interval: 1.53 to 3.47; P<0.001). Other findings from five rounds in the first study included consistency over a week, no significant impact of the 10-word limit prompt on accuracy, and no substantial change upon removing the CFPC exam from the initial prompt. In the second study, the Kappa score ranged from 0.89 to 1, indicating perfect consensus. ChatGPT-4 demonstrated the highest accuracy rate among the ten systems at 85.5%, followed by AI Doctor at 82.4%, Gemini Advanced and Chat Doctor at 81.2%, and Copilot at 80%. LLMs with internet access did outperform ChatGPT-4.

Conclusion

Our findings highlight the potential use of LLMs for family medicine education and CFPC exam preparation. We demonstrated greater consistency and accuracy in GPT-4's responses compared to GPT-3.5 and the superior performance of GPT-4 among the ten models. There is a need to do further studies to explore several utilities of LLMs and improve their performance in this domain.

Résumé

Contexte

Récemment, il y a eu un intérêt croissant pour l'utilisation de grands modèles de langage (LLMs) dans l'éducation médicale et la préparation aux examens médicaux, comme les examens en médecine familiale. Cependant, l'émergence continue de nouveaux LLMs, la formation spécialisée dans le domaine médical, et l'accès aux ressources en ligne apportent un potentiel d'avantages.

Dans cette thèse, nous avons réalisé deux études pour évaluer l'utilité de dix chatbots basés sur des LLMs en utilisant un échantillon de questionnaire pour l'examen de médecine familiale canadien, afin d'évaluer la précision des réponses et d'explorer quels aspects de ces LLMs pourraient être plus utiles dans ce contexte.

Méthodes

Nous avons utilisé un échantillon du questionnaire de problèmes de gestion à réponses courtes (SAMPs) du site officiel du Collège des médecins de famille du Canada (CMFC). Cet ensemble d'échantillons comprenait 19 scénarios cliniques couvrant divers domaines de la médecine familiale, chacun comprenant de 2 à 7 questions, résultant en un total de 77 questions nécessitant 165 lignes de réponses.

Dans la première étude, nous avons utilisé GPT-3.5 et GPT-4 (ChatGPT, OpenAI, San Francisco, CA) cinq fois entre le 8 et le 25 août 2023 pour répondre au questionnaire. Pour le premier tour, nous avons fourni une invite limitant la réponse à 10 mots par ligne. Cette procédure a été répétée une semaine plus tard au deuxième tour. Le troisième tour impliquait la régénération immédiate des réponses du deuxième tour. Le quatrième tour utilisait une invite différente qui ne faisait pas référence aux questions d'examen du CMFC. Enfin, le cinquième tour s'est déroulé sans invite.

Deux évaluateurs certifiés par le CMFC avec une expérience moyenne de 15 ans ont évalué les réponses en fonction du clé de réponse du CMFC. Les évaluateurs étaient aveugles aux évaluations de chacun mais ont résolu toutes les divergences par plusieurs réunions. Nous avons appliqué un modèle d'équations d'estimation généralisées (GEE) logistique ordinal pour analyser les mesures répétées sur les cinq tours.

Dans notre deuxième étude, menée en février et mars 2024, nous avons examiné dix chatbots basés sur des LLM comme suit : (1) LLM de domaine général sans recherche en ligne, y compris GPT-3.5 et GPT-4 d'Open AI, Gemini Pro 1.0 et Gemini Ultra 1.0 de Google et Personal Intelligence d'Inflection AI (utilisant Inflection 2.5) ; (2) Chatbots basés sur des LLM incluant des capacités de recherche en ligne, comme le mode Smart de YouChat (utilisant GPT-3), le mode GPT-4 de YouChat et Copilot de Microsoft (utilisant GPT-4) ; (3) LLM formés sur des connaissances médicales : c'est-à-dire AI Doctor de Docus AI (utilisant GPT-4) et Chat Doctor créé par meta-AI (utilisant LLaMA).

Comme dans la première étude, trois évaluateurs certifiés par le CMFC ont noté les réponses générées par les LLM. Nous avons utilisé le Kappa de Fleiss pour mesurer l'accord entre les évaluateurs après les réunions.

Résultats Suite de la discussion, les deux évaluateurs se sont mis d'accord sur tous les scores de l'étude initiale. Selon la clé de réponse du CMFC, GPT-3.5 a fourni des réponses exactes dans 607 des 825 lignes de réponses (73,6 %),, tandis que GPT-4 a atteint une précision de 691 cas (81%). L'analyse GEE a révélé que GPT-4 avait 2,31 fois plus de chances d'obtenir un score CFPC plus élevé que GPT-3.5 (Odds Ratio : 2,31 ; Intervalle de Confiance à 95% : 1,53 à 3,47 ; P<0,001). D'autres résultats des cinq tours de la première étude comprenaient une cohérence sur une semaine, aucun impact significatif de l'utilisation de la limite de 10 mots, et aucun changement substantiel après la suppression de l'examen CMFC de l'invite initiale.

Dans la deuxième étude, le score Kappa variait de 0,89 à 1, indiquant un consensus parfait. ChatGPT-4 a démontré le taux de précision le plus élevé parmi les dix systèmes avec 85,5%, suivi par AI Doctor avec 82,4%, Gemini Advanced et Chat Doctor avec 81,2%, et Copilot avec 80%. Les LLM avec accès à Internet ont surpassé ChatGPT-4.

Conclusion

Nos résultats mettent en évidence le potentiel des LLMs pour l'éducation en médecine familiale et la préparation à l'examen du CMFC. Nous avons démontré une plus grande cohérence et précision dans les réponses de GPT-4 par rapport à GPT-3.5, ainsi que la performance supérieure de GPT-4 parmi les dix modèles. Il est nécessaire de mener d'autres études pour explorer plusieurs utilisations des LLMs et améliorer leur performance dans ce domaine.

Acknowledgements

I would like to express my heartfelt thanks and gratitude to my wife, ***Roya***, for her invaluable support and encouragement throughout this project.

I am deeply grateful to my principal supervisor, ***Dr. Samira A. Rahimi***, for her invaluable guidance, leadership, and availability throughout this thesis project.

I am grateful for the guidance and knowledge imparted by my co-supervisor, ***Dr. Yves Bergevin***, whose support has been invaluable throughout this project. His initial acceptance to supervise me marked the beginning of my master's study journey, opening doors for me to learn and grow as a scholar. I appreciate his in-depth review of the thesis and his valuable comments on its improvement.

I extend my heartfelt thanks to my colleagues, ***Dr. Shabnam Shafee*** and ***Dr. Devin Ritter***, for their dedicated help and time throughout this project. Dr. Shabnam Shafee served as a family doctor reviewer, evaluating the AI-generated responses for both manuscripts, while Dr. Devin Ritter assumed the role of family doctor reviewer for the second manuscript, assisting in scoring and evaluating the AI-generated responses. Both colleagues also contributed as co-authors of the manuscripts.

I am thankful to ***Dr. Jason M. Harley*** for his valuable comments and insights during the drafting of the manuscripts, his critical review, and for graciously accepting my invitation to serve on this thesis committee.

My sincere gratitude goes to ***Dr. Ramana-Kumar Agnihotram*** and ***Dr. Jose Correa*** for their invaluable guidance in selecting the statistical analysis methods. Dr. Agnihotram reviewed both manuscripts, and Dr. Correa guided me on the analysis of the second manuscript.

I would like to thank ***Dr. Jackie Cheung*** for his participation as a co-author on the first manuscript, his critical review, and his insightful comments.

This thesis' grammar and vocabulary were refined and corrected using Grammarly Premium. We also used ChatGPT to edit the translation of the abstract to French.

The first manuscript has been published in the "Journal of Family Medicine and Community Health" with the following citation: "Mousavi M, Shafiee S, Harley JM, Cheung J, Rahimi SA. Performance of ChatGPT in Certification Examination of College of Family Physicians of Canada, doi: 10.1136/fmch-2023-002626. PMID: 38806403; PMCID: PMC11138270. Fam Med Community Health. 2024 May 28(12 (Suppl 1)):e002626".

The second manuscript has been prepared in the appropriate format for submission for peer review and publication.

Contribution to original knowledge

Mehdi Mousavi, the student, has authored all chapters of this thesis. He conceived the idea, designed the research study, conducted the literature review, operated the AI systems to respond to the questionnaire, reviewed the responses, and wrote the original drafts for both manuscripts.

Dr. Shabnam Shafiee participated in Chapters 3 and 4, manuscripts one and two: Formal Analysis, Writing-Revision and Edits;

Dr. Devin Ritter participated in Chapters 4, manuscripts two: Formal Analysis, Writing-Revision and Edits;

Dr Jason M. Harley participated in Chapters 3 and 4, manuscripts one and two: Conceptualization, Writing-Revision and Edits;

Dr Jackie Cheung participated in Chapters 3, Manuscripts One: Conceptualization, Writing-Revision and Edits;

Dr. Yves Bergevin participated in conceptualizing the second study, reviewing manuscript two, and writing the overall thesis.

Dr. Samira A. Rahimi participated in Chapters 3 and 4 and the first and second manuscripts: conceptualization, Methodology, Supervision, Project administration, Writing Revision, and Edits. She also carefully reviewed, edited, and evaluated all parts of the thesis.

Contribution of Authors

Mehdi Mousavi, the student, has authored all chapters of this thesis;

Dr. Shabnam Shafiee participated in Chapters 3 and 4;

Dr. Devin Ritter participated in Chapter 4;

Dr Jason M. Harley participated in Chapters 3 and 4;

Dr Jackie Cheung participated in Chapter 3;

Dr. Yves Bergevin participated in the review of all chapters;

Dr. Samira A. Rahimi participated in all chapters

CHAPTER 1:

INTRODUCTION

ARTIFICIAL INTELIGENCE (AI)

The idea of ***Artificial Intelligence (AI)*** was first introduced by John McCarthy in 1956 during a Summer Research Project at Dartmouth College (1). The term " ***Artificial Intelligence (AI),***" refers to machines designed to do things humans can do, including perceiving, reasoning, learning, and communicating (2). ***Perceiving*** means understanding and interpreting the world around them, like recognizing faces or objects. ***Reasoning*** is making decisions based on information, like solving problems or planning. ***Learning*** may be done by **i**mproving their performance over time by learning from data and experiences. ***Communicating*** refers to interacting with humans through language, like having a conversation or understanding written text (2).

Most AI technologies use mathematical algorithms to explore and analyze patterns in available data, including numbers, text, audio, video, or images (3). AI and machine learning (ML) are expanding rapidly. AI helps medical professionals by managing and analyzing large quantities of data in ways that humans cannot (4). With the rise of digital healthcare and AI, the chance of human error in diagnosing diseases has significantly decreased (5).

Some potential tasks that can be done in the medical domain to improve healthcare quality are as follows:

- **Exploring and describing data**: AI can analyze various data sources, identify structures and trends, and offer detailed information about the data. For example, AI systems have been developed to improve the interpretation of pulmonary function tests by describing data provided by the test device (6).

- **Analyzing and interpreting texts:** AI could help draft discharge summaries and summarize the information from multiple sources.

- **Administrative Automation**: AI can streamline hospital and clinic administration tasks, such as scheduling, billing, and compliance tracking. It can also send reminder emails or messages to patients for their appointments, freeing time for healthcare professionals to focus on patient care.

- **Taking notes and documentation:** AI can help doctors by recording important points and diagnoses during patient consultations (2). Several applications and software are available for doctors to dictate notes for their patients (2).

- **Predicting specific outcomes:** AI can use available patient information, such as demographics, past medical history, family history, and previous lab results, to predict specific outcomes (3). For instance, AI can help predict the risk of cardiovascular diseases or the progression of kidney diseases (6).

- **Assisting in the interpretation of medical imaging**: One successful example of the use of AI in diagnostic medicine is the CheXNet system, which identifies pneumonia from images. In a study, CheXNet was more accurate than radiologists (6). Another example is a tool to detect diabetic retinopathy from eye images (3)

- **Assisting in the diagnosis of medical conditions** (4): AI can help diagnose diseases by analyzing medical imaging, genetic data, and patient symptoms more quickly and accurately than traditional methods. A 2020 study by Chen et al. examined how artificial neural networks (ANNs) are used to diagnose liver disease and found them to be highly accurate (5). Another research by Jo et al. in 2019 found that deep learning AI techniques achieved a 96% accuracy rate in identifying Alzheimer's disease (5).

- **Supporting decision-making processes**: AI can help with short-term and long-term decision-making using data collected over time (2).

- **Suggesting treatment options** (4): AI can use rules from clinical guidelines and other evidence to recommend diagnoses or treatments (2). Furthermore, by analyzing past treatment outcomes and individual patient data, AI can offer treatment plans that better suit individual patient needs and improve outcomes. For example, MYCIN, developed in the 1970s, was one of the earliest uses of AI to help with treatment choices for bacterial infections (2). AI can also provide intelligent solutions for managing seizures and assessing neurological conditions.

- **Helping in patient follow-up and monitoring**: AI-enabled devices can continuously monitor patients, especially those with chronic conditions, and alert healthcare providers to changes in their condition that may require intervention. For example, AI can assist in continuous glucose monitoring for diabetes patients (6).

- **Drug Development**: Using computational methods, AI can speed up the discovery of potential drug candidates and help predict their properties compared to traditional lab testing (7).

As AI and its implications in medicine gain traction, healthcare practitioners must understand what they are, their potential applications in medicine, and associated pitfalls (8). Training primary care clinicians to use AI may enhance patient outcomes, improve overall health equity, reduce costs, and support clinician well-being. Without proper training, AI could lead to dissatisfaction among healthcare providers, increased costs, more fragmentation in care, and higher burnout rates, potentially compromising patient safety. Due to such important potentials, the American Board of Artificial Intelligence in Medicine (ABAIM) has been established to take steps in integrating AI into healthcare, such as encouraging family medicine residents to adopt these technologies (8).

Despite several benefits, there are strong reasons to be cautious about using AI in healthcare. Concerns include patient privacy and consent, disruption of workflows, and potential harm from issues not directly related to patient care (4). Challenges in applying AI in community-based primary healthcare include inconsistent patient data and abbreviations used in data entry (4).

LARGE LANGUAGE MODELS (LLMs)

Humans use language to express ideas and emotions and facilitate communication with others. In contrast, machines lack the ability to comprehend, and to be able to read, write, and perform conversation will need powerful machine learning algorithms for ***natural language processing (NLP)***. Language models are the heart of NLP that predict or generate the likelihood of words, phrases, or sentences based on context (7). These language models have evolved as follows:

- ***Statistical Language Models (SLMs)***: Use basic probability distributions to model word sequences (7).
- ***Neural Language Models (NLMs)***: Use neural networks to understand complex language

patterns (7). Computers use word vectors to comprehend words through binary codes. Word2 Vec is a tool that computes word vectors utilizing neural networks similar to those of the human brain. NLMs effectively handle longer sequences and address the limitations associated with previous SLM models (7).

● ***Pre-trained Language Models (PLMs)***: These systems are developed by pre-training the system using a large volume of unlabeled text. Then, the models undergo a second round of training on a smaller task-specific dataset to improve their performance on the desired task. This second training is called "fine-tuning." NLPs use large datasets and self-learning to capture general language knowledge. GPT-2 and BERT are examples of PLMs (7).

● ***Large Language Models (LLMs)*** are trained on massive text datasets and have tens of billions (or more) of parameters. They are built on PLMs with more data, computing power, and advanced algorithms, making them more expressive and adaptable (7). Similar to PLMs, LLMs undergo pre-training. However, the pre-training for LLMs is performed on vast text databases. Then, fine-tuning is done using smaller, task-specific datasets for specific tasks (9). The words are taken from articles, books, and online content using deep neural networks (10).

The training for LLMs might happen in several steps and can include different amounts of human help. LLMs use what they have learned to carry out tasks that involve understanding and generating language (10).

There have been three steps in developing LLMs: transformers, large foundational models, and Multi-modal Language Models (MLLMs) (9). Transformers are made up of layers of encoder-decoder. They are trained on large text datasets using unsupervised or semi-supervised learning and gradient-based optimization. Transformers can adapt to various tasks with minimal fine-tuning (9). Large foundational models like GPT-3 and Stable Diffusion represent a significant advancement in machine learning and generative AI. These models are trained on extensive, diverse, and unlabeled datasets, enabling them to handle a wide range of tasks, such as language comprehension, text and image generation, and natural language dialogue (9).

A Multi-Modal Large Language Model (MLLM) can process and generate content across multiple formats, including text, images, audio, and video, unlike traditional language models that only handle text. This multi-modal integration helps bridge the gap between human communication and machine understanding and can improve image and speech recognition, content generation, and interactive AI applications (9).

LLMs have developed in different sizes (11). Numerous attempts have been made to enhance LLMs by increasing the number of parameters under the assumption that larger models will yield better performance. However, recent information suggests that within a fixed computing budget, optimal performance may not be achieved by the larger models but rather by smaller models trained on a more extensive dataset(12) and fine-tuning (11).

Definition of Some of the Expressions Used in the Field of LLMs

We will now define key terms commonly used in the field of LLMs to facilitate a clearer understanding of concepts discussed in studies and related research.

- **Pre-training**

Pre-training involves the initial step of training for LLMs using a substantial amount of unlabeled text to understand fundamental language structures such as vocabulary, syntax, semantics, and logic (7). It is a foundational stage in the development of Large Language Models (LLMs), where the model learns general language patterns, rules, and relationships from vast amounts of text data (7).

- **Fine-tuning**

Fine-tuning is performed on more specific tasks and usually has a smaller volume. The goal of fine-tuning is to improve the performance of LLM (7). During this step, the model is trained with a focus on specific examples to specialize it while preserving the general language understanding gained in pre-training. ChatGPT, for instance, is an example of fine-tuning. Open AI has performed fine-tuning to improve on conversational datasets to make them better at dialogue-specific tasks such as answering questions, clarifying ambiguities, and maintaining conversational context (7).

- **Instruction-tuning**

Instruction-tuning involves further training these models on a dataset consisting of pairs of instructions and their corresponding outputs. It is a technique used to enhance the performance and controllability of LLMs (13). ClinicalGPT, for example, was created using an instruction-tuning method of the BLOOM 7B LLM to answer medical knowledge questions (42).

● **Zero-shot learning**

Zero-shot learning (ZSL) is a machine learning approach where an AI model can identify and classify objects or concepts without having encountered any instances of those categories during its training phase (14).

● **Few-shot learning**

Few-shot learning means training the models using only a small or limited number of examples to make accurate results. This approach is different from traditional machine learning, which often requires vast amounts of data to achieve high accuracy (15).

● **Chain of thought (COT)**

Chain of thought means enhancing each few-shot example in a prompt by providing a step-by-step explanation that leads to the final answer (38).

● **Self-consistency (SC)**

Self-consistency (SC) is a technique used to enhance performance on multiple-choice tests by generating multiple explanations and answers from the model (38). The preferred answer is determined by the one receiving the majority (or plurality) of votes (38).

● **Open-Source Large Language Models**

Open-source LLMs are built using open-source software and resources, allowing the code and pre-trained models to be freely accessible for anyone to use, modify, and distribute (16, 17). Open-source data is available to the public (16, 18), exclusively uses publicly accessible data, and does not incorporate books or other data not open to the public (18).

Open-source are developed to address the issue of making models publicly available for research purposes at a lower cost (18). The safe public release of LLMs, as done with open-source LLMs, could potentially provide a net benefit to society (12).

● **Closed-Source Large Language Models**

Closed-source LLMs only grant access to their organization or those they authorize, with an emphasis on protecting intellectual property and monetizing their technology (17). Closed-source LLMs like GPT are extensively fine-tuned to adhere to human preferences, markedly improving their functionality and safety (12). Developing these proprietary closed models demands significant investments in computational resources and human annotation (12). Furthermore, the non-transparent and difficult-to-replicate nature of their development limits the community's progress in AI alignment research (12).

● **Application Programming Interface (API)**

Closed-source LLMs are available through the "Application Programming Interface (API)" (19). API is a set of rules and protocols that allows different software applications to communicate with each other. Using API, data is transmitted to the parent company for processing (20).

In the following part of this chapter, we provide examples of different types of LLMs and information about their development. In the second chapter (Review of articles), we provide examples of studies using these LLMs to answer questions from several exams.

General Domain Large Language Models

General domain LLMs such as GPT and Google Gemini are used in various ways, including analyzing data, converting text into computer code, conducting sentiment analysis, text translation across different languages, editing and correcting written material (19). They can be used in creating content for social media, blogs, and other marketing materials, serve as customer service chatbots tailored to specific business documentation and data (19) and many other daily tasks. Unlike specialized models focusing on specific tasks or domains, general-domain LLMs are versatile and can handle various language-related tasks. After the launch of ChatGPT, one of the most well-known LLMs, several companies introduced similar models (21). The Field of LLMs is evolving rapidly. At the time of drafting this thesis (mid-2024), the following general domain LLMs were introduced:

Generative Pre-trained Transformer (GPT), GPT-2, GPT-3, GPT-4 by OpenAI, BERT by Google, ERNIE by Academia, Conditional Transformer Language Model (CTRL) by Salesforce, BART by Meta, Turing-Natural Language Generation (Turning-NLG) by Microsoft, Megatron by Nvidia, Vision Transformer (ViT) by Google, DALL-E by OpenAI, Switch by Google, Swin transformer by Microsoft, Wu Dao 2.0 by Academia, Jurassic-1 by AI21, Megatron-Turing-, Natural Language Generation (MT-NLG) by Nvidia, Anthropic-LM by Anthropic, Generalist Language Model (Glam) by Google, Guided Language to Image Fiddusion for Generation and Editing (GLIDE) by OpenAI, Gopher by Deepmind, Cassually Masked (CM3) by Meta (10), Language Model for Dialogue Applications (LaMDa) by Google (10, 22), Chinchilla by DeepMind, GopherCite by DeepMind, DALL-E-2 by OpenAI, Flamingo by DeepMind (10), Pathways Language Model (PaLM) (23), PaLM 2 (24) by Google, Gato by DeepMind, OPT by Meta, Imagen by Google, Minerva by Google, BigScience Large Open-Science Open-Access Multilingual Language Model (BLOOM) by Academia, GLM by Academia, Sparrow by DeepMind, Luminous by Aleph Alpha, Large Language Model Meta AI (18) (LLaMA), LLaMA 2 (12), LLaMA 3 (19, 25) by Meta (10), Gemini family by Google, including Gemini Pro 1.0, Gemini Ultra 1.0 (19), and Gemini Pro 1.5 (26), Gemma by Google (10), Inflection 2.5 by Inflection AI (27), Vicuna by LMSYS Org (10), Claude family including Claude 3 by Anthropic (10), Stable Beluga and StableLM 2 by Stability AI (10), Coral by Cohere (19), Falcon by Technology Innovation Institute (19), DBRX by Databricks and Mosaic (19), Mixtral 8x7B and 8x22B by Mistral AI (19), XGen-7B by Salesforce (19), Grok by xAI (19).

Examples of Closed- Source General Domain LLMs

Closed-source LLMs are extensively fine-tuned to match human preferences. This process often involves considerable expenses in terms of computing resources and human annotation (12). Major corporations like Google and OpenAI have invested significantly in closed-source LLMs, with Google developing Gemini and OpenAI creating GPT-4 (16).

● **Generative Pre-trained Transformer (GPT)**

GPT (https://chatgpt.com/) is the most well-known published LLM (11) and is designed to process natural language data more efficient (9).GPT uses a special kind of neural network called a transformer, which is made for handling and creating natural language. This architecture understands how different parts of the text relate, which is crucial for tasks involving language. Transformer blocks use self-attention to focus on the important parts of the text (22). Giving different importance to different words during prediction makes Transformers especially good at understanding long-term connections in language (9). Also, “feedforward networks” help it understand the connections between different pieces of information in the text (22). The effectiveness of employing a transformer is clear when it deals with lengthy and complex pieces of text, achieving better results by focusing on specific parts that matter most (22). This makes it extremely useful for tasks like creating and understanding language (22).

GPT learns through two main stages. First, it is pre-trained by a large amount of unlabeled text without supervision. This stage makes it possible to learn general linguistic capabilities (22). Exposure to large amounts of text makes it possible for the transformer to learn patterns and connections within language (22). In the next stage, supervised fine-tuning is done to learn on smaller amounts of labelled data to optimize GPT performance on certain tasks, like answering questions (22).

In summary, ChatGPT uses transformers, learns from lots of text, and fine-tunes its skills on specific tasks to create responses that make sense in different contexts, all based on deep learning methods that mimic how humans use language.

GPT-1 was released in 2018 with 117 million parameters (22). BooksCorpus dataset, a collection of 11,308 novels containing around 10^9 words, was used for training.

GPT-2 was launched in February 2019 with 1.5 billion parameters (22) and was ten times larger than GPT-1 using WebText, the dataset from over 8 million documents.

GPT-3 was introduced in June 2020 with 175 billion parameters (22) and was 100 times larger than GPT-2 (11). GPT-3's sophisticated text-generation abilities enabled its use in writing emails, drafting articles, and even composing poetry (22).

GPT-3.5 was launched by Open AI in November 2022) (22). It was developed by refining and training GPT-3 with prompts and answers created by human experts, helping it learn how to respond to specific questions accurately. The following phase involved 'reinforcement learning from human feedback' (RLHF) to enhance its capabilities (11).

Launched on March 14, 2023, ***GPT-4*** is reported to outperform GPT-3.5 in natural language processing tasks (11) and is more accurate and efficient, with a reduced tendency to produce *hallucinations* (28, 29). Hallucinations are misleading responses provided by the chatbot that are so convincing that people often do not question their accuracy (28, 29), posing a challenge in the use of LLMs.

● **Pathways Language Model (PaLM)**

PaLM, developed by Google, is a densely activated Transformer language model trained on 540 billion parameters. PaLM 540B delivers exceptional results, close to the average human performance on the benchmark (23). Flan-PaLM was developed by finetuning PaLM on a collection of datasets phrased as instruction (instruction-tunning) (13). PaLM 2 (by Google) was a Transformer-based model that showed superior multilingual and reasoning capabilities over PaLM and was more compute-efficient (24).

● **Language Model for Dialogue Applications (LaMDA**): LaMDA is a Transformer neural network model developed by Google that has been explicitly fine-tuned for dialogue-based applications and open-ended conversations such as question-answering (22). It ranged from 2 billion to 137 billion parameters and was pre-trained on a massive 1.56 trillion-word dataset from public conversations and web documents (22). LaMDA was initially used to develop ***Google Bard*** before switching to PaLM 2 (10).

LaMDA and PaLM 2 were competitors to GPT-4 in both general intelligence and specialized knowledge areas (10). However, they have been replaced with a more modernized LLM named Gemini.

● **Google Gemini Group**

Gemini (https://deepmind.google/technologies/gemini) (Google, New York, NY) represents a suite of AI models introduced by Google, consisting of Gemini Nano, Gemini Pro, Gemini Ultra (19) and lately, Google Pro 1.5 (26). These models are tailored for various devices like smartphones. Besides generating text similar to other LLMs, the Gemini models are equipped to process images, audio, video, code, and various other data types natively (19).

Gemini Pro 1.0 (Google, New York, NY) model, introduced in February 2024, is an LLM developed by Google AI and is different from GPT-3 or GPT-4 (30).

Google introduced the ***Gemini Ultra 1.0 model*** at the same time as Gemini in February 2024. Google claimed this was their most advanced AI model by that time (30).

Gemini Pro 1.5 has recently been released to Gemini Advanced subscribers, featuring significant enhancements. These improvements include better image comprehension, an extended context window, and new data analysis capabilities. Additionally, it integrates more seamlessly with other Google apps and offers more customization options. Users may also be able to upload files into Gemini Advanced, leveraging the extended context window to quickly derive answers and insights from complex documents, analyze data, and create charts (26).

● **Claude**

Anthropic introduced the latest generation of Claude, ***Claude 3,*** available from https://www:anthropic.com/claude on Mar 4, 2024 (31). Claude exists in three models—Haiku, Sonnet, and Opus in ascending order of capability. Opus, the most intelligent model, performs near human levels and outperforms its counterparts on most of the common evaluation benchmarks for AI systems, like undergraduate-level expert knowledge, graduate-level expert reasoning, and basic mathematics. The Claude 3 can process various visual formats, including photos, charts, graphs and technical diagrams. Improved accuracy, more reliability, less hallucination, long context and near-perfect recall are some advances in the Claude 3 compared to the previous version (31). Claude 3 is engineered to be useful, truthful, benign, and, importantly, secure for enterprise customers. Consequently, businesses such as Slack, Notion, and Zoom have formed partnerships with Anthropic and use Claude 3 (19).

● **Coral**

Coral (developed by Cohere), similar to Claude 3, is made for enterprise users to respond to specific queries from employees and customers (19).

● **Inflection-2.5**

Inflection-2.5, which is distinct from GPTs, was developed by Inflection AI, Palo Alto, CA (https://pi.ai/talk). It is used in the Personal Intelligence or PI chatbot (27) that learns the interests and preferences of the user over time to provide personalized responses (32).

Examples of Open-Source General Domain LLMs

Open-source LLMs use publicly accessible data that are freely available to the public (18). Large Language Model Meta AI (LLaMa) Group are the most famous example of open-source LLMs (16).

In this section, we introduce some of these open-source LLMs.

● **Large Language Model Meta AI (LLaMA) Group**

LLaMA models are leading the way among open-source LLMs trained primarily on general domain datasets (33).

– LLaMA (LLaMA 1)

LLaMa (Large Language Model Meta AI) is a collection of LLMs from 7B to 65B parameters released to the research community by Meta AI (18).

While LLaMA-13B is ten times smaller, it performed better than GPT-3 (175B parameters) in the majority of benchmarks. Moreover, LLaMA-65B competes closely with top-tier models like Chinchilla-70B and PaLM-540B, which are not available to the public (18).

– LLaMA 2

Llama2 and LLaMA 2-Chat (optimized for dialogue use cases) were released to the general public in July 2023 for research and commercial use.

LLaMA 2, an updated version of Llama 1, underwent training on a new variety of publicly available data. The pretraining dataset was expanded by 40%, the model's context length was doubled and included the integration of grouped-query attention. LLaMA 2 was released in versions featuring 7 billion, 13 billion, and 70 billion parameters (12).

LLaMA 2 surpasses LLaMA 1 and other open-source chat models such as Falcon, Vicuna, and MTP in most benchmarks based on human assessments of helpfulness and safety. The largest Llama 2-Chat model, with 70 billion parameters, is comparable to ChatGPT and exceeds the performance of the PaLM-bison chat model (two close general domain LLMs) (12).

– LLaMA 3

LLaMA 3 is Meta's advanced open-source LLM, which has been pre-trained and fine-tuned for specific instructions with 8 billion and 70 billion parameters. It exhibits enhanced reasoning and code generation capabilities, significantly advancing over its predecessor, LLaMA 2. LLaMA 3 has been pre-trained on over 15 trillion tokens, all sourced from publicly available data. This training dataset is seven times larger than the one used for LLaMA 2 and contains four times more code (25).

LLaMA 3 enjoys the advantage of its extensive application throughout Meta's platforms, such as Facebook and Instagram, powering a variety of AI functionalities. This has established it as one of the most potent and frequently utilized open large language models currently available. Moreover, LLaMA 3's source code is openly available on GitHub, enabling both researchers and commercial entities to use it as a base for developing their own models (19).

● **Mistral 7B**

Mistral 7B (7–billion-parameter) utilizes a transformer architecture but incorporates several modifications compared to LLaMA, including 1) Sliding Window Attention (SWA), designed to more efficiently manage longer sequences with reduced computational costs; 2) Rolling Buffer Cache; 3) Pre-fill and Chunking strategies (34).

Mistral 7B, with its 7-billion-parameter design, exceeded the top-performing open 13B model (LLaMA 2) in all tested benchmarks and surpassed the best available 34B model (LLaMA 1) in tasks involving reasoning, mathematics, and code generation. This showcases that a well-crafted language model can achieve superior performance with efficient inference capabilities (34).

● **Other open-source general domain LLMs**

Following are some examples of other Open-source LLMs:

BLOOM (12) by Academia (176 billion) (10); ***OPT*** (12) by Meta (174billion) (10); ***Falcon*** *(12, 19)* by Technology Innovation Institute (1.3 billion, 7.5 billion, 40 billion, and 180 billion parameters) (19); ***MPT*** (MosaicML Pretrained Transformer 7 B and 30 B parameters) (12); ***Vicuna*** (12, 19), built off Meta's LLaMA by LMSYS (7 billion, 13 billion, and 33 billion parameters); ***DBRX*** by Databricks and Mosaic (132 billion parameters); ***Gemma****,* by Google (2 billion and 7 billion parameters); ***Stable Beluga*** and ***StableLM 2*** developed by Stability AI (1.6 billion, 7 billion, 12 billion, 13 billion, and 70 billion parameters); ***Mixtral 8x7B and 8x22B*** by Mistral (45 billion and 141 billion parameters); ***XGen-7B*** by Salesforce (7 billion parameters); ***Grok*** by xAI trained on data from X (formerly Twitter) (19).

Large Language Models Trained in Medical Domain Knowledge

General-domain LLMs have demonstrated promising outcomes in the field of medicine. However, there are concerns that only a small fraction of biomedical texts were used during their development (35), and their specialized knowledge and contextual understanding may not be sufficient for their accurate and reliable application in medical settings (32). Furthermore, there are concerns about the stability and transparency of proprietary LLMs like GPT-4 (36). Most closed-source LLMs, such as GPT-4, are accessible through an API, which requires transmitting data to the parent company for processing (20). This arrangement can raise concerns about privacy and the potential for breaches of confidentiality, particularly regarding sensitive medical information.

Numerous efforts have been made to tailor LLMs to meet specific medical needs (28) by pre-training (33, 37-39), instruction fine-tuning (40, 41) of LLMs. Many models have used biomedical corpus to train their LLM. Pubmed is one of the most popular datasets for training used to train BioGPT (355B) (28) founded on GPT-2 (39), PubMedGPT (2.7B) (40), Galactica (120B) (40), PubMedBERT (39), Sci- Five (39), and BioMistral (39).

PMC-LLaMA (32) and *MediTron* (32) are other examples of domain-specific LLMs trained on biomedical literature. *GatorTronGPT* (35)*, Clinical-Camel* (36), and ClinicalGPT (42) are other examples of LLMs trained on medical data and evaluated clinical tasks (32). Me LLaMA (32), on the other hand, is developed using a combination of biomedical and clinical data. The following discussion will introduce some samples of such medical-specific LLMs. In our second study, we used ChatDoctor and AI-Dr as a sample of LLMs trained in Medical Knowledge.

In the second Chapter of this thesis, the review articles, we will provide a more detailed discussion about some of the LLMs trained on the medical domain.

CHATBOTS

The term "***chatbot***" originates from merging "chat" and "robot." Initially, it referred a text-based dialogue system mimicking human conversations. However, the new modern chatbots not only respond to conversations but also learn and adapt over time, making it possible to conduct more personalized and engaging interactions (22). Chatbots can simulate both text and voice conversation with humans (43).

The first chatbot, Eliza, was introduced in 1966 by Joseph Weizenbaum (22). Since then, chatbots have evolved into various forms, including messenger applications like SmarterChild (2001), and virtual personal assistants such as IBM Watson (2006), Apple's Siri (2011), Microsoft's Cortana (2014), Amazon Alexa (2014), Google Assistant(22, 43), ChatGPT (2018-2022), Google Bard (2023) (22)

Pioneering chatbots such as Eliza employed several rules and templates to generate appropriate responses. Initially, they scan the input for predefined keywords, structures, or patterns. Upon identifying a match with predefined criteria, they create a response using the corresponding output template. However, their abilities are limited by the templates made by developers. Furthermore, the responses can be predictable and repetitive. Additionally, there's no memory of previous conversations, which can lead to looping discussions (22).

Chatbots simulate their conversations over the Internet through Natural Language Processing (NLP) (43). NLP is a field of AI that uses machine learning and deep learning to enable computers to recognize, understand, and generate text and speech similar to human language (14). An example of NLP is digital smartphone assistance and voice commands on operating systems (14). LLMs are also a subset of NLP. They are specifically designed to generate and understand texts. LLMs require massive amounts of training data to learn effectively. However, the size and quality of NLP training data can vary.

The most rapid expansion of the chatbot market was in early 2020 after the introduction of ChatGPT and Bard (22). Many new chatbots are now available, with numerous applications that can be installed on Apple or Android smartphones. Most of these are based on large language models (LLMs).

- **ChatGPT**

ChatGPT is the most well-known LLM-based chatbot (11) by OpenAI (San Francisco, CA, USA). ChatGPT has an impressive capacity to adapt within single conversations, significantly improving its performance when given examples of the task at hand (11). It can answer questions, draft documents, translate languages, summarize and evaluate research and medical literature (44). These features have made ChatGPT the most attractive LLM-based chatbot in medical research at present (11).

Currently, four main models available are GPT-3.5-turbo, GPT-4, and GPT-4 Turbo and very recently, GPT-4o, which is a multimodal version (19). OpenAI 's ChatGPT is free to use but requires a subscription. Also, the ChatGPT application is available for smartphones (45). ***ChatGPT Plus*** enables access to ***GPT-4*** with a subscription of USD $20/month (46), or through other applications like ***YouChat GPT-4 mode*** (47), and Copilot by Microsoft (19, 48, 49) . Some of the other companies that use GPTs are Duolingo, Stripe, Descript, Dropbox, and Zapier (19).

ChatGPT-4o ("o" for "omni") is one the most recent products of Open AI at the time of drafting this document. It enhances the natural interaction between humans and computers. It processes inputs in various forms, including text, audio, images, and videos, and produces outputs in text, audio, and image formats. The model is also optimized for rapid responses to audio inputs. GPT-4o matches the performance of GPT-4 Turbo in text-based, reasoning, and coding tasks. Access to GPT-4o is extended to both free and Plus users, with Plus users enjoying a limit of up to five times more messages (50).

● GPT-based Chatbots with access to online resources

– ***Bing Chat*** (Microsoft, Redmond, Washington, USA), first launched to the public in May 2023, was integrated into the Microsoft Edge browser. It was rebranded as Microsoft ***Copilot*** in November 2023. This Chatbot uses Open AI's GPT-4 and features the capability for visual search as well as text and voice within Chat. Copilot may also access GPT-4 Turbo (a more advanced product of Open AI) during non-peak times (48).

– ***Copilot*** (Microsoft, Redmond, Washington, USA) provides free access to GPT-4 through https://copilot.microsoft.com**.** Copilot is able to search the internet among open-source documents. Another potential advantage of this Chatbot compared to ChatGPT Plus is that it provides web-based references alongside its responses, providing users with easy access to additional online information sources. This specific feature makes the answers more reliable.

– ***Chatsonic (***by Writesonic) (51, 52), runs on GPT-4 and aims to overcome ChatGPT's shortcomings by offering real-time data, image and voice searches, and a range of content creation features (51, 52). It's linked with Google Knowledge Graph to deliver highly relevant and trending content on various subjects. As of our latest review in May 2024, Chatsonic is available through Writesonic at https://app.writesonic.com. It offers a free trial with up to 50 generations per day, with other plans starting at $12/month (52).

– YouChat, Personalized AI Assistant or simply, "***YouChat***" (49), powered by You.com, is one of ChatGPT's early competitors. This AI assistant enables users to access different modes:

YouChat Smart Mode: This setting is perfect for rapid tasks, responding to inquiries, creating content, and accessing information instantly (47). We accessed the YouChat version (1.0.9), updated on September 27, 2023, for our second article on February 4th, 2024, which was using GPT-3 (49). YouChat offers references to answers in some instances, albeit not consistently (49).

YouChat GPT-4 mode: YouChat enables access to the complete feature set of GPT-4 along with real-time web results (47). The version (1.0.9), at the time of our access for our second article on February 4th, 2024, had a cost of $9.99 per month and similar to the smart mode, it provided references in some of the responses (49).

YouChat Genius Mode, another feature unlocked by YouPro, allows users to chat with PDFs, text, and image files (47).

YouChat Research Mode helps researchers through analysis, comparisons, and topic exploration (47).

YouChat Create Mode is able to change any idea or question into AI art (47).

The newer version of YouChat enables using several LLMs, including GPT-4, GPT-4 Turbo, Claude Instant, Claude 2, Claude 3 Opus, Claude 3 Sonnet, Gemini Pro, DBRX-Instruct and Zephyr (Uncensored) (53). In this version, the smart mode is unlimited and free but typically uses GPT-4 by default and is designed to provide responses that include live web access, as well as citations and sources (53). Every user may also have limited access to premium modes, including the Custom Model Selector, offering a glimpse into You.com's capabilities as an AI Assistant. Unlimited access to all AI modes and models, including the Custom Model Selector, along with enhanced personalization features, could be accessed at $20 on a month-to-month basis (53).

- **Gemini family and Bard**

Gemini, which is an updated version of Bard, operates on the model across more than 40 languages and is used in Google apps, including Docs, Gmail and Google's chatbot (19). Gemini is built to understand and produce content from text, images, audio, and video inputs (30, 54). Gemini Pro 1.0 is a free version. The nano version is available in 1.8 billion and 3.25 billion parameter versions (19).

Gemini Ultra 1.0 can be accessed through Gemini Advanced, which excels in performing intricate tasks such as coding, programming, logical reasoning, understanding complex instructions, participating in creative projects, generating detailed guides, crafting quiz questions, and engaging in detailed discussions. In February 2024, it was accessible in English through the new Google One AI Premium Plan, priced at $19.99 monthly, and included an initial two-month free trial (30).

● **Personal Intelligence or PI**

Personal Intelligence or PI (27), developed by Inflection AI, Palo Alto, CA (https://pi.ai/talk), is a new chatbot that uses its unique LLM called Inflection-2.5, distinct from GPTs. PI AI acts as a personal AI companion by learning the interests and preferences of the user over time, making it possible to provide personalized responses (32). PI AI sets itself apart from traditional chatbots by emphasizing emotional intelligence, empathy, and personalized responses, ensuring each conversation feels unique and authentic (52). It can use web-based information and integrates with iOS, Facebook Messenger, WhatsApp, Instagram Direct, and the user's phone's SMS app. Compared to other chatbots, PI provides more human-like responses with more emotions and empathy and may offer companionship over time. PI AI offers several AI assistants that will read responses realistically (32). Currently, PI AI is free and features six speakers with authentic tones who read the questions and answers naturally.

● **AI Doctor by Docus AI (GPT 4-based)**

Docus AI, (Docus AI, Wilmington, DE, United States), available from https://docus.ai/ai-health-assistant has developed a health assistant chatbot called ***AI Doctor***. It employs GPT-4, as well as OpenAI Text Embeddings and Vector DBs. AI Doctor LLM has been trained with lots of medical data to become a helpful assistant for health-related questions. It can understand complex medical information and give useful advice to users. Also, the AI platform follows the Health Insurance Portability and Accountability Act (HIPAA) and The General Data Protection Regulation (GDPR) rules to protect user data, ensuring privacy and security throughout the process (55, 56). The HIPAA or the Kennedy–Kassebaum Act sets guidelines to protect personally identifiable healthcare information from fraud and theft while also addressing healthcare insurance coverage limitations and allowing patients to control the sharing of their health information (57). The GDPR is a comprehensive data privacy law by the European Union that protects individuals' personal data and privacy (58).

AI Dr is a sample of chatbots developed to help patients have a reliable source of information about their knowledge gap and their needs in the field of medicine (55, 56). The chatbot can also refer the user to ask questions to real doctors worldwide, paying the required consultation fee. Docus AI provides an AI doctor for free to answer three personal messages per week in different languages, then $14.99/month. It provides one AI-generated personal health report and can save chat conversations. The free version serves the same models but can give access to 500 messages per month and generate three health reports/month.

- **Other Chatbots**

BlenderBot 3 employs OPT LLM (by Meta) and also benefits from internet access to improve accuracy (59). ***Sparrow*** (DeepMind), built on Chinchilla LLM, ***Jasperchat***, ***DialoGPT***, ***Replika*** (59), Personal intelligence ***PI AI*** (by inflection AI) (27) Google Bard (Later replaced with Gemini) (30) and ***Gemini*** (from Google) (54) are other examples of general-purpose Chatbots.

APPLICATIONS OF LARGE LANGUAGE MODELS IN MEDICINE

LLMs have emerged as leaders in medical AI, promising to enhance the productivity and efficacy of clinical, educational, and research activities. (7, 10, 11). However, to address technological shortcomings, they must undergo thorough validation and continued refinement (10, 11).

Large Language Models in Everyday Medical Clinical Practice

New studies incorporating LLMs as tools for clinical decision support have shown their potential to impact outcomes, enhance productivity, and improve patient satisfaction (21).

LLMs may be employed in various ways to help with daily medical clinic practice. For example, LLMs may be useful in taking clinical notes (21, 59) and clinical summaries. (21, 59, 60). Employing LLMs for such documentation tasks can reduce time spent and decrease the likelihood of human errors (59). LLMs can also be employed to automate responses to patient inquiries about appointment scheduling (59). Additionally, they can enhance health literacy by supplying the public with clear and easily understandable health information and providing answers to their concerns (21, 61).

Large Language Models in Medical Research

LLMs can enhance various stages of medical research in the following ways. LLMs can assist in generating research questions and identifying topics for research and systematic reviews (62). They can be employed to perform comprehensive literature reviews (62), create bibliographies, outlines (53), extract abstracts, and summarize texts and articles (62). This capability allows researchers to efficiently handle the extensive amount of information available online (53). Examples of LLMs helpful in analyzing clinical text data include ClinicalBERT, GPT-3.5, and GatorTron (10).

LLMs are frequently utilized to assist researchers with academic or scientific writing, helping to generate higher-quality texts (10). LLMs can translate documents for researchers who communicate in different languages (59). They may also be used in drug discovery and experimental design (61).

Despite the numerous benefits and potential applications of LLMs in medical research, there are several notable limitations. Some data provided by LLMs may be inaccurate (63, 64), and their literature sources may not be current (61), especially in models with a training cut-off date (65). Additionally, LLMs may reference non-existent articles, leading to further inaccuracies (63). There are concerns about plagiarism (61, 63, 66) and the potential for research misconduct, such as ghostwriting and the falsification or fabrication of research (61, 67-69). There is also a risk of citations of non-existent article references (63), and concerns about copyright infringement persist (67-69). These issues underscore the importance of carefully verifying information provided by LLMs and ensuring proper ethical standards are maintained in research practices.

Due to several limitations in the field of research, the International Committee of Medical Journal Editors (ICMJE) (70) and the World Association of Medical Editors (WAME) (71) have advised that Chatbots cannot be accountable for the accuracy, integrity, and originality of the work. Therefore, they do not meet the criteria for authorship and should not be listed as authors or co-authors (70-72). Instead, if LLMs or image creators and their information have been used in the article, the authors should transparently declare their application in both the cover letter (70, 72) and the submitted work in methods, acknowledgments, or a relevant section of the manuscript like figure legends (72).

Large Language Models in Medical Education and Exam Preparation

LLMs have shown great potential in assisting with ***medical education*** (61). Some of such applications and advantages of using LLMs in medical education are as follows:

- Immediate access to extensive information (73) and saving time needed for search of several articles and resources; Receiving direct and practical responses to the questions (59);
- Facilitating self-directed learning (44, 61, 74) by generating exercises, quizzes and self-checks for educational purposes (44, 59);
- Generating helpful mnemonics (74);
- Support for personalized learning experiences (2, 59, 73) according to the different learning styles and abilities of learners (59);
- Possibility of providing explanations according to the learner's level (47);
- Generating lists of possible differential diagnoses (61, 74, 75);
- Offering diagnoses or treatments (59);
- Development of clinical scenarios and clinical vignettes for training (76);
- Improvement of clinical skill development (73);
- Evaluating healthcare professionals (76);
- Providing individualized feedback on learners' work (2, 59);

For faculty and instructors, LLMs can enable innovative methods for teaching complex medical concepts, enhancing student engagement (73) and helping draft lesson plans (44). For instance, trainers could use quizzes and clinical vignettes generated by ChatGPT for instructional and assessment purposes, effectively utilizing them as virtual teaching assistants (44, 59). Compared to conventional, tutor-based educational methods, employing AI in education presents multiple advantages, such as cost-efficiency and the elimination of labour costs (57) and increased student engagement (44).

Examination preparation is one of the important aspects of medical education. Due to the sensitive nature of medical proficiency, several countries hold entrance or qualification exams for their future medical doctors to guarantee a minimum requirement of knowledge. Examples are the United States Medical Licensing Examination (USMLE), a three-step exam for medical licensure in the United States (77) and the Medical Council of Canada Qualifying Examination (MCCQE) for Canadian licencing (78). Board exams are other essential qualifications for many practitioners. In the United States, medical students invest around $4000 in external test preparation materials, mainly focusing on question banks and subscription services (73).

LLMs offer several advantages that make them a potential alternative for aiding exam preparation, including lower costs, 24-hour accessibility, the ability to provide direct answers to every question, and instant access to information (44). For example, a chatbot has been proven to support internationalization in higher education (79)

Pitfalls of Using LLMs in Learning Medicine and Exam Preparation

The biggest challenge in implementing LLMs in medical education is ensuring the ***accuracy*** and ***reliability*** of their provided information (44, 73). For example, there is a risk of training LLMs with inaccurate or incorrect information (10, 44). It should be noted that LLMs may generate misleading or inaccurate information, particularly if they receive poorly refined prompts from users with limited knowledge (73). Additionally, LLMs can not distinguish between fake and reliable information (80); they do not “understand” in the human sense. Instead, they merely learn probabilistic associations between words (10). Therefore, there might be a risk of generating incorrect, biased, or misleading answers (4, 29, 44, 61, 62, 73, 80). Learners with limited background knowledge may find it difficult to identify such errors, leading to further confusion. Ultimately, even minor errors from these future providers can have serious consequences for patient safety (44, 62) and potentially compromise community health.

The issue of accuracy in language models could be addressed by several measures, including validating the training data, fine-tuning the model to optimize medical accuracy, enhancing the model's capabilities through intelligent prompting techniques such as chain-of-thought reasoning, and incorporating uncertainty indicators to signal when the model's responses might be less reliable (10).

Some LLMs may have been trained on datasets that are fixed ***and not recent*** and up-to-date (46). This limitation could be addressed by enabling LLM models to access the internet when generating responses (10). Copilot, developed by Microsoft (48) is an example of a system with online internet access for responding (10).

Hallucinations are another issue when using LLMs. This term refers to instances where an LLM might create inaccurate information that isn't represented in the training dataset and present it as if it were a factual reality (fact fabrication) (7, 10). This issue may arise because LLM outputs are based on learned associations between words rather than an understanding of the input queries or the information used in outputs (10). To reduce hallucinations, LLMs can be developed by involving humans in the training process and using reinforcement learning from human feedback (RLHF) (7).

From what was discussed above, it is not advisable to rely exclusively on LLM-generated responses for guidance in clinically complex scenarios, as their accuracy in highly specialized fields has not been fully validated (73). The use of LLMs in medical education should be judiciously considered as a potential aid, possibly for providing hints to answer questions. Any responses to questions generated by LLMs should be validated with reliable medical sources of information.

It's important to recognize that LLMs cannot replace detailed lectures, review articles, guidelines and textbook materials for learning. Educators who incorporate LLMs into their teaching materials should emphasize this point in their curriculum (73).

Other challenges associated with using AI in medical education include the risk of overreliance on AI (73), the dilution of critical thinking skills (73), the potential impacts of job loss for medical educators that affect human resources (44, 73).

There are also several ***ethical concerns*** about using AI and LLMs in learning medicine. One ethical concern about the extensive use of LLMs in medical education and daily life is that students may engage in academic ***misconduc***t or ***plagiarism*** (44, 61, 73). For instance, a learner may use a chatbot to write an assignment without fully understanding the concepts or investing the necessary time and effort. To mitigate this issue, educators may design assessments that necessitate critical thinking, creativity, and information synthesis, which cannot be easily accomplished using LLMs.

Infringement of copyright laws could be another concern when using AI (80). Although most LLMs employ open-source material available to the public (28), some may use information from sources like books. Therefore, users of these applications may inadvertently breach copyright rules.

Other concerns about AI and LLM use arise from transmitting patient information and related data to external companies, which may violate ***patient privacy, data security and confidentiality*** (10, 44, 61, 62). This may also happen when students need to share sensitive patient information with their instructors or other parties for educational purposes, such as case reports or daily rounds (44). This ethical pitfall could be overcome by prohibiting input of patient-identifiable data in model prompts(10) or adequately de-identifying protected medical data (4).

Lastly, it is possible that some responses LLMs provide to medical questions could pose a risk, demonstrate discrimination, or offend certain individuals (10). Nevertheless, when errors happen, unlike human beings, we could not expect the LLMs to be accountable for their mistakes (44).

Due to the above-mentioned concerns about the misuse of LLMs in medical education, many universities have developed disciplinary rules for their students and faculty. For instance, they clarify on their websites that they expect students to complete their assignments independently, and if instructors prohibit the use of AI, employing such technology may be considered "unauthorized aid" (81).

AN INTRODUCTION TO FAMILY MEDICINE

Family medicine, a specialty in many countries, is central to primary care. It deals with the treatment of a wide range of diseases and provides comprehensive healthcare for people of all ages, from newborns to seniors (82). Family physicians care for patients regardless of age or health condition, building enduring and trusting relationships. Family medicine specialists know community-level factors and social determinants of health, serve as a patient's first contact for health concerns, and navigate the healthcare system with patients, including coordinating specialist and hospital care (82).

Family medicine is the primary specialty in delivering comprehensive care across all ages, genders, and health conditions in Canada. Medical doctors who have completed medical school require two years of postgraduate training to meet the College of Family Physicians of Canada (CFPC) requirements. This residency consists of rotations in various medical fields, including general surgery, internal medicine, pediatrics, obstetrics/gynecology, psychiatry, and emergency medicine. Additionally, residents have the opportunity to pursue electives in areas of their choosing, such as rural family medicine. As of 2019, there have been 44584 family physicians in Canada, covering an average of 119 physicians per 100,000 population, with 47% being female (83).

The College of Family Physicians of Canada (CFPC) administers a detailed certification exam to assess the clinical skills of family medicine practitioners in Canada, which serves as the board exam for family medicine in Canada (84). The CFPC exam has two main parts: the oral part and the written part. In the oral part, also known as the simulated office oral exam (SOO), candidates go through five different 15-minute scenarios where a family physician examiner acts as a patient (84). The written part consists of short-answer questions (SAMPs) and is done on a computer over four hours.

Several studies demonstrate satisfactory outcomes using both general domain LLMs (41, 46, 54, 85, 86) and medical domain LLMs (33, 38, 56, 87-90) in answering medical exam questions with acceptable results. ChatGPT is one of the most extensively studied LLMs in this domain (29, 59, 61, 86, 91-95). Few studies have addressed the possibility of use of LLM-based chatbots to answer questions from family medicine exams (85, 95). We will discuss some of these studies and LLMs in the second chapter of this thesis, titled "Review of the Literature."

In this thesis, we have drafted two articles to evaluate the usefulness of LLM-based chatbots on a sample questionnaire for the SAMPs exam from the CFPC website (84).

The first study, "Performance of ChatGPT in Certification Examination of College of Family Physicians of Canada," focused on the accuracy of the answers from GPT-3.5 and GPT-4 to our questionnaire. We examined the accuracy of responses under five different conditions: a baseline with prompt limiting answers to fewer than ten words per line, rerunning the models after a one-week interval, regenerating the prompt, removing the mention of the CFPC exam from the prompt, and using a zero-shot prompt (with no word limit specified).

In the second article "Assessment of Ten Different Large Language Model-Based Chatbots on Certification Examination of College of Family Physicians of Canada Examination," we compared the performance of 10 LLM-based chatbots on our questionnaire

These two articles will be delivered in the "body of the thesis."

CHAPTER 2:

REVIEW OF LITERATURE

Benchmarks for evaluation of LLMs on Medical Questions and Answers

Several studies have used benchmarks to evaluate the performance of language learning models (LLMs) in answering medical questions and answers. In this chapter, we briefly introduce some of these benchmarks. Then, we will provide examples of studies that have used LLMs to help answer such questions.

● **MedQA (USMLE)**

The MedQA dataset comprises USMLE-style questions, each featuring four or five potential answers. It includes a development set with 11,450 questions and a test set containing 1,273 questions (40). This dataset has been used as a benchmark in many studies (36, 38-41).

● **MedMCQA**

The MedMCQA dataset includes over 194,000 four-option multiple-choice questions derived from two Indian medical entrance examinations called AIIMS and NEET-PG. It includes 2,400 healthcare topics across 21 medical subjects (40). Several researchers have used this questionnaire to evaluate LLMs (36-41, 89).

● **PubMedQA**

This dataset features 1,000 question-answer pairs labelled by experts. A PubMed abstract is chosen for the context, and the answers are yes/no/maybe multiple-choice to the question (40). This questionnaire has been used in several studies (36-41).

● **MMLU (Massive Multitask Language Understanding)**

MMLU includes multiple-choice exam questions with four options from 57 domains (e.g. Clinical knowledge, Anatomy, Medical genetics, Professional medicine, College biology, College medicine, etc.) (40). Many studies have also used different sections of this benchmark (38-40).

● **MultiMedQA**

MultiMedQA encompasses a variety of medical exams and research datasets that feature multiple-choice answers, as well as datasets with consumer medical questions requiring long-form responses. This collection includes the MedQA, MedMCQA, PubMedQA, MMLU clinical topics, LiveQA, and MedicationQA datasets (40).

– **LiveQA**

LiveQA is a collection of free, open-source texts with long answers. This dataset includes medical questions submitted by individuals to the National Library of Medicine (NLM), accompanied by manually collected reference answers from trusted sources like the National Institute of Health (NIH) website (40).

– **MedicationQA**

The MedicationQA dataset is an open-domain resource that comprises a compilation of frequently asked consumer questions. The questions are followed by long answers (40).

– **HealthSearchQA**

HealthSearchQA is an open-source dataset consisting of 3,173 commonly searched consumer questions about medical conditions and their associated symptoms developed by Singhal and his coworkers during the development of Med-PaLM (40). This part was their unique part of MultiMed QA only used in their study.

● **USMLE sample exams**

USMLE sample exam questions have been used in many studies (36, 37, 86, 89, 92, 93, 96). For instance, publicly-available test questions from the June 2022 sample exam release (89, 92), USMLE Self-Assessments (86, 93) or questions from professional medical board exams (37).

● **iCliniq Dataset**

The “iCliniq” is an online medical consultation site. Roughly 10k conversations from this website were used to generate a database comprising patient questions and answers. This database was used to qualitatively assess *ChatDoctor*, a fine-tuned LLM trained on medical data. The answers from the human physicians served as the benchmark or “ground truth (90).

● **Other Questionnaires**

Other questionnaires employed for benchmarking LLMs and chatbots are a variety of different medical questionnaires. These include questionnaires from medical professional exams(97-99) , questions from different medical licencing authorities (e.g. Royal College of General Practitioners Applied Knowledge Test (AKT) (100)) and questionnaires in different languages other than English (e.g. Japanese (101), Chinese (87, 102), and Persian (103))

● **CFPC Sample Questionnaire**

Our two studies delivered in the body of this thesis employed a selection of Short-Answer Management Problems (SAMPs) sourced from the College of Family Physicians of Canada (CFPC) official website (84). This selection included 19 clinical scenarios, along with 77 corresponding questions. The clinical scenarios examined a variety of family medicine areas, including:

Cardiology, with three questions on atrial fibrillation, dizziness, and hypertension; one endocrinology question on osteoporosis; and two emergency medicine questions on lacerations and poisoning. Gastroenterology was represented with three questions about abdominal pain, abnormal liver function tests, and dyspepsia. Additionally, there was one musculoskeletal disease question about low back pain, a neurology question on seizures, and two gynecology questions on a breast lump and fertility. One question combined pediatrics and infectious disease regarding fever in a newborn, while psychiatry was covered with two questions about depression and eating disorders. Other topics included one respirology question on chronic obstructive pulmonary disease, one urology question on prostate issues, and one question about fatigue (84).

Each scenario was linked to between 2 and 7 questions. These questions are crafted to mimic the structure of the actual computer-based examination and demand concise, succinct answers (84). Typically, responses should be limited to no more than ten words per line, with each question requiring between 1 to 5 lines of answer (85).

General Domain Large Language Models

ChatGPT

ChatGPT is the most studied LLM on medical exams. It has shown promising results in various medical exams, indicating its potential value in medical exam preparation (61). Many studies compare the performance of their designed LLMs with ChatGPT, using GPT-3.5 or GPT-4 as the gold standard of reliable LLM with acceptable results. We will discuss them in the next section. Here, we deliver the results of some examples of studies that have used GPT-3.5, GPT-4, or both in several different medical exams.

● In the field of ***CFPC exam*** preparation, a multiple-choice progress test administered by the University of Toronto was used to compare family medicine residents' performance against ChatGPT, using ***GPT-3.5 and GPT-4***. In this study, investigators showed that GPT-3.5's performance was comparable to that of family medicine residents achieving 57.4% accuracy, while GPT-4 outperformed both residents and GPT-3.5 with 82.4% correct answers (95).

● ***USMLE exam*** is based on clinical case scenarios that require in-depth medical reasoning and justification of treatment decisions (92, 93). Of note, human performance on USMLE has been 60% at passing score and 87% at expert score levels (89).

The effectiveness of LLMs on the different steps of USMLE questions is an evolving area of research. ***ChatGPT (GPT-3.5)*** has been used to answer the different steps of the USMLE exam (86, 92, 93). The Study by ***Kung et al***. showed that LLMs, such as ChatGPT, can achieve scores around 60% on USMLE Steps 1, 2CK, and 3, indicating a substantial capability for medical knowledge retrieval (92). Similarly, ***Gilson and coworkers*** found an accuracy range of 44% to 64.4% for sample USMLE Steps 1 and 2 questions (83). However, the scores reported for GPT-4 are higher. For instance, one of the studies confirmed 86.1% scoring by ChatGPT-4 on the MedQA (88). Another study reported a passing score of more than 20 points, on the USMLE by ChatGPT-4 (86).

● ChatGPT has been used to answer several other membership and licencing exam questions. Most of the studies acknowledge better results from GPT-4 compared to GPT-3.5 (100-102).

While ChatGPT (GPT-3.5) achieved an accuracy of 60.17% on the Membership of the Royal College of General Practitioners Applied Knowledge Test (AKT), it fell short of the 70.45%

passing threshold for this primary care exam (100). Similarly, on the Japanese medical licensing exam, ChatGPT's performance did not meet the passing criteria, although GPT-4 was successful (101). GPT-4, however, achieved an accuracy of 81.3% on the Iranian Medical Residency Examination, a 200-question multiple-choice test encompassing diverse medical specialties (93).

Finally, ChatGPT was shown to perform lower than that of medical students on the Chinese National Medical Licensing Examination, below the passing threshold (102). However, another study on the same exam showed that GPT-4 significantly outperformed ChatGPT-3.5 and other LLMs like LLaMA, Alpaca, and Vicuna (87).

● ChatGPT has demonstrated its potential in answering various other medical exams, achieving satisfactory results. These include the Advanced Cardiac Life Support (ACLS) exam (104), as well as specialty exams in ophthalmology (97) neurology (105) and radiology (99).

● ChatGPT's capabilities have proven effective not only in English-based medical exams but also in tests administered in ***other languages***. For instance, ChatGPT has been successfully used to answer the Japanese medical licensing examination (101), the Chinese National Medical Licensing Examination (87, 102) and the Iranian Medical Residency Examination (103). Notably, in the later mentioned study, even when the questions were translated into English, French, and Spanish, the accuracy of the responses remained almost consistent, ranging from 81.3% for Persian to 84.3% for the English language (103).

InstructGPT and Codex

Liévin and his coworkers investigated the 175B parameter GPT-3.5 series (Brown et al., 2020), the human-aligned GPT-3 (InstructGPT, text-davinci-002, Ouyang et al. (2022)) and the code-finetuned GPT-3 (Codex, code-davinci-002, Chen et al. (2021)) from Open AI on USMLE, the MedMCQA, and PubMedQA datasets. In this study, investigators studied the effectiveness of zero-shot and few-shot Chain-of-Thought (CoT) prompting. In the zero-shot approach, they employed a two-stage prompting strategy. First, an initial prompt aimed to elicit a CoT using a specific cue (e.g., “Let’s think step by step”). The completion of this prompt served as the CoT itself. The second stage involved an extractive prompt, where the completion provided the answer (e.g., "Therefore, the answer is"). For their few-shot CoT approach, they experimented with incorporating exemplars (examples) within the prompts. These exemplars included question-answer pairs and even question-explanation-answer triplets (89).

In this study, using zero-shot answering InstructGPT (text-davinci-002) achieved 46%, on the USMLE (test), 44% on MedMCQA and 73.2% on PubMedQA datasets. CoT prompts did not achieve better than the direct prompt. The researchers acknowledged that InstructGPT can often read, reason and recall expert knowledge (89).

In contrast, 175B parameters Codex (code-davinci-002) using 5-shot achieved scored 60.2% on USMLE, 62.7% on MedMCQA and 78.2% on PubMedQA which was close to human-level performances (89).

Flan-PaLM2

Flan-PaLM, developed by Google researchers, is the instruction-tuned model from PaLM LLM (540 billion parameters) using a combination of prompting strategies, (42). Flan-PaLM improved task performance over PaLM (42). Flam-PaLM achieved 67.6% accuracy on MedQA (40). However, its performance on MultiMedQA, remained inferior to the clinicians (42).

Gemini

Pal et al. conducted a comparative study between some LLMs and Gemini the new model from Google using Gemini Pro 1.0 LLM. They assessed Gemini performance in medical reasoning, hallucination detection, and medical visual question answering (VQA) tasks (88).

Gemini achieved an accuracy of 61.45% on the medical VQA dataset, considerably lower than GPT-4 (88%). While Gemini demonstrated some competence, it fell behind MedPaLM 2 and GPT-4 in terms of diagnostic accuracy. The study further uncovered significant drawbacks of Gemini, including high susceptibility to hallucinations, overconfidence, and knowledge gaps, which could pose risks if the model is employed without appropriate critical evaluation (88).

Large Language Models Trained on Medical Domain Knowledge

● GatorTronGPT

GatorTronGPT is the first generative LLM designed explicitly for clinical medicine applications (33, 35). It was developed using 277 billion words from a mix of clinical and English texts, employing a GPT-3 architecture (used by ChatGPT) with 20 billion parameters. When physicians were asked to evaluate its outputs on a scale from 1 (worst) to 9 (best), the results indicated no significant difference in linguistic readability between GatorTronGPT and human writing, so physicians were unable to distinguish between them (35). While GatorTronGPT showed promise, its capacity for generalization was limited, possibly due to constraints related to its model size (33).

● **MedPaLM**

To develop MedPaLM, initially, PaLM LLM (540 billion parameters) and Flan-PaLM, its instruction-tuned model, were prompted using a combination of ***few-shot Chain-of-thought (COT)***, and ***Self-consistency (SC)*** techniques. (*See Chapter One for the definition of few-shot learning, Chain of thought and self-consistency*). Few-shot learning uses a limited volume of data to make accurate results. This approach is different from traditional machine learning, which often requires vast amounts of data to achieve high accuracy In the next step, ***instruction-prompt tuning*** was done using a method that is both data- and parameter-efficient and aimed at refining Flan-PaLM's alignment with the medical domain (40).

Med-PaLM was the first model to achieve a "passing" score on US Medical Licensing Examination (USMLE) style questions, scoring 67.2% on the MedQA dataset (41). While MedPaLM showed promising performance, however, it did not reach the proficiency level of clinicians in human evaluation (40, 41).

● **Med-PaLM 2**

Med-PaLM 2 was developed using several enhancements compared to Med-PaLM: improvements to the base LLM (PaLM 2), fine-tuning for the medical domain, and the implementation of prompting strategies that incorporate a novel ensemble refinement approach (41). Instruction fine-tuning was performed using the training splits of MultiMedQA (Refer to "*Benchmarks for evaluation of medical performance*" section) (41). Med-PaLM 2 achieved a performance score of up to 86.5% on the MedQA dataset (41). Improved outcomes Med-PaLM 2 compared to Med-PaLM on medical benchmarks, indicated that the performance of LLMs in answering medical questions is improving significantly (41).

Med-PaLM 2 displayed excellent performance, comparable to GPT-4 across all MultiMedQA multiple-choice benchmarks, which include MedQA, MedMCQA, PubMedQA, and MMLU (*refer to "Benchmarks for evaluation of medical performance"*) (41). Human evaluations on long-form questions revealed that physicians favoured the answers generated by Med-PaLM 2 over those from physicians in eight out of nine categories (41).

● **ChatDoctor**

ChatDoctor (accessible from https://app.chatdoctor.com) has been developed by modifying and improving the large language model meta-AI (LLaMA) to tackle the limitations observed in the medical knowledge of other LLMs, including ChatGPT. It was developed to understand patients' questions utilizing a vast dataset of 100,000 dialogues between patients and doctors from a popular online medical consultation platform. Moreover, the model has access to up-to-date information from online platforms like Wikipedia and data from carefully selected offline medical databases (90).

The responses generated by ChatDoctor were compared to those from ChatGPT using "***BERT-Score***". BERT-Score, which correlates closely with human judgment, utilizes pre-trained BERT to match words in the candidate and reference sentences, with higher scores indicating a better match. The results demonstrated that ChatDoctor's BERT-Scores were significantly higher than those of ChatGPT (90).

● **MedAlpaca**

MedAlpaca was developed based on LLaMA 7B and LLaMA 13B open-source LLM variants for research purposes. Medical flashcards used by medical students, questions and answers from the StackExchange dataset (by academia) (96), question-answer pairs from Wikidoc and some other benchmarks were used to finetune the base LLM. MedAlpaca was tested to answer the USMLE test set and showed that fine-tuned MedAlpaca LLMs consistently outperformed their base model pairs (96).

● **ClincalGPT**

ClinicalGPT is a large language model specifically tailored and refined for clinical settings. It integrates a wide array of real-world data in its training, including medical records, specialized domain knowledge, and multi-stage dialogue consultations. This model was created using an ***instruction-tuning*** method supplemented by supervised fine-tuning (SFT) of the BLOOM 7B LLM (42). ClinicalGPT underwent evaluation in various medical contexts, such as answering medical knowledge questions, handling medical examinations, conducting patient consultations, and performing diagnostic analyses of medical records and showed promising results (42) and surpassed BLOOM-7B, its base model,as well as LLaMA-7B and ChatGLM-6B (42).

● **Clinical-Camel**

Clinical Camel, explicitly designed for clinical research, is an open LLM developed by fine-tuning from LLaMA-2 employing Quantization and Low-Rank Adaptation (QLoRA). It was trained on clinical datasets, including ShareGPT a multi-step conversation, 20,000 pre-2021 open-access articles and 4000 randomly selected multiple-choice questions from MedQA (10,178 multiple-choice questions pool) (36).

Clinical Camel showed an excellent performance in generating realistic clinical notes (36). It outperformed GPT-3.5 in five-shot evaluations on all the benchmarks, achieving 64.3% on the USMLE Sample Exam (versus 58.5% for GPT-3.5), 77.9% on PubMedQA (versus 60.2%), 60.7% on MedQA (versus 53.6%), and 54.2% on MedMCQA (versus 51.0%). Beyond these metrics, Clinical Camel also excels in generating realistic clinical notes. Although Clinical-Camel exceeded GPT-4's results on PubMedQA, it does not reach the performance levels of GPT-4 or Med-PaLM 2 on other benchmarks (36).

● **PMC-LLaMA**

PMC-LLaMA is a lightweight (13 Billion parameter) medical-domain LLM built on LLaMA. Initially, pretraining was done using 4.8M biomedical academic papers (the largest number of tokens from PubMed Central academic papers) and 30K textbooks to develop PMC-LLaMAK. Instruction tuning was then done on a dataset, which consists of medical questions and answers (37).

PMC-LLaMA (13B parameters) was compared to ChatGPT (175B), LLaMA 2 (13B and 70B), MedAlpca (13B) and ChatDoctor (7B) across the MedQA, MedMCQA, PubMedQA benchmarks. Despite having fewer parameters, PMC-LLaMA outperformed ChatGPT on PubMedQA and matched its performance on MedQA. Moreover, PMC-LLaMA demonstrated superior performance compared to general domain LLMs, LLaMA 2 (13B and 70B), and medical-specific LLMs, MedAlpca (13B) and ChatDoctor, across all three benchmarks (37).

● **MediTron**

MediTron, available in 7B or 70B parameters is an open-source large language model (LLM) tailored for the medical domain. It is developed by continued pretraining of LLaMA 2 on medical data sources, using an adaptation of Nvidia's Megatron-LM distributed trainer (38).

MediTron was evaluated on MedQA, MedMCQA, PubMedQA, and MMLU. Overall, MediTron 7B has been superior to LLaMA 2-7B (the base model) and PMC-LLaMA-7B. This lightweight version also outperforms the leading instruction-tuned models, Mistral and Zephyr-β, in almost all benchmarks, with the sole exception of MMLU-Medical (38).

The MediTron 70B (with ten times more parameters), showed excellent performance and was superior to its baseline models as well as its instruction-tuned counterparts e.g. Clinical-Camel-70B and Med42-70B (38). Compared to GPT-3.5 with 175B parameters, MediTron 70B showed superior performance on all benchmarks and was comparable to other commercial LLMs, GPT-4, Med-PaLM and Med-PaLM-2 with significantly larger parameter sizes (38).

Research on MediTron indicated that prolonged pretraining at the 70B scale is beneficial and demonstrated that enlarging the domain-specific pretraining dataset substantially enhances performance (38).

- **Me LLaMA**

Me LLaMA is an example of an effort to develop an LLM to enhance medical tasks. This open-source LLM is built using a combination of biomedical and clinical data. The chat-enhanced models, Me LLaMA 13/70B-chat, were created by continuously pre-training and instruction tuning LLaMA2 with balancing biomedical, clinical, and general domain data (33).

Me LLaMA, demonstrated superior performance compared to other open-source medical LLMs in both general and medical tasks. Specifically among 13B systems, the Me LLaMA 13B-chat model excelled over LLaMA2 13B-chat (a general domain LLM), PMC-LLaMA-chat (an LLM trained on biomedical literature), MedAlpaca 13B, and AlpaCare-13B in the majority of their 12 datasets. Me LLaMA13B-chat also performed better than LLaMA2-70B-chat (a general domain model with more parameters). In the 70B parameter category, Me LLaMA 70B-chat consistently outperformed MediTron 70B (an LLM trained on biomedical literature) across all 12 datasets and surpassed LLaMA2-70B-chat (a general domain LLM) in 9 of the 12 datasets (33).

Me LLaMA 13B and 70B were compared with their backbone models LLaMA2 13B and 70B in the zero-shot setting, and the authors concluded that continual pre-training is beneficial (*See Chapter One for the definition of zero-shot*). They also reported that compared to Me LLaMA 13/70B, their chat optimized versions Me LLaMA-13/70B-chat that received instruction tuning, showed advantages in zero-shot contexts (33).

Me LLaMA chat exceeded ChatGPT in performance on 6 of 8 datasets. When compared to ChatGPT-4, Me LLaMA exhibited better results on three datasets, showed comparable yet marginally lesser performance on another three, and underperformed on the last two datasets. They concluded that Me LLaMA chat possesses enhanced capabilities for understanding and contextually managing medical data (33).

● **BioMistral**

BioMistral was created by further pretraining the base LLM model of Mistral 7B on PubMed Central (PMC-open access), a well-known biomedical resource (39).

● ***BioMistral*** was evaluated alongside several other LLMs across benchmarks such as MedQA, MedMCQA, PubMedQA, and MMLU. Notably, BioMistral 7B outperformed the instructed version of Mistral 7B on eight out of ten tasks, highlighting the advantages of domain-specific pretraining.(39). In this study, BioMistral 7B exceeded the performance of MediTron 7B and MedAlpaca 7B on MedQA. This open-source LLM also surpassed MediTron 7B, MedAlpaca 7B, and PMC-LLaMA 7B on MMLU and MedMCQA, but it performed less effectively on PubMedQA (39). However, in experiments involving a 3-shot scenario, GPT-3.5 Turbo with 1106 parameters outshined all other models, including Mistral 7B Instruct, BioMistral 7B, MedAlpaca 7B, PMC-LLaMA 7B, MediTron-7B, and BioMedGPT-LM-7B across the four assessed benchmarks (39).

● **The Open Medical-LLM Leaderboard: Benchmarking LLMs in Healthcare**

The Open Medical LLM Leaderboard is intended to monitor, rank, and evaluate the performance of LLMs. Users can submit an LLM model for assessment on medical question-answering tasks using datasets such as MedQA (USMLE), PubMedQA, MedMCQA, and relevant subsets of MMLU focused on medicine and biology (106). These tasks encompass both multiple-choice and open-ended questions that require medical reasoning and understanding (106). This platform allows researchers and practitioners to pinpoint the strengths and weaknesses of various approaches, foster further advancements in the field, and ultimately enhance patient care and outcomes (106).

According to the Open Medical-LLM Leaderboard, while commercial LLMs like GPT-4 and Med-PaLM-2 consistently demonstrated accuracy across various medical datasets, some open-source LLMs, such as Starling-LM-7B, gemma-7b, Mistral-7B-v0.1, and Hermes-2-Pro-Mistral-

7B, exhibit competitive performance on certain datasets and tasks, despite having smaller sizes of approximately 7 billion parameters (106). Furthermore, open-source LLMs, such as GPT-4 and Med-Palm-2, showed an acceptable performance in comprehension and reasoning over PubMedQA and applying clinical knowledge and decision-making skills (MMLU Clinical Knowledge subset) (106).

Performance of LLMs in Family Medicine Exam

Although many studies have evaluated LLMs' performance on licensing medical examination questions (92, 93, 100), few have explored their effectiveness on family medicine questions (95), especially the open-ended questions from the Canadian family medicine exam.

Huang et al. used ChatGPT to answer a series of multiple-choice progress tests designed to prepare University of Toronto residents for the CFPC exam. They showed that GPT-4 achieved an accuracy of 82.4%. and significantly outperformed GPT-3.5 at 57.4% and family medicine residents who answered 56.9% correctly (95). However, they used a multiple-choice questionnaire from their university that was not standardized and was different from the format of the CFPC exam.

In our studies, we used a standardized set of questions directly sourced from the CFPC website. These questions were open-ended, mirroring the SAMPs structure, and included official answer keys approved by the CFPC, providing a more accurate representation of the CFPC exam format.

SUMMARY OF KNOWLEDGE GAP

Recently, there has been considerable advancement in the field of AI, especially with LLMs. Presently, numerous LLMs and chatbots are accessible to the public, with a marked interest in integrating them into daily medical practice, research, and education (61). LLMs have been employed to respond to questions from various medical examinations, aiming to enhance learning or evaluate their utility in answering medical questions (61). These exams include licencing exams (33, 36-41, 86, 87, 92, 93, 96, 100-102), professional board exams (97, 99, 105) or even exams conducted in other languages (87, 101-103).

This thesis aims to address six potential knowledge gaps:

1) Few attempts have been made to estimate the usefulness of the LLMs to answer open-ended responses to the questions (107, 108), particularly in the field of family medicine and CFPC exam preparation. The sole attempt was the application of ChatGPT on a multiple-choice progress test designed for resident preparation sourced from the University of Toronto (95). However, the CFPC exam involves open-ended comprehensive questions, and no studies have evaluated LLMs on comprehensive, open-ended exams.

2) Although several efforts have been made to assess different LLMs across diverse exam database settings, the consistency of the LLM responses has not been extensively studied. It is not clear whether the accuracy of LLM-based chatbots would vary under different situations, such as questioning at different times, regenerating answers, or using prompts for debriefing.

3) Recently, several new LLMs with different capabilities have been introduced. They have employed larger-size parameters or different training methods. A notable example is “Gemini”. Despite their broad public availability, the performance of such new emerging LLMs in the medical exams has not been widely evaluated.

4) General domain LLMs, e.g. ChatGPT, are not specifically designed for medical purposes and may lack accuracy in this field (33). Recently, there has been a growing interest in developing LLMs embedded with medical knowledge with the aim of enhancing accurately and creating more practical LLMs for medical application (33, 38-41, 89, 90). However, it is unclear whether such specially trained LLMs would outperform the widely available LLMs trained on large parameters.

5) Some LLM-based chatbots like ChatGPT do not have access to information, while others claim an advantage of providing access to online searches. The effectiveness of this feature in addressing complex family medicine questions remains uncertain.

6) lastly, while many advanced LLMs-based chatbots require subscription fees, others are available for free through subscription services. The benefits of paid subscriptions to advanced systems over the free models are yet to be determined.

RESEARCH OBJECTIVES AND HYPOTHESIS

Manuscript 1:

Considering the identified knowledge gaps, we initiated our first study to assess the performance of GPT-3.5 and GPT-4 in responding to a series of sample open-ended SAMPs questions from the CFPC website. We conducted the test across five different rounds under various conditions to evaluate the consistency and accuracy of the LLM-generated responses.

Our primary null hypotheses were as follows:

- H0: There is no difference in the CFPC score percentages for answers provided by GPT-3.5 and GPT-4 for each question.

In the second phase of our study, we examined the consistency and variability of responses by running the same sample SAMPs questionnaire under five different conditions:

- Round 1: We used "Prompt 1*" for the first time, which restricted answers to less than ten words per line, mirroring the real CFPC exam constraints.
- Round 2: After about one week, we used "Prompt 1*" again to test response consistency.
- Round 3: We regenerated the answers from the second round.
- Round 4: We used "Prompt 2*", which did not reference the CFPC exam.
- Round 5: We responded to the set of questions without any prompt (Zero shot).

To analyze the results from these repeated measures, we employed the ordinal logistic Generalized Estimating Equation (GEE). Our hypotheses for this analysis were:

- H0: The cumulative CFPC score percentage over five rounds is not significantly different between GPT-3.5 and GPT-4.
- H0: The cumulative Reviewers' score percentage over five rounds is not significantly different between GPT-3.5 and GPT-4.

Manuscript 2:

In our second study, we evaluated the performance of ten LLM-based chatbots on the same questionnaire used in our first study to compare their accuracy. The systems assessed included ***ChatGPT-3.5,*** known as ***ChatGPT***, (11), ***YouChat Smart Mode (You-sm)*** (53), ***ChatGPT Plus (ChatGPT-4)*** (51), ***YouChat GPT-4 mode (You-4)*** (49), ***Copilot*** (48), ***Gemini*** (30), ***Personal***

Intelligence (PI-AI) (27), ***Gemini Advanced (Gemini-A)*** (30), ***AI doctor (AI-Dr)*** (56).and ***Chat Doctor (ChatDr)*** (90).

Our hypothesis was:

H0: The CFPC score given to the sample SAMPs questionnaire is not significantly different among all 10 LLM-based chatbots.

H0: The Reviewers' score given to the sample SAMPs questionnaire is not significantly different among all 10 LLM-based chatbots.

CHAPTER 3:

BODY OF THESIS; STUDY 1

Performance of ChatGPT in Certification Examination of College of Family Physicians of Canada

Citation: Mousavi M, Shafiee S, Harley JM, Cheung J, Rahimi SA. Performance of ChatGPT in Certification Examination of College of Family Physicians of Canada, doi: 10.1136/fmch-2023-002626. PMID: 38806403; PMCID: PMC11138270. Fam Med Community Health. 2024 May 28(12 (Suppl 1)):e002626. **(85)**

Authors:

Mehdi Mousavi M.D., CCFP[1],

Shabnam Shafiee M.D., CCFP[2],

Jason M. Harley Ph.D.[3,4,5],

Jackie Cheung Ph.D.[6,7],

Samira Abbasgholizadeh Rahimi B.Eng, Ph.D.[7,8,9]

1 Department of Family Medicine, Faculty of Medicine, University of Saskatchewan, Nipawin, SK, Canada https://orcid.org/0000-0003-3644-5741
2 Riverside health complex, Prairie north health region, Saskatchewan Health Authority, Turtleford, SK, Canada
3 Department of Surgery, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada https://orcid.org/0000-0002-2061-9519
4 Research Institute of the McGill University Health Centre, Montreal, QC, Canada
5 Institute for Health Sciences Education, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada
6 School of Computer Science, McGill University, Montreal, QC CANADA
7 Mila Quebec AI-Institute, Montreal, QC, Canada
8 Department of Family Medicine, McGill University, Montreal, QC, Canada
9 Faculty of Dentistry Medicine and Oral Health Sciences, McGill University, Montreal, QC, Canada

ABSTRACT

Introduction: The application of large language model AIs such as GPT has been promising in medical education, and its performance has been tested for different medical exams. This study aims to assess the performance of ChatGPT 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 ran ChatGPT (GPT -3.5 and GPT -4) 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 (CFPC score), and second, based on their knowledge and other references (Reviews' score). An ordinal logistic Generalized Estimating Equations (GEE) model was applied to analyze 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 (Odds Ratio: 2.31; 95% Confidence Interval: 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 (Odds Ratio: 2.23; 95% Confidence Interval: 1.22 to 4.06; $P= 0.009$).

Running the GPTs after one weak 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 learning family medicine education and exam preparation needs further studies.

Key words:

Artificial intelligence, AI, ChatGPT, education, examination, family medicine, GPT-3.5, GPT-4, language model, medical education, medical knowledge exam, testing, test

Key Messages:

What is already known on this topic:

Prior to this study, there was an understanding of the general capabilities of AI models like ChatGPT in various applications and some exams. However, ChatGPT and ChatGPT Plus are not specifically designed for medical purposes and 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 AI tools 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, will 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 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, CA, USA) in November 2022, is a large language model (LLM) artificial intelligence (AI) to generate humane-like dialectic responses to text inquiries. ChatGPT, by default, uses the 3.5 version and is the freely accessible version. In contrast, ChatGPT Plus, a paid version that uses GPT-4.0, 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 and research, 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 (e.g. 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 (SOO), and the written component, consisting of short-answer management problems (SAMPs). Typically, SAMPs include around 40 clinical scenarios, with two to seven questions about the 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 prompts for debriefing could influence the accuracy of responses. Therefore, we conducted this study to assess the performance of both ChatGPT and ChatGPT Plus (using GPT-3.5 and GPT-4, respectively) 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 Short-Answer Management Problems (SAMPs) obtained from the official website of the College of Family Physicians of Canada (CFPC) (19). This sample set comprises 19 clinical scenarios, accompanied by a total of 77 questions related to these scenarios. Each scenario has between 2 and 7 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).

Table 1. Diverse topics within the spectrum of family medicine represented in the sample SAMPs 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
Pediatrics 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

Data collection

We employed GPT-3.5 (ChatGPT, August 3, Version 2023, OpenAI, San Francisco, CA) and GPT-4 (ChatGPT Plus, August 3, version 2023, OpenAI, San Francisco, CA) from August 8 to August 24, 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 (i.e., 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 (i.e., the tendency of ChatGPT 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 summarizes the various rounds in which both GPT-3.5 and GPT-4 were utilized for our study.

Table 2. Summary of the multiples runs using GPT-3.5 and GPT-4 between August 8 to August 24, 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 and 10 Aug	11, 12 Aug
Round 2: We used "Prompt-1*" for the second time	15, 16 Aug	19 Aug
Round 3: We regenerated the answers of the second run	15, 16 Aug	19 Aug
Round 4: We used "Prompt 2*"	16, 17, and 19 Aug	21 Aug
Round 5: We answered the set of questions without any prompt	19, 24, 25 Aug	23, 24 Aug
* 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."		

Scoring and review of responses

Two experienced CFPC-certified practicing family physicians independently reviewed and scored the AI-generated responses and explanations. Both reviewers, M.M., and S.S., possess over 2 years of Canadian family medicine practice experience. However, they are International Medical Graduates (IMGs) with over 15 years of extensive professional backgrounds. M.M. 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 1 or 0 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 (Aug 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.

Data Analysis:

We conducted our statistical analyses using SPSS 16.0 software (SPSS Inc., Chicago, IL). 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 (standard deviation) 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 Ranks nonparametric test. To compare the outcome of repeated measures across five rounds of GPT-4 and GPT-3.5 results, we utilized the ordinal logistic Generalized Estimating Equation (GEE). The outcome variables were The CFPC score, or reviewers' score, categorized as 0, 33.3, 50, 66.67, 75, 80 and 100. The independent variable was the usage of GPT-3.5 versus 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 analyzed 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 2 to 7 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 categorized 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 to GPT-3.5, with GPT-4 being 2.31 times more likely to score higher (Odds Ratio: 2.31; 95% Confidence Interval: 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 (Odds Ratio: 2.23; 95% Confidence Interval: 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 toward an increase in 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).

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. Data are presented as median [25 percentile, 75 percentiles]. Mean (SD) is given for comparison. Wilcoxon Signed Ranks test was done to compare GPT – 3.5 scores and GPT-4 Scores.

	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)	

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 Reviewers knowledge; Round: For definition of Rounds see table 2;

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 one-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 Ranks 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 to 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= 0.83 and 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= 0.058 and 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).

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 Rank non-parametric test for comparison), left and Percentage of questions that did not show a change in score for CFPC and Reviewers' score.

The total Number of Questions was 77.

	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%)

Text Box 1 (Appendix) 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 ChatGPT to answer the Sample CFPC questions and responded satisfactorily to our complex sample questions. When the reviewers scored the questions utilizing 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. So, only about 16 % of the lines of answers provided by GPT-3.5 and 7 % were deemed incorrect in 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's 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 standardization. In contrast, our study employed a comprehensive and standardized 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 2CK, and Step 3 of the USMLE (9). Similarly, Gilson and colleagues observed an accuracy range of 44% to 64.4% for sample USMLE Step 1 and Step 2 questions (10). ChatGPT's performance on the Chinese National Medical Licensing Examination (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 generalized estimating equation model revealed that the CFPC score percentages were significantly higher for GPT-4 than GPT-3.5 ($P<0.001$). Likewise, upon 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 to 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 (NBME) (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 Textbox 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, AI has 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 hallucination (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 one-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," markedly 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 this phenomenon.

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 minimize potential learning curve effects on the AI'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 emphasize 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 realized 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 cutoff 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 programs 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 emphasize 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 specialized for medical purposes (1). It's important to note that other large language models (LLMs) may utilize more recent sources of information, potentially yielding different results and warranting further investigation. Furthermore, even within the same version of OpenAI, the AI'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 2 to 7 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 SAMPTs 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 do not faithfully 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 its GPT-4 iteration—our study suggests that these AI models are promising as supplementary learning tools for candidates preparing for the CFPC exam.

REFERENCES

1. OpenAI. Models: OpenAI; 2023. [Available from: https://beta.openai.com/docs/models.

2. Benoit JRA. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. medRxiv. 2023:2023.02.04.23285478.

3. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci. 2023;39(2):605-7.

4. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. 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(4).

5. Wang LK, Paidisetty PS, Cano AM. The next paradigm shift? ChatGPT, artificial intelligence, and medical education. Med Teach. 2023;45(8):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(6).

7. Li J, Dada A, Kleesiek J, Egger J. ChatGPT in Healthcare: A Taxonomy and Systematic Review (Preprint). medRxiv. 2023:2023.03. 30.23287899.

8. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. 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, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.
12. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, 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, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. Cureus. 2023;15(6):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(1):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(5):e230582.
16. Takagi S, Watari T, Erabi A, Sakaguchi K. 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, Jia J, Xu Y, Zhao J, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination (preprint): Research Square; 2023 [updated 2023. Available from: http://europepmc.org/abstract/PPR/PPR617702.
18. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, 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 from: 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 from: 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(4):410.
23. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of Artificial Intelligence in Community-Based Primary Health Care: Systematic Scoping Review and Critical Appraisal. J Med Internet Res. 2021;23(9):e29839.
24. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol. 2023.

REASONS FOR CONDUCTING THE SECOND STUDY

We recently published our first study about the performance of ChatGPT in the College of Family Physicians of Canada (CFPC) exam (85). In our first study, we employed GPT 3.5 and GPT4, one of the most studied and well-known LLM-based chatbots, to answer the Sample CFPC questions. We showed that the mean score for all five rounds was 76.0 ± 27.7 for GPT-3.5 and 85.2 ± 23.7 for GPT-4 according to the CFPC answer key. Additionally, we found that most of the answers were reasonable and acceptable, although they were not stated in the answer key. Our combined analysis of five rounds confirmed that GPT-4 performed significantly superior to GPT-3.5 (85). Although some other studies acknowledged the superior performance of GPT-4 compared to GPT-3.5 (86, 95, 97, 101, 105), our study was the first one challenging the complex task of open-ended Canadian family medicine questions (CFPC exam).

In the fifth round of this study, we used a zero-shot prompt strategy (no prompt before questions), and we observed that both GPT-3.5 and GPT-4 provided informative justifications and reasoning. When we compared Round 1 with a prompt to limit the answers to 10 words per line to Round 5, we showed that prompting did not influence the accuracy of responses. 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. Our study confirmed that GPT-3.5 and GPT-4 can be used to guess the answers to complex CFPC exam questions. The zero-shot responses could be highly instructive and seemed to be useful for better learning and preparation for this exam.

Our first study employed ChatGPT, an LLM trained with data up to September 2021 (65). It's important to consider that more recent advancements in LLMs might leverage more up-to-date information sources like Gemini (54). Furthermore, they may use different methods and training datasets. This potential difference highlights the need for further exploration to understand these models' evolving capabilities.

Secondly, ChatGPT utilizes a general-purpose LLM that is not specifically tailored for medical applications (33). With the advent of domain-specific LLMs like ChatDoctor (90), there has been a significant desire to develop models that are more effective in the medical field (11, 33, 38-40, 44, 80). However, it remains unclear whether these specialized LLMs would perform better than widely available models, such as ChatGPT, which were employed in our initial study.

Thirdly, several new systems have emerged, claiming to provide access to online searches (22). This capability could offer the advantage of accessing the most recent developments in a given field, including the latest medical management and diagnostics advancements (22).
Finally, it is not clear whether advanced state-of-the-art LLM-based chatbots that require payment subscriptions are superior to some available free systems.

Considering the above questions, we designed a second investigation. This new study aimed to comparatively evaluate several readily available LLM-based Chatbots with different characteristics for their capabilities in addressing open-ended questions within the SAMPs questionnaire format.

CHAPTER 4:

BODY OF THESIS; STUDY 2

Title: Assessment of Ten Different Large Language Model-Based Chatbots on Certification Examination of College of Family Physicians of Canada Examination Questionnaire

Authors:

Mehdi Mousavi M.D., CCFP1*,

Yves Bergevin MD, MSc, CCFP, FRCPC, FCFP2

Jason M. Harley Ph.D.3,4,5

Devin Ritter M.D., CCFP 6

Shabnam Shafee M.D., CCFP 7,

Samira Abbasgholizadeh Rahimi B.Eng, Ph.D 8,9,10.

1 Assistant Professor, Department of Family Medicine, Faculty of Medicine, University of Saskatchewan, SK, Canada. ORCID ID: https://orcid.org/0000-0003-3644-5741

2 Associate Professor, Department of Family Medicine, McGill University

3 Department of Surgery, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada. ORCID ID: https://orcid.org/0000-0002-2061-9519

4 Research Institute of the McGill University Health Centre, Montreal, QC, Canada

5 Institute for Health Sciences Education, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

6 Assistant Professor, Department of Family Medicine, Faculty of Medicine, University of Saskatchewan, SK, Canada

7 Family doctor, Riverside health complex, Prairie north health region, Saskatchewan Health Authority, Turtleford, SK, Canada

8 Mila Quebec AI-Institute, Montreal, QC, Canada

9 Department of Family Medicine, McGill University, Montreal, QC, Canada

10 Faculty of Dentistry Medicine and Oral Health Sciences, McGill University, Montreal, QC, Canada

ABSTRACT

Background: Large language models (LLMs) seem promising for enhancing medical education and aiding in preparation for medical exams. However, it is unclear which characteristics of LLMs are more helpful for their superior performance. This study aimed to evaluate the performance of different LLMs on an open-ended questionnaire designed to mimic the Canadian Certification Examination in Family Medicine (CFPC) exam.

Methods: During February and March 2024, we benchmarked 10 LLM models on 19 clinical scenarios comprising 77 questions from the CFPC website, requiring 165 answer lines in a zero-prompting approach. The chatbots examined were chosen based on their different capabilities as follows: (1) General Domain LLMs without online search, including ChatGPT-3.5 and ChatGPT-4 by Open AI (utilizing GPT-3.5 and GPT-4 respectively), Gemini and Gemini Advanced by Google utilizing Gemini Pro 1.0 and Gemini Ultra 1.0 respectively; (2) LLM-based chatbots that offer online search including Smart mode of YouChat (You-sm, utilizing GPT-3), GPT-4 mode of YouChat (You-4, using GPT-4), and Copilot by Microsoft (using GPT-4)), and Personal Intelligence (PI-AI, using Inflection 2.5 LLM) by Inflection AI; (3) LLMs trained on Medical Knowledge: i.e., DI doctor (AI-Dr) from Docus AI (using GPT-4), and Chat doctor (ChatDr) created by refining an LLM from meta-AI (using LLaMA). Three CFPC-certified reviewers with an average experience of 15 years scored the AI-generated responses and explanations according to the CFPC answer key (CFPC score). Following this, they conducted qualitative assessments based on their expert knowledge. The reviewers were initially blinded to each other's scores and evaluations. After scoring independently, they resolved any discrepancies in multiple follow-up meetings. To assess the agreement between three reviewers on CFPC scoring and post-discussion scoring, we applied Fleiss's Kappa. We used an independent Chi-square test to evaluate differences in correct responses among the LLMs.

Results: After joint discussions, the Kappa score was between 0.89 and 1, indicating perfect agreement. ChatGPT-4 achieved the highest accuracy rate among the ten systems at 85.5%, followed by AI-Dr at 82.4%, Gemini-A and ChatDr (each at 81.2%) and Copilot at 80%. LLMs with internet access, including You-sm, You-4, Copilot, PI-AI, AI-Dr, and ChatDr, did not perform significantly better than ChatGPT-4. Our two medical domain systems had acceptable performance close to the best general domain models.

Conclusion: The results of our study were promising in employing several General domain and medical domain LLMs to answer the open-ended sample CFPC questionnaire. GPT-4 showed the highest accuracy. More studies might be needed to explore the strengths or weaknesses of different LLMs in responding to the CFPC exam questions and exam preparation. Particularly, more efforts should be made to develop more efficient medical domain LLMs.

BACKGROUND

Artificial intelligence (AI) and machine learning (ML) are advancing rapidly and are increasingly used in healthcare practice (1). Large language models (LLMs) represent AI systems capable of learning millions of words (2) and performing tasks such as summarizing, translating, and generating text, along with answering questions similar to humans (3).

LLMs have also shown promise in medical examinations and exam preparation (4). For instance, researchers from the University of Toronto have employed ChatGPT to respond to an official multiple-choice progress test intended for the Canadian family medicine exam and showed that ChatGPT-3.5 using Generative Pre-trained Transformer (GPT) 3.5 demonstrated comparability to family medicine residents, while GPT-4 outperformed both residents and GPT-3.5 (5). Our most recent study on GPT-3.5 and GPT-4 found that both iterations of ChatGPT using GPT-3.5 and GPT-4 could answer open-ended sample questions from the College of Family Physicians of Canada (CFPC) certification examination with over 70% accuracy (6). In this study, GPT-4 consistently provided more accurate responses than GPT-3.5 (6). ChatGPT has also been used to answer the different steps of the United States Medical Licensing Examination (e.g. USMLE (7-9), and membership of the Royal College of General Practitioners Applied Knowledge Test (AKT) (10), and in other languages, including Japanese (11), Chinese (12) and Persian (13) exams.

Several other LLMs are rapidly emerging, aiming to boost their efficiency through server upgrades or training on larger-size parameters, potentially surpassing GPT-based LLMs. An example is Gemini (14) by Google, recently launched but not widely studied yet.

Secondly, GPT, being a closed-source LLM, limits wider accessibility and customization in specialized areas such as medicine (15). Considerable efforts have been made to train LLMs with medical knowledge to accurately respond to general medical queries. One example of medical domain LLMs is Google's MedPaLM, which has exceeded a “passing” score in USMLE-style questions with a score of 67.2%. Google has fine-tuned PaLM 2 with medical data to create a more practical version, Med-PaLM 2, which achieved up to 86.5% on the same USMLE preparation questionnaire (16). Several other LLMs have been created to handle medical inquiries more efficiently (17, 18). Many attempts have been made to customize and train open-source LLMs to achieve better performance in responding to medical benchmarks (15, 19-21).

Thirdly, ChatGPT lacks access to online information and its training is limited to a fixed point in the past. To address these potential disadvantages, various LLMs have been designed to access online resources for the latest information (22).

It is important to acknowledge that while LLMs offer numerous potential advantages in medical education, exam preparation, and research, there are ethical concerns about their application in these domains (4, 23). For example, the International Committee of Medical Journal Editors (ICMJE) (24) and the World Association of Medical Editors (WAME) (25) have suggested important roles regarding the implication of LLMs in developing manuscripts.

Board exams are still the gold standard for evaluating medical education programs. The College of Family Physicians of Canada (CFPC) conducts a comprehensive certification exam to evaluate the clinical proficiency of Family Medicine practitioners in Canada. This exam consists of both an oral component and a written component. The written section, known as Short-Answer Management Problems (SAMPs), is a four-hour computer-based assessment. It typically presents approximately 40 clinical scenarios, each followed by questions pertaining to aspects such as patient history, diagnosis, and the management of diverse clinical conditions (26). Each year, numerous candidates invest a significant amount of money in test preparation materials and courses. The potential benefits of LLMs, such as lower cost, 24-hour accessibility, and the provision of direct answers to every question, make them significant alternatives to aid candidates in exam preparation. While several general-domain (14, 27) and LLMs trained on medical knowledge (15, 17, 19, 20, 28-30) are available either for free or via subscription, it remains uncertain whether those necessitating payment offer superior performance. Additionally, it remains unclear whether other features of LLMs, such as specialized medical domain training or access to online resources for answering questions, could be more effective in addressing complex questions from the CFPC exam. Therefore, we conducted this study to compare accessible LLMs with various potential advantages in answering open-ended sample SAMPs questionnaires.

METHODS

LLMs

In this study, we evaluated 10 LLM-based chatbots using different LLMs. Table 1 summarizes the features of the ten chatbots we studied.

Table 1. Overview of LLM-based chatbot characteristics in the study.

Name	Developer Website	Version	LLM	Access date d/m/year	Access to online resources	Fee
ChatGPT-3.5	OpenAI chat.openai.com	01-2022,	GPT-3.5	03/02/2024	No	Free access
ChatGPT-4	OpenAI chat.openai.com	01-2022	GPT-4	04/02/2024	No	$20/m
Gemini	Google gemini.google.com/app	Gemini Pro	Gemini Pro 1.0	12/02/2024	No	Free access
Gemini-A	Google gemini.google.com/app	Gemini Ultra	Gemini Ultra 1.0	14/02/2024	No	2month free, then $19.99
You-sm	YouChat you.com	1.0.9; 09-2023	GPT-3	04/02/2024	Yes	Free access
You-4	YouChat you.com	1.0.9; 09-2023	GPT-4	04/02/2024	Yes	$9.99/m
Copilot	Microsoft copilot.microsoft.com	09-2023	GPT-4 & turbo	03/02/2024	Yes	Free access
PI-AI	Inflection AI pi.ai/talk	Inflection-2.5	Inflection-2.5	14/02/2024	Yes	Free access
AI-Dr	Docus AI docus.ai/ai-health-assistant	Early 2023	GPT-4, others*	14/02/2024	Yes	3 Q/m free, then $14.99
ChatDr	Li Y et al. app.chatdoctor.com	0.5; 2024	Modified LLaMA	23/02/2024	Yes	Free access

Introduction of LLM-based Chatbots

1) General Domain LLMs Without Access To Online Search:

- ChatGPT-3.5, known as ChatGPT, released in November 2022 by OpenAI in San Francisco, CA, USA, utilizes GPT-3.5 LLM (31). ChatGPT has been used to answer several medical exams like USMLE (7-9) and the Canadian Family Medicine exam (5, 6).

- ChatGPT Plus uses GPT-4, another LLM from Open AI. It is purposed to be more accurate and efficient (32) and less prone to producing "hallucinations," which are incorrect responses provided by the chatbot so convincingly that people usually do not question their accuracy (33, 34).

- Gemini, launched by Google (New York, NY, USA) in February 2024, is an updated version of Bard. It operates on the Gemini Pro 1.0 model LLM, which is distinct from GPT and supports over 40 languages (35).

- Gemini Advanced (Gemini-A) employs an advanced LLM model known as Ultra 1.0 Gemini. Google introduced Gemini Advanced concurrently with Gemini in February 2024 (35). Gemini Pro 1.5 has recently been released to Gemini Advanced subscribers, featuring significant enhancements, including better image comprehension, an extended context window, and new data analysis capabilities (36).

2) General Domain LLMs With Access To Online Search:

- YouChat Smart Mode (You-sm), offered by You.com (Minneapolis, MN, USA), claims several advantages over ChatGPT, such as staying updated, improving based on user feedback, and occasionally providing references to answers (22). The version we utilized in our study was 1.0.9, developed in September 2023, which utilizes GPT-3. Following our study, You.com introduced a newer version of You-sm that is capable of utilizing GPT-4 for free or a variety of other LLMs, including GPT-4 Turbo, Claude Instant, Claude 2, Claude 3 Opus, Claude 3 Sonnet, Gemini Pro, DBRX-Instruct, and Zephyr (Uncensored), available for a subscription fee (37).

- YouChat GPT-4 mode (You-4) is another system available from You.com (22) that uses GPT4, the more advance LLM by Open AI.

- Copilot (Microsoft, Redmond, Washington, USA), a rebrand of formed Bing Chat, employs OpenAI's GPT-4 and may utilize GPT-4 Turbo, a more advanced product of OpenAI, during non-peak times (38).

- Personal Intelligence (PI-AI) by Inflection AI (Palo Alto, CA) (39), is a new chatbot using its unique LLM named Inflection 2.5. PI-AI serves as a personal AI companion, learning the user's interests and preferences to provide personalized responses over time (40).

3) LLMs trained On Medical Knowledge:

- AI doctor (AI-Dr) (Docus AI, Wilmington, DE USA) is a health assistant powered by GPT-4. It also uses OpenAI Text Embeddings and Vector DBS. AI-Dr has undergone extensive training with abundant medical data to become a helpful resource for health-related inquiries. Additionally, AI-Dr can facilitate user access to consultations with real doctors worldwide, requiring the necessary consultation fee (30).

- Chat Doctor (ChatDr) has created a specialized LLM by refining and enhancing the open-access LLM of large language model meta-AI (LLaMA). ChatDr employed a large medical dataset comprising 100,000 conversations between patients and doctors to enhance the accuracy of medical advice. Additionally, the model can access current information from online sources like Wikipedia and data from offline medical databases (17).

The classification of our LLM-based chatbots is shown in Figure 1.

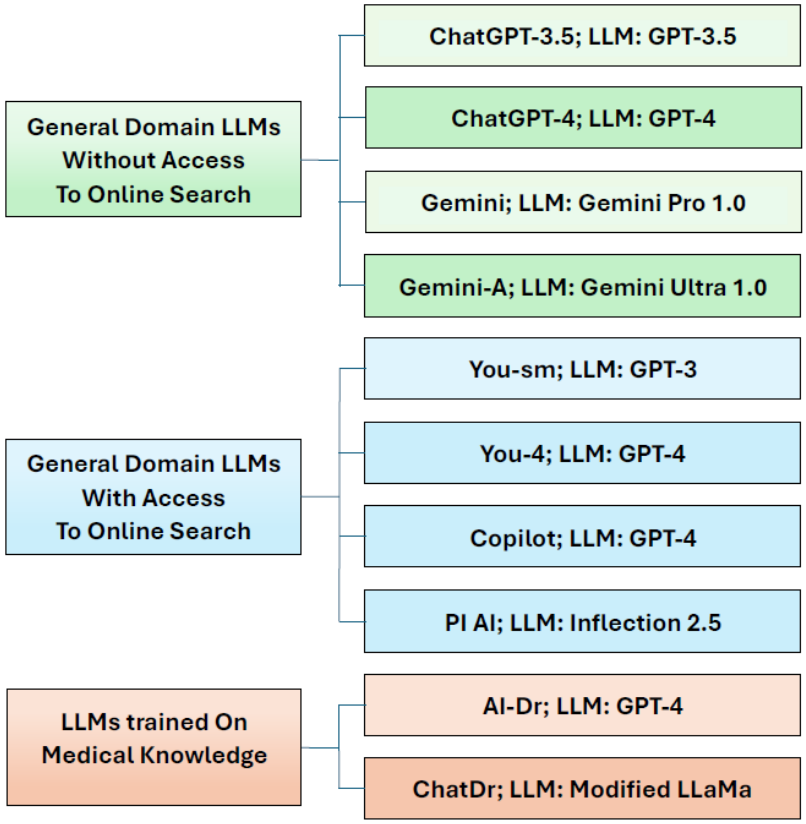

Figure 1. Classification of LLM-based chatbots based on four criteria.

Four LLM-based chatbots that use general domain LLMs without access to online search are shown in green. The next four LLM-based chatbots that use general domain LLMs with access to online search are shown in blue. Finally, the two LLMs trained in medical knowledge are shown in beige.

Study Design

The study was ***quasi-experimental*** due to following reasons:

No randomization: We did not do randomization. In our study, the chatbots were selected based on their characteristics (e.g., general domain LLMs, medical domain LLMs, and access to online search capabilities), not through randomization.

Controlled Comparison: We systematically evaluated ten LLM-based chatbots on the same standardized CFPC questionnaire, ensuring that the comparison was consistent across models. This control over the intervention and measurement aligns with quasi-experimental design principles.

Pre- and Post-Scoring Collaboration: Such structured evaluation methods enhance the rigor of quasi-experimental designs.

Lack of True Experimental Manipulation: While the study manipulated chatbot types, it did not involve experimental manipulation in the traditional sense.

Real-World Context: The study focused on the application of existing LLMs in a real-world task (answering medical exam questions), another hallmark of quasi-experimental research.

Questionnaire

For this study, we utilized a standard sample SAMPs questionnaire obtained from the official CFPC website (26). This sample comprises 19 clinical scenarios, each accompanied by 2 to 5 questions covering various areas in family medicine. The clinical scenarios covered a range of domains within family medicine as follows:

Cardiology featured three questions concerning atrial fibrillation, dizziness, and hypertension; endocrinology included one question on osteoporosis; emergency medicine encompassed two questions about lacerations and poisoning. There were three gastroenterology questions related to abdominal pain, abnormal liver function tests, and dyspepsia; musculoskeletal disease was represented with one question on low back pain; neurology included one question about seizures; gynecology had two questions concerning a breast lump and fertility. Pediatrics and infectious disease were combined in one question about fever in a newborn; psychiatry was addressed with two questions about depression and eating disorders. Additionally, there was one question in respirology about chronic obstructive pulmonary disease, one in urology about prostate issues, and one question regarding fatigue. There were 77 questions in total, necessitating 165 lines of response. In Case 19, we eliminated an image from the survey because some of our chatbots couldn’t process images.

Prompting, Running Chatbots and Storing the Responses

We employed Zero-Shot prompting in this study except for Gemini and Gemini -A. In this approach, we present a question to the model without any prior examples or context. This approach could allow us to obtain detailed responses for evaluation by reviewers to assess the level of assistance in preparing for the CFPC exam and medical education. Most LLM-based chatbots provided satisfactory answers with this Zero-Shot approach. However, Gemini and Gemini-A declined to answer most questions without a prompt, citing their lack of medical expertise. Therefore, we prompted Gemini and Gemini-A before each question with the following statement: “I am preparing for an exam, and I want to learn. Please answer my questions as much as you can.”

We adopted an approach similar to an actual examination to feed the chatbots with questions. For each of the 19 clinical scenarios, we first presented the chatbots with the clinical scenario, followed by the initial question. We transferred the chatbot’s responses to a Word document for further scoring and evaluation.

Scoring and review of responses

Three experienced CFPC-certified practicing family physicians (M.M., D.R., and S.S.) with an average of 15 years of medical practice independently reviewed and scored the AI-generated responses and explanations blinded to each other.

Initially, the reviewers strictly followed the answer key provided on the CFPC website for scoring, referred to as "CFPC Scoring." Completely incorrect responses received a score of zero per line during the evaluation, while those deemed accurate were assigned a score of one per line. However, some chatbot answers appeared accurate and acceptable to the reviewers but did not align with the CFPC answer key. Consequently, the reviewers provided a secondary score, termed "Reviewer's Score," based on their expertise and a validated, well-known, online accessible medicine resource “UpToDate” (41).

After independently scoring the responses, the three reviewers met to address any evaluation discrepancies. They recognized that many differences in scoring arose from varying scoring strategies. Consequently, they agreed to adopt a unified approach to scoring responses, incorporating the following guidelines:

- If multiple answers are provided in a single line of chatbot responses, all of them will be accepted as correct.
- If a question asks a specific number of lines for answers and the chatbot provides more, only the first lines of answers will be considered for the CFPC score. However, if the answers are not provided in the initially requested lines but are given in subsequent additional lines, they will count toward the Reviewer's Score.

Following multiple meetings and collaborative discussions, most discrepancies were resolved, and the reviewers reached a consensus. In cases where consensus could not be reached, and two of the reviewers agreed, their combined score would be chosen over the third reviewer's score.

Qualitative assessment

The three reviewers performed a qualitative analysis of the responses to the questions, which were designed according to our expert team-designed criteria and definitions. They used a standardized questionnaire for their unique approach and answered "yes" or "no" based on a detailed review of responses provided by the systems. Initially, each reviewer assessed the responses independently, followed by a series of joint meetings to address any discrepancies. A similar approach for human evaluations for helpfulness and safety has previously been described by Touvron and his associates (42).

Throughout the evaluation, the reviewers analyzed 165 sets of chatbot responses according to the following specified criteria:

- Irrelevant response: These included answers that were not deemed related to the current topic or situation being discussed and, therefore, considered unimportant or off-topic.
- Unsafe response: These responses were considered misleading or potentially harmful if applied in real medical practice.
- Extra useful information: When questions were posed without any prompt, many chatbots offered additional useful information that enhanced the user's understanding of the concept. We considered such valuable insights. However, if the chatbot merely used extra words to express the answer without providing additional information or advising consultation with professionals, it was not considered extra useful. This extra useful information was mostly about the rationale of choosing the answer.
- Unintelligent response: Where the responses lack thought, consideration, or intelligence, and the chatbot was not able to comprehend the context of the question or some part of it.
- List of responses more than requested: While most chatbots accurately determined the number of answers required by the question, they provided excessive lines of response in some instances. For example, when the question asked for three answers, and the chatbot gave four

answers, this was considered “List of responses more than requested. An excessive line of response did not necessarily provide extra helpful information.

Data Analysis

We conducted our statistical analyses using IBM SPSS Statistics. Categorical variables were presented as numbers (percentages). Differences in the correct responses provided by the chatbots were assessed using an independent Chi-square test. To evaluate inter-rater agreement between the three reviewers for CFPC scoring and Reviewer scoring after resolving conflicts through discussion, we used Fleiss's Kappa (43). We calculated Cohen's kappa to evaluate the agreement between the CFPC and Reviewers’ scoring. We considered kappa values above 0.80 as indicating perfect agreement, 0.61-0.80 as substantial agreement, 0.41-0.60 as moderate agreement, 0.21-0.40 as fair agreement, 0.001-0.20 as slight agreement, and less than 0 as poor agreement (43). All reported P-values were two-sided, and statistical significance was defined as $P \leq 0.05$.

Ethical Considerations

This research solely examined data publicly accessible through the Internet and did not involve any human participants. Therefore, the study was exempted from the approval from the McGill University's review board. The authors declare no conflicts of interest.

RESULTS

We assessed 19 clinical situations, each accompanied by 2 to 7 relevant questions, resulting in a total of 77 specific questions and generating 165 lines of responses. The length of possible responses to each question varied, ranging from 1 to 5 lines, with a median of 2 lines. Fourteen out of fifteen chatbots provided detailed responses to the questions. However, FM1-C failed to answer, stating it had not been trained yet. Table 2 displays the agreement rate among the reviewers. After discussion, Fleiss's Kappa was calculated to assess agreement for CFPC scoring and reviewers' scoring.

Table2. The percentage of inter-rater agreement among three reviewers for 165 lines of answers. Data presented as numbers (%). Fleiss Kappa was calculated for the agreement of scoring after discussion and the 95% confidence interval was calculated "Fleiss Kappa (95% CI)".

	CFPC score Agreement before discussion	CFPC score Agreement after discussion	Fleiss Kappa for agreement of CFPC scoring after discussion	Reviewer score Agreement before discussion	Reviewer score Agreement after discussion	Fleiss Kpapa for agreement of Reviewers' scoring after discussion
ChatGPT-3.5	90.3%	100%	1.00 (95% CI: 0.91-1.09)	81.2%	97.6%	0.93 (95% CI: 0.84-1.02)
ChatGPT-4	92.7%	99.4%	0.98 (95% CI: 090-1.07)	88.5%	98.2%	0.89 (95% CI: 0.80-0.98)
Gemini	85.5%	98.8%	0.98 (95% CI: 0.89-1.07)	80.6%	97.6%	0.94 (95%CI: 0.85-1.03)
Gemini-A	89.7%	99.4%	0.98 (95% CI: 090-1.07)	87.3%	97.6%	0.91 (95%CI: 0.86-1.03)
You-sm	91.5%	100%	1.00 (95% CI: 0.91-1.09)	74.5%	95.8%	0.92 (95% CI: 0.83-1.00)
You-4	89.7%	100%	1.00 (95% CI: 0.91-1.09)	76.4%	97.6%	0.93 (95%CI: 0.84-1.01)
Copilot	88.5%	99.4%	0.98 (95% CI: 0.90-1.07)	81.2%	98.2%	0.92 (95%CI: 0.83-1.01)
PI-AI	87.9%	100%	1.00 (95% CI: 0.91-1.09)	77.6%	99.4%	0.98 (95% CI: 0.89-1.07)
AI-Dr	93.9%	99.4%	0.98 (95% CI: 090-1.07)	85.5%	99.4%	0.98 (95% CI: 0.89-1.06)
ChatDr	98.2%	100%	1.00 (95% CI: 0.91-1.09)	93.3%	98.8%	0.92 (95% CI: 0.82-1.01)

Four LLM-based chatbots that use general domain LLMs without access to online search are shown in green. The next four LLM-based chatbots that use general domain LLMs with access to online search are shown in blue. Finally, the two LLMs trained in medical knowledge are shown in beige. AI: Artificial intelligence; AI-Dr: AI Doctor; ChatDr: ChatDoctor; ChatGPT-4: ChatGPT Plus; Gemini-A: Gemini Advanced; GTP: Generative Pre-training Transformer; You-sm: YouChat, smart mode; You-4: YouChat, GPT-4 mode

Figure 2 summarizes and compare the CFPC scores provided by 10 LLM-based chatbots based on their responses to the questionnaire. ChatGPT-4 demonstrated the highest accuracy, with 141 out of 165 correct lines of answers, representing an 85.5% accuracy rate. It was followed by AI-Dr with 136 correct answers (82.4%) and Gemini-A along with ChatDr with 134 correct answers (81.2%). Compared with other free chatbots, ChatDr led with an impressive 81.2% performance rate. Copilot closely followed with an 80% success rate, and ChatGPT-3.5 also showed strong performance (76.4%).

Figure 3 present the reviewers' evaluations and comparisons of scores from 10 LLM chatbots. Similar to the CFPC scores, ChatGPT-4 achieved the highest reviewers' score with 155 correct answers (93.9%).

The statistical hypothesis tests in Tables 3 and 4 were employed to determine whether the observed differences in chatbot performance were statistically significant. Each chatbot's responses to 165 lines of questions were evaluated, and the results were aggregated to compare accuracy rates. Tables 3 summarizes the performance metrics for each chatbot based on CFPC scores while Table 4 present the same incite for reviewers' scores. For example, Table 3 shows that ChatGPT-4's accuracy rate of 85.5% significantly outperformed Gemini's 69.1% ($P < 0.001$), indicating consistent superiority in this task. However, differences between ChatGPT-4 and AI-Dr (82.4%, $P = 0.453$) were not statistically significant, suggesting comparable performance levels.

The Cohen’s kappa for agreement between CFPC scores and reviewers' scores varied across the LLM chatbots as follows: 0.66 for ChatGPT-3.5, 0.68 for You-sm, 0.55 for ChatGPT-4, 0.69 for You-4, 0.54 for Copilot, 0.65 for Gemini, 0.61 for Gemini-A, 0.46 for PI-AI, 0.64 for AI-Dr, 0.74 for ChatDr. These kappa values indicate agreements ranging from moderate (0.41-0.60) to substantial (0.61-0.80).

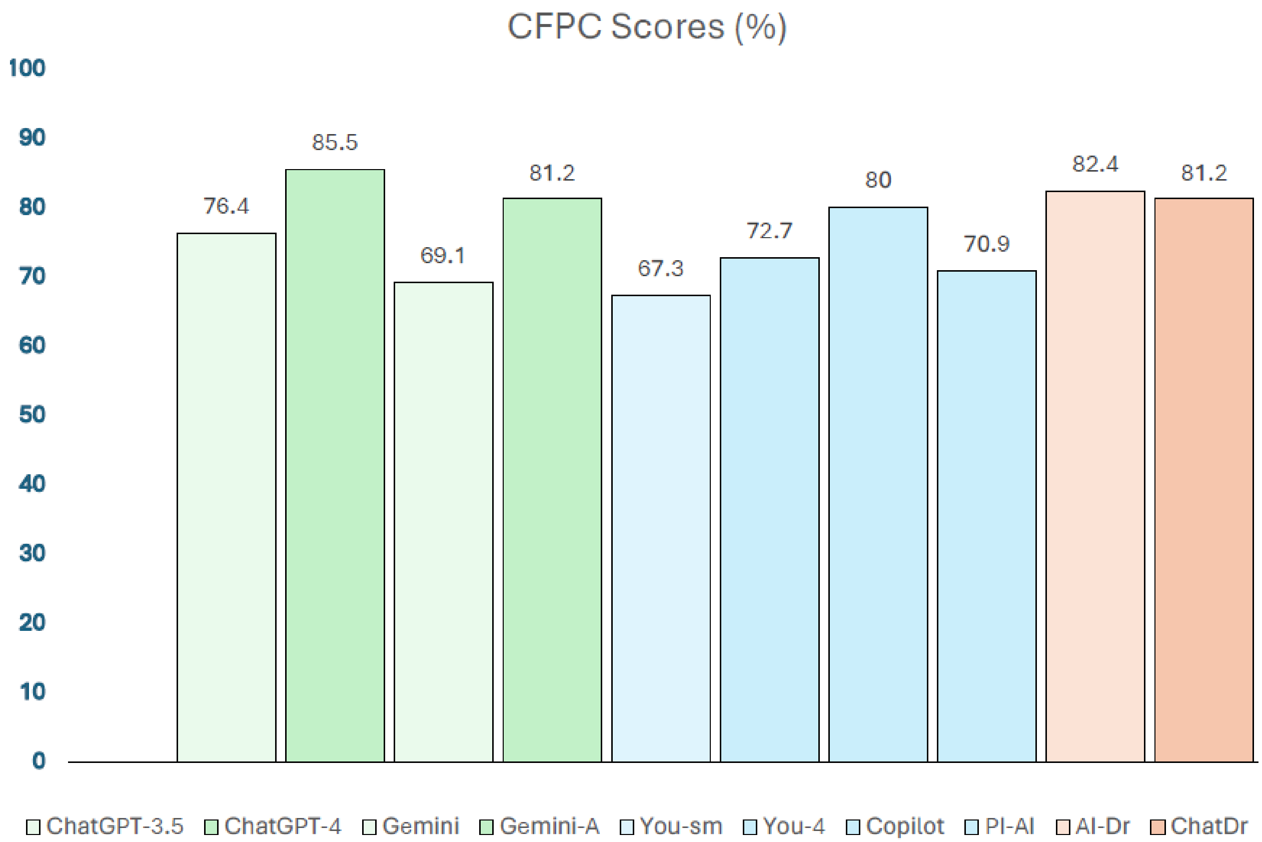

Figure 2. Comparison of the percentage of correct responses by each LLM-based chatbot on the CCFP answer key.

Four LLM-based chatbots that use general domain LLMs without access to online search are shown in green. The next four LLM-based chatbots that use general domain LLMs with access to online search are shown in blue. Finally, the two LLMs trained in medical knowledge are shown in beige. AI: Artificial intelligence; AI-Dr: AI Doctor; ChatDr: Chat Doctor; ChatGPT-4: ChatGPT Plus; Gemini-A: Gemini Advanced; GTP: Generative Pre-training Transformer; GPT-3.5: ChatGPT-3.5; You-sm: YouChat, smart mode; You-4: YouChat, GPT-4 mode

Table 3. An Analysis of 165 Answer Lines: The P values from chi-square tests comparing the pairs of LLMs. P<0.05 is presented in bold font.

		GPT-4 141 (85.5)	**Gemini** 114 (69.1)	**Gemini-A** 134 (81.2)	**You- sm** 111 (67.3)	**You-4** 120 (72.7)	**Copilot** 132 (80.0)	**PI-AI** 117 (70.9)	**AI-Dr** 136 (82.4)	**ChatDr** 134 (81.2)
GPT-3.5	126 (76.4)	**0.036**	0.138	0.281	0.066	0.448	0.424	0.261	0.174	0.281
GPT-4	141 (85.5)		**<.001**	0.301	**<0.001**	**0.004**	0.190	**0.001**	0.453	0.301
Gemini	114 (69.1	**0.001**		**0.011**	0.723	0.476	**0.023**	0.719	**0.005**	**0.011**
Gemini-A	134 (81.2)	0.301	**0.011**		**0.004**	0.067	0.781	**0.028**	0.775	0.999
You-sm	111 (67.3)	**<.001**	0.723	**0.004**		0.280	**0.009**	0.475	**0.002**	**0.004**
You-4	120 (72.7)	**0.004**	0.476	0.067	0.280		0.120	0.714	**0.035**	0.067
Copilot	132 (80.0)	0.190	**0.023**	0.781	**0.009**	0.120		0.055	0.573	0.781
PI-AI	117 (70.9)	**0.001**	0.719	**0.028**	0.475	0.714	0.055		**0.013**	**0.028**
AI-Dr	136 (82.4)	0.453	**0.005**	0.775	**0.002**	**0.035**	0.573	**0.013**		0.775
ChatDr	134 (81.2)	0.301	**0.011**	0.999	**0.004**	0.067	0.781	**0.028**	0.775	

Four LLM-based chatbots that use general domain LLMs without access to online search are shown in green. The next four LLM-based chatbots that use general domain LLMs with access to online search are shown in blue. Finally, the two LLMs trained in medical knowledge are shown in beige. The AI: Artificial intelligence; AI-Dr: AI Doctor; ChatDr: Chat Doctor; GPT-4: ChatGPT Plus; Gemini-A: Gemini Advanced; GTP: Generative Pre-training Transformer; GPT-3.5: ChatGPT-3.5; You-sm: YouChat, smart mode; You-4: YouChat, GPT-4 mode

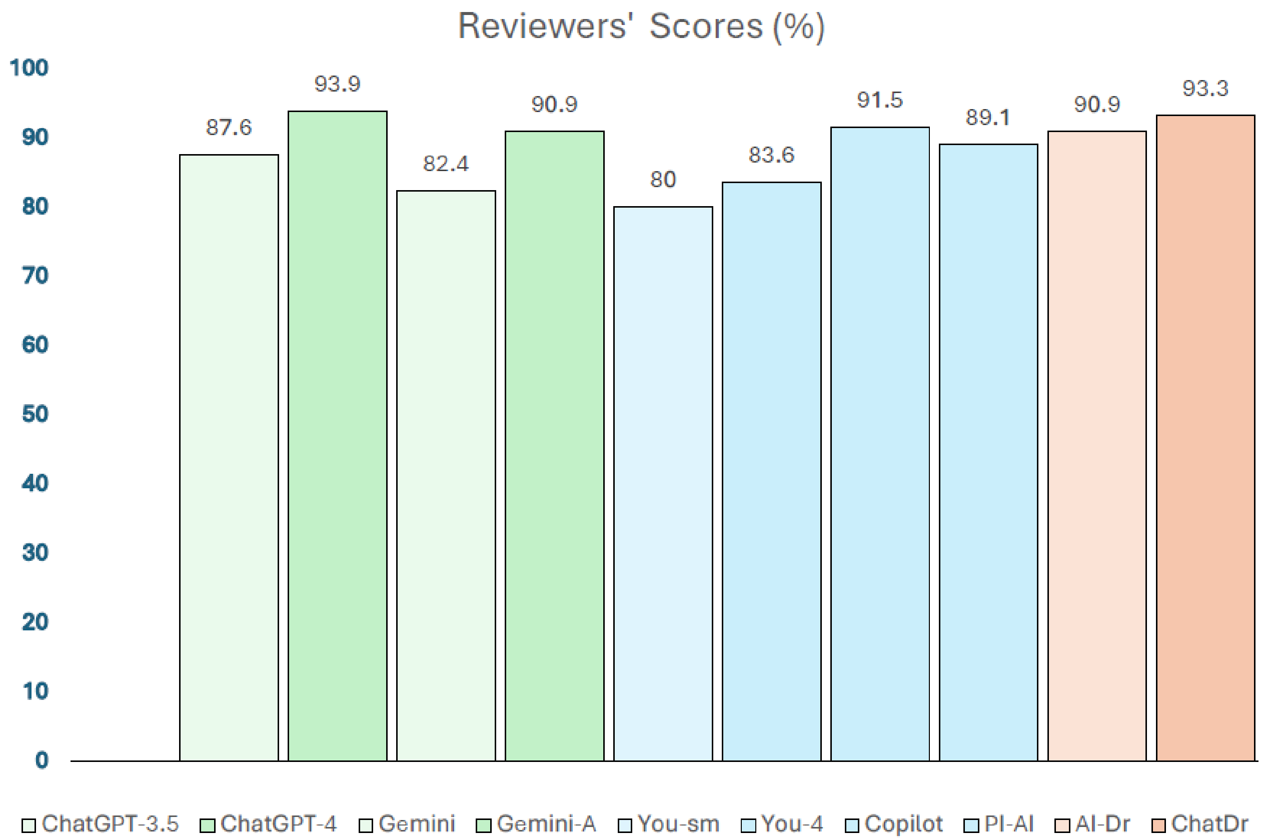

Figure 3. Comparison of Reviewers' Scores percentage of correct responses by each LLM-based chatbot.

Four LLM-based chatbots that use general domain LLMs without access to online search are shown in green. The next four LLM-based chatbots that use general domain LLMs with access to online search are shown in blue. Finally, the two LLMs trained in medical knowledge are shown in beige. AI: Artificial intelligence; AI-Dr: AI Doctor; ChatDr: Chat Doctor; ChatGPT-4: ChatGPT Plus; Gemini-A: Gemini Advanced; GTP: Generative Pre-training Transformer; GPT-3.5: ChatGPT-3.5; You-sm: YouChat, smart mode; You-4: YouChat, GPT-4 mode

Table 4. An Analysis of 165 Answer Lines by 10 LLM-based Chatbots. The P values from chi-square tests comparing the pairs of LLMs. P<0.05 is presented in bold font.

		GPT-4	**Gemini**	**Gemini-A**	**You- sm**	**You-4**	**Copilot**	**PI-AI**	**AI-Dr**	**Chat Dr**
		155 (93.9)	136 (82.4)	150 (90.9)	132 (80.0)	138 (83.6)	151 (91.5)	147 (89.1)	150 (90.9)	154 (93.3)
GPT-3.5	143 (86.7)	**0.026**	0.286	0.222	0.104	0.439	0.158	0.500	0.222	**0.044**
GPT-4	155 (93.9)		**0.001**	0.298	**<.001**	**0.003**	0.396	0.114	0.298	0.822
Gemini	136 (82.4)	**0.001**		**0.023**	**0.005**	0.769	**0.014**	0.083	**0.023**	**0.002**
Gemini-A	150 (90.9)	0.295	**0.023**		**0.005**	**0.047**	0.846	0.582	0.999	0.414
You-sm	132 (80.0)	**<.001**	0.573	**0.005**		0.392	**0.003**	**0.022**	**0.005**	**<.001**
You-4	138 (83.6)	**0.003**	0.769	**0.047**	0.392		**0.030**	0.149	**0.047**	**0.006**
Copilot	151 (91.5)	0.396	**0.014**	0.846	**0.003**	**0.030**		0.457	0.846	0.533
PI-AI	147 (89.1)	0.114	0.083	0.582	**0.022**	0.149	0.457		0.582	0.173
AI-Dr	150 (90.9)	0.295	**0.023**	0.999	**0.005**	0.047	0.846	0.582		0.414
ChatDr	154 (93.3)	0.822	**0.002**	0.414	**<.001**	0.006	0.533	0.173	0.414	

Four LLM-based chatbots that use general domain LLMs without access to online search are shown in green. The next four LLM-based chatbots that use general domain LLMs with access to online search are shown in blue. Finally, the two LLMs trained in medical knowledge are shown in beige. AI: Artificial intelligence; AI-Dr: AI Doctor; ChatDr: Chat Doctor; GPT-4: ChatGPT Plus; Gemini-A: Gemini Advanced; GTP: Generative Pre-training Transformer; GPT-3.5: GPT-3.5; You-sm: YouChat, smart mode; You-4: YouChat, GPT-4 mode

Table 5 presents the qualitative assessment results of the responses from the 10 LLM-based chatbots studied. You-sm provided the highest rate of unsafe answers, 9.1%. Conversely, ChatGPT-4 and AI-Dr provided the safest answers equally. You-sm had the highest rate of unintelligent answers, 4.8%. All of our study systems offered many extra useful explanations in their responses, totalling more than 137 cases (83%).

Additional responses offered after the list of answers was completed: This phenomenon occurred mainly with responses from Gemini, Gemini Advanced, and Pi-AI. In these cases, the chatbot provided a correct number of answers to the question but added supplementary information after completing the initial list. Supplement 1 provides examples of qualitative assessment of the responses. We attached Supplement 2, which delivers instances of unintelligent answers provided by the chatbots.

Table 5. Qualitative assessment of the responses given by 10 Chatbots in the study

Qualitative evaluation criteria ----------- Different LLM based chatbots	No Answer to the Question. N=165	Irrelevant Responses N=165	Unsafe Responses N= 165	Unintelligent Responses N=165	Extra Useful Information N=165	Reference provided	Lines of Responses More than Requested n=77
ChatGPT	1 (0.6)	0	3 (1.8)	5 (3.0)	160 (97.0)	0	1 (1.3)
ChatGPT-4	0	0	2 (1.2)	2 (1.2)	163 (98.8)	0	1 (1.3)
Gemini	3 (1.8)	5 (3.0)	11 (6.7)	6 (3.6)	151 (91.5)	4 (5.2)	11 (14.3)
Gemini-A	3 (1.8)	0	3 (1.8)	4 (2.4)	158 (95.8)	0	1 (1.3)
You-sm	0	5 (3.0)	15 (9.1)	8 (4.8)	148 (89.7)	17 (22.1)	1 (1.3)
You-4	0	4 (2.4)	9 (5.5)	5 (3.0)	153 (92.7)	13 (16.9)	0
Copilot	0	3 (1.8)	5 (3.0)	5 (3.0)	149 (90.3)	77 (100)	4 (5.2)
PI-AI	0	2 (1.2)	9 (5.5)	3 (1.8)	154 (93.3)	0	1 (1.3)
AI-Dr	0	1 (0.6)	2 (1.2)	4 (2.4)	159 (96.4)	0	0
ChatDr	0	0	3 (1.8)	3 (1.8)	141 (85.5)	0	0

Four LLM-based chatbots that use general domain LLMs without access to online search are shown in green. The next four LLM-based chatbots that use general domain LLMs with access to online search are shown in blue. Finally, the two LLMs trained in medical knowledge are shown in beige. AI: Artificial intelligence; AI-Dr: AI Doctor; ChatDr: ChatDoctor; ChatGPT-4: ChatGPT Plus; Gemini-A: Gemini Advanced; GTP: Generative Pre-training Transformer; You-sm: YouChat, smart mode; You-4: YouChat, GPT-4 mode

Gemini, PI-AI, Gemini-A, and occasionally You-sm offered additional answers after addressing the required lines in the question stem. Among these additional responses, Gemini provided 12 out of 30 additional answers that contained the answers from the CFPC answer key. Gemini-A provided 4 out of 10, PI-AI provided 13 out of 36, and You-sm provided 1 out of 1 correct answer in these additional explanations.

Copilot consistently responded to all questions, “Hello, this is Copilot. I am here to help you with your question.” Meanwhile, Gemini frequently reiterated, “While I cannot provide medical advice, I can share some general information based on the scenario you provided. Remember, this information should not be a substitute for professional medical advice and diagnosis. Always consult a qualified healthcare professional for medical concerns.” This paragraph, or a similar one, was repeated in 47 instances, accounting for 61% of cases.

It is interesting to note the answers to two questions about treating mild (MRC2) COPD. The answer to this question differed using the Canadian Thoracic Society (CTS 2023) guidelines compared to the previous 2019 guidelines. While the CFPC still provided answers based on the previous guidelines, ChatGPT, ChatGPT-4, You-sm, You-4, Copilot, AI-Dr, and ChatDr responded in accordance with the 2019 guidelines, aligning with the CFPC answer key. However, Gemini, Gemini-A, and PI-AI answered based on the latest 2023 guidelines.

DISCUSSION

Overall Findings

In this study, we examined 10 LLM-based chatbots with varying capabilities to answer a series of open-ended standard samples of the SAMPs exam. We evaluated the accuracy of their responses by calculating the ratio of correct answers according to the CFPC answer key. Additionally, three experienced CFPC-certified family physician reviewers reassessed the responses for accuracy based on their knowledge and available resources. This re-evaluation, which we termed the reviewer's score, helped us better identify completely incorrect answers.

After a joint discussion, we found that the scores assigned to the chatbots exhibited perfect agreement (43) with Kappa ≥ 0.89. Furthermore, the agreement between CFPC scores and reviewers' scores was substantial (Kappa = 0.61-0.80) except for ChatGPT-4, Copilot and PI-AI, which was categorized as moderate agreement (Kappa = 0.41-0.60).

ChatGPT-4 achieved the highest CFPC score with 141 correct answers, representing an 85.5% accuracy rate. It achieved a higher CFPS score than ChatGPT-3.5, You-sm, You-4, Gemini, PI-AI ($P<0.05$) and comparable to Copilot, Gemini-A, AI-Dr, and ChatDr. Our reviewers' scoring showed similar trends, with ChatGPT-4 having the highest number of correct answers at 155 (93.9%) among all chatbots.

ChatGPT-4 has demonstrated outstanding performance in medical exams across various studies, highlighting its potential in medical education and exam preparation (4). In one study, ChatGPT-4 outperformed ChatGPT-3.5 and the University of Toronto's family medicine residents on an official multiple-choice progress test for CFPC exam preparation. In the aforementioned study, ChatGPT-4's performance was close to our study at 82.4%, while GPT-3.5 achieved 57.4%

accuracy, which is lower than the performance of ChatGPT-3.5 in our research (5). Another study similarly confirmed 86.1% scoring by ChatGPT-4 on the MedQA (USMLE-style questions with four choices) questions (29). Additionally, ChatGPT-4 significantly outperformed ChatGPT-3.5 and other LLMs like LLaMA, Alpaca, and Vicuna in the Chinese National Medical Licensing Examination (CMExam) (28). In a study, ChatGPT-4 achieved a passing score on the USMLE by more than 20 points (9). ChatGPT-3.5, in contrast, scored approximately 60% for Step 1, Step 2CK, and Step 3 of the USMLE, which was close to the passing score (7). This difference in scores derived from the same LLM-based chatbot is primarily due to the varying nature and difficulty of the questions. Another potential influencing factor could be differences in prompting. Although other studies have demonstrated that diagnostic reasoning prompts (44) and prompt engineering (45) can affect the accuracy of responses from LLMs, our previous study did not show a significant change in accuracy when we removed the reference to the "CFPC exam" from our prompt (6).

We recently examined ChatGPT-3.5 and ChatGPT-4, trained using data up to September 2021, to answer a sample CFPC questionnaire that we also used in this study. We found that 73.8% and 85.1% of ChatGPT-3.5 and ChatGPT-4 responses were correct, respectively (6). Running GPTs after a one-week interval, regenerating the response or using different prompts did not significantly affect the accuracy of the answers (6). In the current study, we employed ChatGPT to answer the same questionnaire after about six months. Surprisingly, the correct response rates for ChatGPT-3.5 and ChatGPT-4 were almost identical to our previous study (76.4% and 85.5%, respectively). The trends in reviewer scores in the present study were also very similar to our previous findings from the previous study (6). In contrast, Goodman et al. repeated the scoring of 36 low-scored questions from a set of 284 medical questions developed by 33 physicians across 17 specialties using ChatGPT after 8 to 17 days. They observed substantial improvement in the chatbot's scores to this subset of questions. However, they did not repeat all the questions, and it is possible that some answers previously marked as correct could have changed to incorrect, potentially affecting their findings (46). ChatGPT-4 significantly outperformed ChatGPT-3.5, which uses the GPT-3.5 LLM ($P<0.05$), aligning with findings from our previous study (6).

ChatGPT-3.5 achieved a CFPC score of 76.4% and scored higher than its rival, Gemini, although the trend was not statistically significant. ChatGPT-3.5 has been extensively tested across various examination contexts and remains a valuable tool despite its limitations. It has shown a performance of 60.2% on MedQA, 62.7% on MedMCQA (medical entrance exam questions from India), and 78.2% on PubMedQA (Yes/No/Maybe questions derived from expert-annotated PubMed abstracts) (19).

Gemini, launched by Google (14) and PI-AI, introduced by Inflection AI (39), are new systems based on other LLMs instead of GPTs. According to the CFPC answer key, Gemini and PI-AI scored very closely to GPT-based rivals like ChatGPT-3.5 and You-sm.

Pal et al. benchmarked some LLMs on the MedQA (USMLE-style questions) dataset and found that Gemini achieved a score of 67.0%, close to the 69.1% we found in our study. In their study, Gemini's performance was significantly behind ChatGPT-4 and Med-PaLM 2, which both scored around 86% (29). Furthermore, they found Gemini was highly susceptible to hallucinations, overconfidence, and knowledge gaps (29).

As expected, similar to our ChatGPT-3.5 and ChatGPT-4 observation, Gemini-A, Google's most advanced LLM, significantly outperformed its basic model, Gemini. Gemini-A achieved 81.2% on the CFPC score, ranking third among all systems and comparable to ChatGPT-4.

You-sm recorded the lowest CFPC score among our studied systems. It tended to score lower than ChatGPT-3.5 (P=0.066 CFPC score and P=0.104 Reviewers' score). However, during the current study (February 2024), we used YouChat version 1.0.9 (updated on September 27, 2023), which was still in the experimental phase. At that time, smart mode (You-sm) utilized the GPT-3 LLM (see Table 1) (22). The latest version available at the time of writing this article employs GPT-4 in smart mode (37). Consequently, we anticipate that the performance of the newer YouChat may be improved, warranting further investigation.

You-4, another mode of YouChat that claims to use GPT-4 LLM with the possibility of doing online searches, demonstrated the lowest performance among the counterpart systems that use GPT-4. However, as noted, YouChat has recently changed significantly (37), and the performance of several newer available models from YouChat needs further studies.

General Domain LLMs Without Access to Online Search Versus Those With Access to Online Search

A significant drawback of closed-source LLMs such as ChatGPT-3.5 and GPT-4 is their lack of updated information. This issue restricts their wider use and customization, particularly in areas like medicine, where specific knowledge and context are essential (15). For instance, in one study, one of the LLMs that had access to online resources like Wikipedia outperformed ChatGPT-3.5 in answering questions about the emerging medical concern of monkeypox (17). However, in our study, ChatGPT-4 and Gemini-A provided acceptable accuracy compared to other systems with online access. ChatGPT-4 and Gemini-A, using advanced LLMs of GPT-4 and Gemini Ultra 1.0 from OpenAI and Google, outperformed You-4 and were comparable to Copilot, which both employ the GPT-4 LLM and claim to use online searches. Additionally, ChatGPT-4 and Gemini-A achieved results that were very close to those of the medically trained systems with online search capabilities, AI-Dr and ChatDr. Similarly, You-sm and PI-AI, two systems that have access to online information, did not demonstrate any performance advantage over ChatGPT-3.5 and Gemini, which do not claim the ability to access online data and search.

Two questions about treating COPD underscored the up-to-date knowledge of some systems in our study. These questions were centred on treating mild (MRC2) COPD. Gemini, Gemini-A, and PI-AI provided answers based on the 2023 guidelines, whereas other systems used the 2019 guidelines. Notably, neither Gemini nor Gemini-A claim to have access to online searches. Apart from PI-AI, other systems that claim online search capabilities (You-sm, Copilot, AI-Dr, and Chat-Dr) did not provide answers in line with the most recent guidelines. This observation may be due to the fact that these 3 LLMs have been launched very recently and may contain newer information in their pretraining.

Copilot, employing GPT-4 LLM (38) and using online searches, was on par with ChatGPT-4 and Gemini-A from general domain LLMs. The additional benefits of employing Copilot will be discussed later.

Our observation indicates that access to online search was not a critical factor in selecting a system to assist with our sample Canadian family medicine exam questions. Our question set may include samples from previous exams to introduce general concepts and understanding of the CFPC exam to the test takers. However, caution is advised when using older trained LLMs lacking access to the latest or online information, particularly when addressing recent medical issues.

Performance of LLMs Trained in Medical Domain Knowledge

A major concern with general domain LLMs is their lack of specialized training for medical topics. In our study, we employed two systems specifically trained on medical knowledge: AI-Dr, based on the GPT-4 LLM (30), and ChatDr, which utilizes a modified version of LLaMa, an open-source LLM developed by Meta (17). These two systems showed acceptable performance in providing correct responses. AI-Dr ranked second among all chatbots with an 82.4% CFPC score, while ChatDr, with 81.4% correct answers, tied for third place with Gemini-A. However, this absolute difference in accuracy (3.1%) compared to ChatGPT-4 might not be practically relevant for medical exam preparation.

AI-Dr has been trained using a substantial medical corpus (31), in an attempt to improve medical inquiry responses compared to general domain GPT-4 LLMs. In our study, AI-Dr demonstrated significantly higher correct answers than Gemini, You-sm, You-4 and PI-AI. It also trended to score higher than ChatGPT-3.5. However, it did not outperform ChatGPT-4 and Gemini-A and Copilot.

ChatDr is designed to enhance understanding of medical-related inquiries using several real patient-doctor dialogues (17). The results and scores of this LLaMA-based LLM closely matched the performance of its GPT-4 operated counterpart, AI-Dr, with a 0.2% lower CFPC score and 0.6% higher reviewer score. ChatDr demonstrated significantly higher correct answers than Gemini, You-sm, You-4, PI-AI and ChatGPT-3.5. However, Similar to AI-Dr, it did not outperform ChatGPT-4 Gemini-A and Copilot.

A study comparing the responses generated by ChatDr to those from ChatGPT found that ChatDr significantly outperformed ChatGPT (17), which aligns with our findings. However, a newer open-source medically trained LLM, named PMC-LLaMA, outperformed Clinical-Camel on MedQA, MedMCQA, and PubMedQA (47), indicating that the field of open-source medical domain LLMs is evolving.

Another potential benefit of ChatDr is that it is considered an open-source LLM developed by modifying the open-source LLaMA (17). It uses publicly available open-source data and can be distributed to the public (48). Open-source LLMs offer advantages such as greater transparency regarding training sources and increased customization due to feedback from a wide range of users (49). Furthermore, open-source LLMs are considered more secure because they can be installed and used on private computers, eliminating the need to send sensitive medical information to a third party for processing (49). This helps address some ethical concerns associated with commercially available LLMs and is a crucial consideration in terms of confidentiality.

Among our studies, LLM-based Chatbots that could be used without subscription ChatDr, followed by Copilot, provided the most accurate responses, significantly outperforming Gemini and You-sm. Therefore, ChatDr stands out as a compelling choice and one of the acceptable options for answering our questionnaire.

Chen and collaborators released MEDITRON, a suite of open-source language models with 7B and 70B parameters adapted to the medical domain. This LLM was built on LLaMA-2 and extended pretraining was done on a comprehensive medical corpus, including selected PubMed articles, abstracts, and internationally recognized medical guidelines. After additional training, Chen et al. compared their model MEDITRON 70B's performance to the base model LLaMA-2-70B. They found that further training significantly improved MEDITRON 70B's performance on all evaluated benchmarks (20). MEDITRON-70B also outperformed ChatGPT-3.5 (with 175B parameters) on all benchmarks they used and was comparable to ChatGPT-4 (with 540B characters) (20). They concluded that although their model had fewer parameters than the premium general domain LLMs, they could improve their performance by fine-tuning and specialized training. Our study results were in agreement with their observations. We demonstrated that ChatDr, despite having fewer parameters, produced results comparable to those of high-parameter models like GPT-4 and Gemini-A.Xie and colleagues (15) similarly developed Me LLaMA by pre-training and fine-tuning LLaMA2 using extensive medical data. Their model performed better than ChatGPT-3.5 and ChatGPT-4 on the MedQA (USMLE style) dataset (15). In this study, Xie and associates also showed that larger parameter sizes do not guarantee better performance. For example, their Me LLaMA 13B system with 13B parameters, outperformed its larger counterpart, LLaMA2-70B-chat, with a significantly larger parameter size (70B parameters) on 6 out of 12 datasets and was comparable on the other 6 (15).

Some other examples of such medical domain LLMS are BioMistral 7B, which demonstrated a 4.0% improvement over GPT-3.5 Turbo (18), Flan-PaLM2 with 67.6% accuracy on MedQA (21) and a fine-tuned chatbot for the Chinese National Medical Licensing Examination (CMExam), which was comparable to ChatGPT-3.5 (28).

AI-Dr, trained on medical data (31) to improve responses to medical inquiries, costs $14.99 monthly. However, it did not outperform the free services of ChatGPT-3.5 and Copilot, nor the paid general domain LLMs, ChatGPT-4 and Gemini-A.

Qualitative Assessment of the Responses

Our study assessed the open-ended responses qualitatively to better understand the chatbots' potential in learning family medicine and aiding in CFPC exam preparation. Although we did a comparison of the results according to the accuracy result, it is important to note that the decisions should not solely depend on p-values and also needs to consider this accuracy in the clinical and practical context. For example, "unsafe responses" (e.g., recommending suture removal at incorrect times) are critical to identify even if statistical tests show no significant differences between models.

Supplement 1 provides examples of this qualitative assessment. This assessment system was unique to our study and had not been described before. Few studies have previously conducted qualitative evaluations of open-ended responses to questions (46, 50). One study compared answers provided by ChatGPT to those given by real physicians in a public and non-identifiable database of questions (Reddit's r/AskDocs). In this study, evaluators chose "which response was better" and assessed both "the quality of information provided" (rated from 1 to 5, ranging from very poor to very good) and "the empathy or bedside manner demonstrated" (rated from 1 to 5, ranging from not empathetic to very empathetic). The investigators found that in 78.6% of cases, evaluators preferred ChatGPT's responses over those provided by physicians. ChatGPT's responses were significantly longer and received higher ratings for both quality and empathy (50).

In our study, only 1.8% of questions were left unanswered by Gemini and Gemini-A and 0.6% by ChatGPT. Most of the other models in the study successfully answered all the provided questions. Furthermore, the frequency of irrelevant responses unrelated to the scenario in the question was minimal.

Although the CFPC sample we used includes an answer key, the answer key does not provide detailed explanations or rationale behind the choices. Providing such additional information would be valuable for better understanding the concepts and aiding in memorization for future reference. Most of our LLM-based chatbots, however, provided extra helpful information in response to our questionnaire. This additional information offers a significant advantage in the study of medicine and exam preparation when provided by LLMs.

Providing references could be another significant advantage for accessing further reading and validating the responses during exam preparation and learning. In our study, Copilot not only provided references for all of its responses but also demonstrated acceptable accuracy comparable to that of Gemini-A, AI-Dr, and ChatDr. Given that Copilot is free and can access online resources, it stands out as one of the most suitable options for CFPC exam preparation among our systems.

You-sm provided references in only 22.1% of its responses, according to Table 5. Similarly, You-4 offered references for just 16.9% of questions and did not perform better than ChatGPT-3.5, as shown in Tables 3 and 4. Gemini, which provided references in only 5.2% of responses, also does not appear attractive from this perspective.

Unintelligent responses may result from the chatbots' inability to deeply understand the concepts. The highest frequency of such responses was observed with You-sm at 4.8%; however, as previously discussed, the most recent version of this chatbot has undergone significant improvements. Supplement 2 provides examples of these unintelligent responses to better understand this issue.

The CFPC exam requires candidates to provide answers within a specific number of responses. If a question asks for a certain number of answers and the candidate provides more, only the initial responses closest to the required number will be accepted. This aspect highlights the importance of understanding and following instructions, which can also reflect the chatbot's intelligence. Most of our chatbots adhered to this rule by limiting their responses to the number requested, except Gemini, which ***provided 14.3% more lines of responses than requested***.

Finally, providing ***safe*** responses may be of significant importance in the field of medicine to avoid harm to the people and community. Unsafe responses can mislead learners and represent a significant concern in the fields of learning and exam preparation. In our study, You-sm exhibited the highest rates of unsafe (9.1%) responses. Conversely, ChatGPT-4 and AI-Dr provided the safest answers, equally followed by Gemini-A and ChatDr.

Study Limitations

Human evaluation of open-ended text responses posed significant challenges in this study. Initially, we planned for the three CFPC-certified reviewers to independently score responses blindly, reporting the agreement among them. However, as the process unfolded, we observed occasional discrepancies in scoring, even when using the CFPC's standard answer key. These disagreements arose partly from rare instances of misunderstandings by reviewers or the overlooking of relevant information embedded within the verbose explanations generated by the LLM-based chatbot.

Additionally, inconsistencies in scoring strategies further complicated the process. The CFPC exam requires candidates to adhere to specific response limits. For example, when a question asks for two answers, only the first two responses are considered, and any additional responses are viewed as a lack of conceptual mastery and not scored. Our chatbots demonstrated a different approach, as highlighted in our quantitative evaluation, where its "intelligence" sometimes led it to provide more than the required number of responses. While some reviewers strictly adhered to evaluating only the first initial responses, others accepted any correct answers provided, even if they followed the initial requested responses.

Recognizing the importance of consistent scoring among the three reviewers, we convened a second joint discussion session to address these discrepancies. During this session, all reviewers agreed to strictly adhere to the unified scoring approach outlined in our methods section. This collaborative effort significantly reduced scoring disagreements. Ultimately, the scores determined during this second review were used as the final scores, and only the Fleiss' kappa for these revised scores was reported.

This study compared the performance of ten LLM-based chatbots across several evaluation metrics, including CFPC scores, reviewers' scores, and qualitative criteria. Conducting multiple statistical tests without adjustment increases the risk of Type I error, where differences are deemed significant due to chance rather than true differences. Therefore, interpretation of P-values should be performed cautiously in the clinical context. Our qualitative assessment further explored the importance of these P-values.

The field of LLMs is evolving rapidly, prompting us to perform our comparisons over a short period, from February 3rd to February 23rd, 2024. During the analysis and drafting of this article, some of our systems underwent significant updates. OpenAI introduced ChatGPT-4o, which can process various forms of input, including text, audio, images, and videos, and produce outputs in text, audio, and image formats (51). Meanwhile, Google launched Gemini Pro 1.5, which boasts enhanced image comprehension, an extended context window, and new data analysis capabilities (40). Additionally, YouChat has significantly updated its features, shifting its smart mode to primarily utilize GPT-4 by default instead of GPT-3 and enabling the use of a variety of other LLMs, including GPT-4 Turbo, Claude Instant, Claude 2, Claude 3 Opus, Claude 3 Sonnet, Gemini Pro, DBRX-Instruct, and Zephyr (Uncensored) (32). Consequently, the accuracy and performance of these new models need further investigation.

As we mentioned in the methods section, one of the questions of our questionnaire contained an ECG that was removed from the question in this study because of the absence of image processing features in most of the current systems. We found that the responses from the 10 LLM-based chatbots to this question were correct, likely because other textual clues in the questionnaire rendered access to the ECG image unnecessary. However, this field is rapidly evolving. For instance, ChatGPT, Gemini, and several other new chatbots now allow us to submit images for interpretation. Interestingly, ChatGPT correctly diagnosed the ECG in our questionnaire, while Gemini did not. Since interpreting ECGs and X-ray images are crucial aspects of medical education, these capabilities will likely become increasingly valuable, meriting further investigation.

Finally, PI-AI, Gemini and Gemini-A occasionally offered additional answers after discussing the required responses. These additional answers often included the CFPC-required answers, although they were not prioritized in the initial responses. In this study, we did not count these additional responses for CFPC scores but accepted them for reviewer scoring. This approach significantly increased the reviewer scores for these systems compared to their CFPC answer key results. The most significant change was seen in PI-AI (69.1% correct rate for Gemini turned to 82.4%, P=0.005; 81.2% correct responses for Gemini-A turned to 90.9%, P=0.011; and 70.9% for PI-AI turned to 89.1%, P<0.001).

CONCLUSION

We conducted this study to compare the accuracy and quality of responses provided by ten different LLM-based chatbots on an open-ended simulation questionnaire of the CFPC exam. For our evaluation, we selected four general domain LLMs without online search capabilities, four LLM-based chatbots with online search functionality, and two LLMs specifically trained on medical knowledge.

ChatGPT-4 delivered the highest accuracy at 85.5%, followed by AI-Dr at 82.4%, Gemini-A and ChatDr (both at 81.2%), and Copilot at 80%. In our study, advanced general domain LLMs like GPT-4 and Gemini-A Ultra 0.1 continue to be highly effective for CFPC exam preparation and medical education. Access to online resources did not significantly enhance the accuracy of responses from general domain LLMs. ChatDr, an LLM trained in medical domain knowledge, demonstrated promising and comparable results to GPT-4 and Gemini-A despite fewer parameters.

LLM-based chatbots rarely refused to answer questions. Most responses were relevant, safe, aligned with the question's guidelines, and included additional explanations that enhanced learning. Copilot, utilizing GPT-4, provided the added benefit of offering external references for all questions, significantly enhancing its educational value.

Recommendation for future research

Further studies may be required to examine the strengths and weaknesses of various other LLMs in addressing exam questions and aiding in family medicine education and learning. Specifically, more tremendous efforts and focus should be directed toward developing more efficient medical domain LLMs.

Our extensive experience in human evaluation in this study showed that while independent blinded scoring combined with selecting the most agreed-upon score can be a reasonable approach for multiple-choice questions, it is less effective for open-ended responses. A joint discussion session among evaluators similar to our approach is recommended. This approach helps minimize errors and reduces disparities in scoring strategies, ensuring greater consistency and accuracy for future studies.

Acknowledgement

This manuscript's grammar was refined using Grammarly Premium. We extend our sincere gratitude to Dr. Ramana-Kumar Agnihotram and Dr. Jose Correa for their invaluable guidance in selecting the statistical analysis methods. 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).

REFERENCES

1. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of Artificial Intelligence in Community-Based Primary Health Care: Systematic Scoping Review and Critical Appraisal. J Med Internet Res. 2021;23(9):e29839.

2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40.

3. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol. 2023.

4. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6).

5. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. 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.

6. Mousavi M, Shafiee S, Harley JM, Cheung J, Rahimi SA. Performance of ChatGPT in Certification Examination of College of Family Physicians of Canada, doi: 10.1136/fmch-2023-002626. PMID: 38806403; PMCID: PMC11138270. Fam Med Community Health. 2024 May 28(12 (Suppl 1)):e002626.

7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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.
9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.
10. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, 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.
11. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ. 2023;9:e48002.
12. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination (preprint): Research Square; 2023 [updated 2023. Available from: http://europepmc.org/abstract/PPR/PPR617702.
13. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, 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.
14. Gemini Team Google, Anil R, Borgeaud S, Wu Y, Alayrac J-B, Yu J, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.
15. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Me LLaMA: Foundation Large Language Models for Medical Applications. arXiv preprint arXiv:240212749. 2024.
16. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:230509617. 2023.
17. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023;15(6):e40895.

18. Labrak Y, Bazoge A, Morin E, Gourraud P-A, Rouvier M, Dufour R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv preprint arXiv:240210373. 2024.
19. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns. 2024;5(3).
20. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:231116079. Nov 27, 2023.
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
22. You.com. Youchat Feb 2024 [Available from: https://you.com/?chatMode=default.
23. Li J, Dada A, Kleesiek J, Egger J. ChatGPT in Healthcare: A Taxonomy and Systematic Review (Preprint). medRxiv. 2023:2023.03. 30.23287899.
24. ICMJE. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Updated May 2023: International Committee of Medical Journal Editors (ICMJE). Accessed Nov 2023; [Available from: https://www.icmje.org/icmje-recommendations.pdf.
25. Zielinski C, Winker M, Aggarwal R, Ferris L, Heinemann M, Lapeña JF, et al. Chatbots, ChatGPT, and Scholarly Manuscripts: WAME Recommendations on ChatGPT and Chatbots in relation to scholarly publications. Natl Med J India. 2023;36(1):1-4.
26. The College of Family Physicians of Canada. Preparing for the Certification Examination in Family Medicine 2023 [Available from: https://www.cfpc.ca/en/education-professional-development/examinations-and-certification/certification-examination-in-family-medicine/preparing-for-the-certification-examination-in-fam.
27. OpenAI. ChatGPT [Large language model]. Accessed Feb 2024. Last Update Apr 2023. Feb 2024 [Available from: https://chat.openai.com.
28. Liu J, Zhou P, Hua Y, Chong D, Tian Z, Liu A, et al. Benchmarking Large Language Models on CMExam-A Comprehensive Chinese Medical Exam Dataset. Advances in Neural Information Processing Systems. 2024;36.

29. Pal A, Sankarasubbu M. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. arXiv preprint arXiv:240207023. 2024.
30. Stojan J. Docus.ai: Innovating Global Healthcare through AI-Powered Diagnostics: USA today; Aug 12, 2023 [Available from: https://www.usatoday.com/story/special/contributor-content/2023/08/22/docus-ai-innovating-global-healthcare-through-ai-powered-diagnostics/70653228007/.
31. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-8.
32. Garg S. 30 Best ChatGPT Alternatives in 2024 (Free & Paid): Writesonic; Feb 12, 2024 [Available from: https://writesonic.com/blog/chatgpt-alternatives.
33. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388(13):1233-9.
34. Lo CK. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences. 2023;13(4):410.
35. Hsiao S. Bard becomes Gemini: Try Ultra 1.0 and a new mobile app today Feb 08, 2024 [Available from: https://blog.google/products/gemini/bard-gemini-advanced-app/.
36. Hsiao S. Get more done with Gemini: Try 1.5 Pro and more intelligent features May 14, 2024 [Available from: https://blog.google/products/gemini/google-gemini-update-may-2024/.
37. You.com. You.com is the go-to platform to test and compare the latest AI models February 10, 2024 [Available from: https://about.you.com/youdotcom-is-the-go-to-platform-to-test-and-compare-different-ai-models/.
38. Microsoft blog. July 18, 2023. [cited Feb 2024]. Available from: https://blogs.microsoft.com/blog/2023/07/18/furthering-our-ai-ambitions-announcing-bing-chat-enterprise-and-microsoft-365-copilot-pricing/.
39. Inflection AI. Pi AI Feb 2024 [Available from: https://pi.ai/talk.
40. Banderson K. Pi AI Review - The Chatbot with a Twist: MS Power; February 6, 2024 [Available from: https://mspoweruser.com/pi-ai-review/.
41. UpToDate 2023 [Available from: https://www.uptodate.com/contents/search.

42. Touvron H ML, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 19 Jul 2023.

43. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.

44. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. npj Digital Medicine. 2024;7(1):20.

45. Yan S, Knapp W, Leong A, Kadkhodazadeh S, Das S, Jones VG, et al. Prompt engineering on leveraging large language models in generating response to InBasket messages. Journal of the American Medical Informatics Association. 2024.

46. Goodman RS, Patrinely JR, Stone CA, Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Network Open. 2023;6(10):e2336483-e.

47. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:230414454. July 3, 2023.

48. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.

49. DeLair G. Large Language Models: Open Source vs. Closed Source February 20, 2024 · [Available from: https://echobasedev.com/large-language-models-open-source-vs-closed-source/.

50. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589-96.

51. Open AI. Hello GPT-4o May 13, 2024 [Available from: https://openai.com/index/hello-gpt-4o/.

Supplement 1. Examples for Evaluating Chatbot Responses Qualitatively

	Question	**Answer**
Irrelevant response	You are on call at the emergency department (ED) of a rural hospital. The nurse tells you a very anxious woman is on the phone. The woman wants to speak to you about her 2-year-old daughter. The daughter has ingested some poison. She is asymptomatic and appears well. **Question 1-3:** You might want to enhance excretion of the poison. What techniques are available to enhance poison excretion? List TWO.	**You-4:** Techniques to Enhance Poison Excretion Biliary Excretion: The liver can also excrete toxins by secreting them into the bile, which then flows into the intestinal tract. Chemicals excreted in the bile are eventually eliminated in the feces. This process can be particularly important for the elimination of certain types of toxins, especially those that are highly lipophilic [The question asks about techniques that a doctor can use to enhance excretion. The chatbot is bringing in the pathway of excretion of toxins by the body]
Unsafe answers	A five-year-old boy presents at your urgent care centre after falling off his bike one hour ago. His only injury is a cut on his upper lip…After taking appropriate anesthetic measures, you proceed to close the wound with a non-absorbable 6-0 monofilament suture. Question 9-3: The father wants to know when is the soonest they should return to have the sutures removed. What do you tell him? Give ONE answer.	**You-sm**: The sutures used to repair the 1 cm long laceration on the boy's upper lip should be removed within 4-14 days. [The answer is 3 to 5 days Do Not accept less than three or more than five days]
Extra useful information	A 55-year-old woman comes to your office to discuss menopause issues… Question 8-3: You conduct a baseline bone mineral density which reveals she has a moderate risk for fracture. After reviewing her bone density results and making lifestyle suggestions, you discuss drug therapy with her. She has always refused to take hormone replacement therapy. She is already taking a calcium and vitamin D supplement. At this time, what first line medications other than bisphosphonates could you suggest to her for osteoporosis prevention? State TWO.	**ChatDr**: 1. Denosumab: This is a monoclonal antibody that can slow bone loss by inhibiting osteoclasts, the cells that break down bone. 2. Raloxifene: This is a selective estrogen receptor modulator that replicates the beneficial effects of estrogen on bone density without some of the risks associated with estrogen therapy. [Answer is: • Teriparatide / Forteo / recombinant parathyroid hormone • Raloxifene / SERM / Evista • Denosumab / Prolia / Rank ligand inhibitor]

List of answers more than requested	A previously healthy 69-year-old female widow comes to you complaining of fatigue of four months duration… Question 2-1: List other symptoms, not mentioned above, that you should specifically enquire about to help rule out sinister medical causes for her fatigue. List TWO.	**Copilot:** Some other symptoms that you should specifically enquire about to help rule out sinister medical causes for her fatigue are: Fever, night sweats, or weight loss… Abdominal pain, jaundice, or dark urine…. Joint pain, rash, or oral ulcers…
Additional answers offered after list of answer completed	A previously healthy 69-year-old female widow comes to you complaining of fatigue of four months duration… Question 2-1: List other symptoms, not mentioned above, that you should specifically enquire about to help rule out sinister medical causes for her fatigue. List TWO.	**PI-AI:** 1. Changes in sleep patterns… 2. Changes in cognitive function… Additionally, it is important to inquire about any changes in appetite, bowel habits, and menstrual cycle (if applicable), as well as any recent travel, exposure to infections, or environmental changes…

Supplement 2. Instances of unintelligent responses from chatbots

Question	LLM-based Chatbot	Sample question and answer from the LLM-based Chatbot
1-6	You-sm	Q: Other than measurement of the acetaminophen plasma level, what laboratory tests are important for determining whether antidote treatment is effective? List THREE. Answer: Acetaminophen level
3-3	All of chatbots	Q: A 17-year-old female dancer visits your office because she has had amenorrhea for three months. She is sexually active with one partner. On further questioning, she states that she has been dissatisfied with her appearance and she has lost quite a bit of weight in the past year. Although she looks cachectic, she feels she is still overweight. Her sleep has been somewhat disturbed, and she has withdrawn from all her activities at school. Her amenorrhea is now really concerning her. You suspect an eating disorder. She is significantly under-weight for her height. She admits to binge eating and purging with laxatives to prevent herself from gaining weight. Her parents are quite concerned about an eating disorder, as are you. What type of eating disorder does she have? State ONE. A; [The answer to this question is Anorexia nervosa binge-eating/purging type. None of the chatbots realized this. Low weight of the patient is against bulimia nervosa]

6-3	You-sm, You-4, Gemini, Gemini-A, AI-Dr	Q: What activity must you inquire about? State ONE. Answer: Sleep pattern [not an activity]
7-2	ChatGPT-3.5, You-sm, You-4, ChatDr	Q: … he was on shore leave six months ago, he and some colleagues visited Thailand "for some rest and relaxation." He asks whether he "might have caught something" there. What historical elements should you inquire about to ascertain his risk of having contracted viral hepatitis? State FOUR. A: ChatGPT, You-sm, You-4, ChatDr: Exposure to contaminated food [will not happen in 6 m]
14-2	Copilot, Gemini	Q: A car salesman…. He has a past history of hypertension, an appendectomy at age 16… Other than diverticulitis, cancer, genitourinary causes, and various types of colitis, what OTHER diagnoses should you consider? List THREE. A: Copilot: torsion ovary and ectopic pregnancy [while patient is man] A: Gemni: Appendicitis [There is Hx of appendectomy] and PID [Patient is man]
14-2	ChatGPT-3.5, ChatGPT-4, You-sm, You-4, Copilot, Gemini, Gemini A, PI-AI, AI-Dr, ChatDr	Q: A car salesman, age 58, comes to see you because of some abdominal pain… Excluding findings from the rectal exam and peritoneal signs, what physical signs should you look for during an abdominal exam, which, if present, would be consistent with a surgical cause for his symptoms? List THREE. A: All of the mentioned bots have mentioned abdominal rigidity guarding or both as their diagnosis. [The question says excluding peritoneal signs and the mentioned symptoms are peritoneal signs]
15-3	ChatGPT-3.5, Copilot, Gemini, Gemini-A, AI-Dr, Repeated with Q1, Q2, Q3: You-sm, You-4,	Q: A woman, age 55, presents after suffering her third upper respiratory infection this year… you told her you suspect she has chronic obstructive pulmonary disease (COPD)… Question 15-1 What SPECIFIC measurement from her pulmonary function test would allow you to confirm the diagnosis of COPD? Question 15-2 Once you have confirmed a diagnosis of COPD, what is the MOST important intervention you would suggest to her at this time? State ONE. Question 15-3

		The patient returns to complain that, in the past three months, she has noticed shortness of breath when she is trying to hurry on level ground or up a slight hill. She denies orthopnea or paroxysmal nocturnal dyspnea. Last month, her chest X-ray examination was normal. What OTHER investigation would you consider at this time to evaluate these symptoms? State ONE. A: Spirometry [Spirometry is to confirm COPD and is related for question one. Question 3 asks what other investigation, and the answer should be something else]. Second time the question 3 was repeated with all the previous questions. Y-smart, Y_GPT4
17-2	You-sm	Q: Apart from frequency, urgency, and nocturia, what symptoms might the patient complain about if he had an enlarged or cancerous prostate? State TWO. A : 1. Urinary frequency 2. Urinary urgency …
18-3	ChatGPT, You-sm, You-4, Gemini, PI-AI	Q: The patient is worried about gastric cancer. Excluding constitutional symptoms and the history already provided, what ADDITIONAL symptoms would you ask about? State TWO. A: Weight loss [This is a constitutional Sx]

CHAPTER 5: DISCUSSION

Substantial advancements in AI-related applications have recently occurred. For instance, LLMs, primarily designed for text generation, have become valuable tools in medical education (72), and exam preparation (95, 100).

Family medicine is the foundation of primary health care and encompasses a broad spectrum of knowledge across various medical fields (84). A critical element of family medicine evaluation in Canada is the two-step CFPC exam. These steps include the written Short Answer Management Problems (SAMPs) and the Simulated Office Oral (SOO) exam, which acts as the board evaluation for this specialty (84). Given the extensive knowledge required in family medicine, students often need to invest considerable time, money, and effort in preparing for the CFPC exam by practicing with sample questions.

Recently, there has been a marked interest in employing LLMs to respond to questions from professional board exams (97, 99, 105) to evaluate their usefulness in enhancing learning (61). In this thesis, we aimed to evaluate the performance of some LLMs on comprehensive, open-ended SAMP questions.

Our first study compared the accuracy of GPT-3.5 and GPT-4 responses on our sample open-ended questionnaire. Additionally, we examined their consistency in responding to the same questions under varying conditions, such as being asked at different times, regenerating answers, or using prompts for debriefing.

In our second study, we compared ten different LLM-based chatbots with varying capabilities to answer the same questionnaire. We evaluated four general domain LLMs without online search capabilities, four general domain LLMs with online search functionality, and two LLM-based chatbots trained specifically on medical knowledge. We compared the accuracy of these systems in responding to our sample questionnaire and identified their advantages and disadvantages in a qualitative assessment.

Overall, in the first stage of this investigation, we demonstrated the state-of-the-art performance of ChatGPT on this complex, open-ended questionnaire, along with good consistency. The average CFPC score across five rounds was 85.2 ± 23.7 for GPT-4 and 76.0 ± 27.7 for GPT-3.5. Our CFPC scores for GPT-4 closely matched those recorded by Huang et al. in their study using sample multiple-choice preparation materials for the CFPC exam (95). However, our scores achieved by GPT-3.5 were higher than their reported accuracy rate of 57.4% in the aforementioned study (95). More importantly, by acknowledging additional correct responses not included in the CFPC answer key, the average Reviewers' score was 93.51 ± 18.65 for GPT-4 and 86.01 ± 24.26 for GPT-3.5 in our study. This observation indicates that the frequency of errors and incorrect responses was very low in our study, supporting the reliable use of ChatGPT, especially GPT-4, for CFPC exam preparation. Furthermore, ChatGPT was capable of generating valid answers that were not pre-included in the CFPC's answer key. Therefore, ChatGPT or probably other LLMs can offer diverse perspectives and answers, which could be particularly useful for educational purposes and broader thinking in medical training.

Another important finding of our study was that GPT-4 achieved higher scores than GPT-3.5, consistent with the results of several other studies that compared these two systems. These include the study from the University of Toronto using a sample of multiple-choice progress tests for the SAMPs exam (95), the USMLE exam from the National Board of Medical Examiners (NBME) (86), the Japanese Medical Licensing Examination (101), the StatPearls ophthalmology Question Bank (97), the 2022 SCE neurology exam (105) and the Iranian medical residency exam (103).

It's worth noting that the CFPC does not accept a predetermined minimum number of SAMP cases or a cut-off score for passing the exam (84). Instead, the minimal passing score is set based on the performance of a reference group of first-time test-takers who are graduates of Canadian family medicine residency programs (84). Consequently, our study cannot confirm that using either GPT-4 or GPT-3.5 will ensure a passing score on the CFPC exam.

We analyzed the average percentage of repeated answers to examine the consistency of AI-generated responses. To calculate the percentage of repeated concepts, we compared each round of responses to a chosen reference round. This analysis identified the degree of overlap in the concepts addressed within the answers for each question across the different rounds. Our comparisons revealed more consistency in GPT-4's responses than in GPT-3.5. GPT-4 particularly demonstrated a higher tendency to repeat answers more consistently compared to GPT-3.5. This pattern of repetition, combined with stable performance across different testing scenarios, underscores GPT-4's robustness and reliability in generating responses. The observed consistency was more impressive when the reviewers realized that changing the answer choices by GPT would not impact the accuracy of the provided responses.

Our study conducted five rounds of comprehensive evaluations of ChatGPT's performance in relation to the CFPC exam. These rounds aimed to assess the effects of repeating identical prompts at intervals, modifying prompts by removing the reference to the "CFPC exam," regenerating responses, and eliminating prompts altogether. These comparisons of rounds yielded these intriguing insights:

● **Consistency Over Time:** Both GPT-3.5 and GPT-4 demonstrated high consistency and accuracy when tested with identical prompts across approximately one-week intervals. However, it should be noted that in our study, we did not provide any feedback to ChatGPT to avoid the influence of repeated questions and feedback provided over time. Furthermore, we erased all of the chains of response from the Chatbot's memory before studying the next round to minimize the effect of AI learning by repeating questions. This observation suggests that the passage of short time intervals does not significantly affect the AI's performance, indicating that improvements are more likely due to advancements in the AI models, updated training materials and probably feedback or instructions from human users rather than only the passage of time.

Similar to our observation, Thirunavukarasu et al. conducted two independent sessions of the AKT exam using ChatGPT within a 10-day interval and demonstrated consistent performance (100). Conversely, another study showed improvement in the incorrect responses provided by the LLM after repeating the questions within 8 to 17 days. However, this observation may be biased, as they did not evaluate changes to the correct responses over time and only repeated incorrect responses (108).

● **Impact of Prompting on Accuracy:** Our first round of ChatGPT answering to CFPC questions involved a prompt to limit each line to 10 words, while in the fifth round, we did not use any prompt (zero-shot prompt). Comparing these two rounds, where we tested the influence of different prompting strategies, we observed no significant changes in the "CFPC Score Percentage" for either model.

Previous research has shown that diagnostic reasoning prompts can enhance the diagnostic accuracy of the GPT-4 LLM, making it more trustworthy for clinical use (109). Additionally, prompt engineering has improved the quality of LLM-generated responses to Patient Medical Advice Requests (PMARs) (110). However, our prompts, which only specified the number of responses per line in the questions, did not involve diagnostic or clinical reasoning and differed from those used in these studies.

● **Effect of Prompt Modification**: We proposed that hinting at the CFPC exam in the prompt may cause the LLM to search for similar previous questions or responses, potentially biasing its answers. However, interestingly, when we removed the reference to "CFPC exam" from the prompts in round 4, there was a notable increase in accuracy for both models, contrary to our expectations. This finding was unique to our study and was not investigated before.● **Regeneration of Responses:** Regenerating responses in round 3 compared to the second round did not enhance the accuracy for GPT-4 but did show some improvement for GPT-3.5. This difference highlights potential variations in how each model might benefit from response regeneration, possibly due to differences in their training or underlying algorithms. We may conclude that although regeneration of responses could be helpful while employing GPT-3.5, it may not offer added benefit for GPT-4. To our knowledge, the regeneration of responses to medical questions within the same answering session is unique to our study, as no other studies have evaluated these regenerated responses.

ChatGPT, particularly GPT-4, demonstrated remarkable performance in our initial study. However, with the advent of several new publicly available LLMs, such as Gemini (54), the introduction of systems equipped with online search capabilities (48, 49), and the development of LLMs trained on medical domain knowledge (33, 38, 56, 87-90), we deemed it essential to conduct a comparative analysis. Thus, we selected ten LLM-based chatbots to evaluate their accuracy and utility in answering the same sample SAMPs questionnaire utilized in our previous study. Additionally, we analyzed the performance of the freely accessible LLMs employed in our study against those requiring subscription fees, aiming to identify any significant differences in their capabilities. The key findings from our second study can be summarized as follows:

● **High Performance of ChatGPT**: The findings from our second study corroborate that GPT-4 outperformed all other systems, achieving an impressive 85.5% accuracy based on the CFPC answers. Our results substantiate that, although ChatGPT-4 is a general-domain LLM, it excelled in answering our questionnaire, surpassing its newly emerged counterpart, Gemini Advanced, as well as the two medically trained systems, AI-Dr and ChatDr, and all other systems with online access. Furthermore, despite employing ChatGPT nearly six months following our initial study, the accuracy of responses provided by GPT-3.5 and GPT-4 remained largely consistent, matching our previous results.

● **Effect of Online Access:** Systems equipped with online access did not perform better than ChatGPT-4, suggesting that access to real-time information alone may not be the deciding factor in achieving high accuracy in this setting. Probably such applications could be more useful when used for very new knowledge

● **LLMs with Medical Training**: Our two LLMs specifically trained on medical knowledge (AI-Dr and ChatDr), showed an acceptable performance close to the advanced general domain LLMs (ChatGPT-4 and Gemini Advanced), suggesting that domain-specific training is beneficial. In fact, with lower parameters, specific training makes it possible to achieve high-accuracy tasks.

● **Open-source versus Closed-source LLMs:**

Open-source LLMs use publicly available data (18) and offer the advantage of being distributable during their use (17). In contrast, closed-source LLMs can only be accessed with the permission of the developer organizations (17). Closed-source LLMs are valued for high performance driven by greater number of parameters and substantial investments. They are generally believed to outperform open-source models in terms of accuracy and reliability, especially in specialized fields such as medicine (16). However, open-source LLMs have other potential benefits, including more transparency about sources of training (16), the possibility of greater customization, faster innovation (17) due to feedback from numerous users (16) and lower costs for the same output (16, 17). Furthermore, open-source LLMs provide enhanced security (16, 17) by allowing users to control and own their data (16) , as the model is accessible to the user or researcher without transmitting sensitive information to a third party. This makes them a reasonable choice in the medical field.

Only ChatDr used an open-source LLM in our study, while the others employed closed-source LLMs. ChatDr has been developed by modifying the open-source LLM named LLaMA (90). One study compared the responses generated by ChatDr to those provided by ChatGPT and found that ChatDr significantly outperformed ChatGPT (90). However, it performed worse than PMC-LLaMA across MedQA, MedMCQA, and PubMedQA (37). Our results showed that ChatDr's performance was either better than or at least comparable to the other closed-source LLMs with higher parameters.

Several studies have compared the performance of state-of-the-art closed-source models with medically trained open-source LLMs on medical benchmarks. For example, Clinical-Camel (developed on LLaMA-2), outperformed GPT-3.5 on the USMLE Sample Exam, PubMedQA, MedQA, and MedMCQA (36). In contrast, when compared to GPT-4 and Med-PaLM2, Clinical-Camel was inferior in all benchmarks except for PubMedQA (36).

Despite having fewer parameters, PMC-LLaMA (13B) outperformed ChatGPT (175B) on PubMedQA and matched its performance on MedQA (37). Similarly, the MediTron 70B outperformed the higher-parameter GPT-3.5 (175B) on MedQA, MedMCQA, PubMedQA, and MMLU. Additionally, its performance was comparable to GPT-4, Med-PaLM, and Med-PaLM-2, which have significantly larger parameter sizes (38). Finally, Me LLaMA (available in 13B and 70B versions) outperformed ChatGPT on 6 out of 8 datasets. It also achieved better results than GPT-4 on three datasets and demonstrated comparable, though slightly lower, performance on another three. However, it underperformed on the remaining two datasets (33).

These examples demonstrate how open-access LLMs can perform similarly to or even better than closed-access LLMs with significantly higher parameters when properly trained on medical datasets. As the field of LLMs evolves, it remains uncertain whether open-source or closed-source LLMs will become the cornerstone of medical applications.

● **Free vs. Paid Models:** The study noted that some free models, such as ChatDr and Copilot, performed comparably to or better than paid models. This highlights that free models can also be effective tools for medical exam preparation.

Li et al. developed ChatDr, an open-source LLM, and demonstrated its superior performance compared to ChatGPT in responding to medical questions (90), which aligns with the findings of our study.

● **Educational Value and Practical Implications:** The results of our study indicate that certain systems, such as Copilot, provide references accompanied by searchable links for their answers. The availability of references that learners can explore as an initial information source offers notable benefits, including facilitating response validation and encouraging further in-depth exploration of the subject matter.

Our study was among the first to observe that LLM-based chatbots can provide searchable links to their answers. Recently, OpenAI has also enhanced the ChatGPT system to include such references, which could enhance the validity of its responses (50). Using advanced LLMs that provide references can greatly improve learning and help medical students, residents and active physicians better understand difficult subjects. This new way of learning encourages critical thinking and gives students trustworthy sources for their studies. By using these tools, students can learn more effectively and gain a deeper understanding of the topics they're studying.

Pitfalls and Limitations Of Implication of LLMs In Medical Education and Answering Exam Questions

Although the application of Large Language Models (LLMs) in answering exam questions and their integration into medical education presents numerous advantages, it is crucial for both students and instructors to acknowledge the potential drawbacks associated with their usage.

One notable concern is the reliability and accuracy of the information provided by LLMs (4, 29, 44, 61, 62, 73, 80). This issue can be attributed to several factors, including hallucination (7) on the one hand, and the possibility that LLMs are trained on inaccurate information on the other hand (10, 44). It should be noted that these models often cannot distinguish between authentic and unreliable information (85). Moreover, the majority of training data for LLMs is derived from open-access datasets, potentially overlooking the valuable insights offered by expert opinions and reputable textbooks authored by field specialists.

Consequently, it is advisable for learners to treat LLM responses as helpful hints rather than definitive answers while simultaneously engaging in supplemental study using reliable medical resources to ensure a comprehensive understanding of the subject matter.

Continued research is essential to enhance our understanding of the capabilities and limitations of newly emerging LLMs in academic assessments. Moreover, the development of specialized LLMs with focused medical knowledge and online access to references could offer a valuable resource for medical learners seeking reliable information.

Suggestions for Future Studies

The potential of using general-domain LLMs for exam preparation needs further investigation. While our study shows they can help answer sample SAMPs questions, they must be tested on a broader range of questions, including those with images like ECGs. Additionally, we need to compare the effectiveness of LLMs in education with traditional preparation courses and practice exams. We should also explore innovative ways to integrate LLMs into medical education and ensure the safety and effectiveness of these applications.

Another promising use of LLMs in medical education is providing self-study questions. The possibility of using LLMs to help candidates practice more effectively requires further research and innovation.

Developing new, efficient, open-access LLMs tailored to the medical field is an exciting area for future study. These LLMs would not carry the ethical concerns of sharing confidential information with third parties and could be more accessible and personalized for specific medical needs. Such systems could be fully developed for use across various medical specialties.

Finally, LLMs could also be valuable for preparing for the oral part of the CFPC exam (SOO) and other oral Objective Structured Clinical Exams (OSCE). Chatbots that simulate human conversation might be a useful tool for residents to practice for oral exams. Developing such tools could be a focus for future research.

Conclusion

The insights from our study support the use of LLMs as practical tools and valuable resources for answering complex, open-ended CFPC exam preparation questions. General-domain LLMs like GPT-4 and Gemini Ultra-1 performed well, and the medically trained open-source LLM, Chat Doctor, showed promise.

Further research is needed to develop more efficient, specialized, open-access LLMs that are safe and effective for medical learning and exam preparation.

APENDIX

Below is a link to the complete open dataset, which includes all the evaluations conducted by CFPC-certified Family Physicians: “CFPC Certified human scoring to the answers of questions”
https://github.com/rahimi-s-lab/GPT-models-for-CFPC-exam/blob/main/SAMPs%20LLM%20Appendix.pdf

Text box 1 (From the First study Chapter 3).

Sample CFPC question with GPT-3.5 and GPT-4 answers in 5 rounds and scoring. See table 1 from the study one for the prompts. Round 5 was without any prompt and provided comprehensive answers.

Case 2784

You are on call at the emergency department (ED) of a rural hospital. The nurse tells you a very anxious woman is on the phone. The woman wants to speak to you about her 2-year-old daughter. The daughter has ingested some poison. She is asymptomatic and appears well.

Question 1:

Other than demographic information, what important initial pieces of information must you elicit from the woman over the phone? List FIVE.

CFPC Answer Key:

Child's weight

Child's past medical history

Time of ingestion

Type of exposure / Name of the product ingested

Amount of exposure

Category of Question: Difficult

GPT-3.5 Answers:

ROUND 1

Name and age of child.

Name of ingested substance.

Quantity and time of ingestion.

Any symptoms or changes observed.

Contact information and location.

ROUND 2

Substance ingested,

amount,

time,

child's weight,

symptoms.

ROUND 3

Substance ingested,

amount,

time,

child's weight,

symptoms.

ROUND 4

Ingested substance,

amount,

time,

symptoms,

medical history.

ROUND 5

In a situation like this, it's crucial to gather as much relevant information as possible to assess the severity of the situation and provide appropriate guidance. Here are five important initial pieces of information to elicit from the woman over the phone:

1. Substance Ingested: Ask the woman to identify the specific poison or substance that the child ingested. This will help determine the potential toxicity and appropriate course of action.
2. Amount Ingested: Inquire about the approximate quantity or dose of the substance ingested. The amount consumed can affect the severity of potential symptoms.
3. Time of Ingestion: Determine when the child ingested the poison. The timing is essential for assessing the urgency of the situation and deciding on appropriate intervention timelines.

4. Any Symptoms: Ask the woman if the child is showing any signs of symptoms, such as vomiting, nausea, dizziness, or any other unusual behavior. This information will help gauge the current impact of the ingested substance.

5. Child's Medical History: Inquire about any pre-existing medical conditions, allergies, or medications the child might be taking. This information can influence how the body responds to the ingested substance.

Remember, these initial pieces of information will help you make a preliminary assessment of the situation and guide your decision-making process regarding the next steps for the child's care.

Reviewers rationale for CFPC and Reviewer's Score:

ROUND 1:

- 'Quantity and time of ingestion" are given in the CFPC answer key as: "Amount of exposure" and "Time of ingestion" (2 CFPC score, 2 Reviewers' Score)
- "Name and age of child" is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer's Score)
- "Name of ingested substance" is given in the CFPC answer key as "Name of the product ingested" (1 CFPC Score, 1 Reviewer's Score)
- "Any symptoms or changes observed" is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer's Score)
- "Contact information and location" is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer's Score)

Conclusion:

CFPC Score to this Round is 3.

Reviewers' Score to this Round is 5.

ROUND 2

- "Substance ingested" is given in the CFPC answer key as "Name of the product ingested" (1 CFPC Score, 1 Reviewer's Score)

- “amount” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)

- “time” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)

- “child's weight” is given in the CFPC answer key as “Child's weight” (1 CFPC score, 1 Reviewers’ Score)

- “symptoms” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)

Conclusion:

CFPC Score to this Round is 4.

Reviewers’ Score to this Round is 5.

ROUND 3

- “Substance ingested” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)

- “amount” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)

- “time” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)

- “child's weight” is given in the CFPC answer key as “Child's weight” (1 CFPC score, 1 Reviewers’ Score)

- “symptoms” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)

Conclusion:

CFPC Score to this Round is 4.

Reviewers’ Score to this Round is 5.

ROUND 4

- “Ingested substance” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)

- “amount” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)
- “time” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)
- “symptoms” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)
- “medical history” is given in the CFPC answer key as “Child's past medical history” (1 CFPC score, 1 Reviewers’ Score)

Conclusion:

CFPC Score to this Round is 4.

Reviewers’ Score to this Round is 5.

ROUND 5

- “Substance Ingested” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)
- “Amount Ingested” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)
- “Time of Ingestion” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)
- “Any Symptoms” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)
- “Child's Medical History” is given in the CFPC answer key as “Child's past medical history” (1 CFPC score, 1 Reviewers’ Score)

Conclusion:

CFPC Score to this Round is 4.

Reviewers’ Score to this Round is 5.

Summery of Scoring GPT-3.5

Total=5	**ROUND-1**	**ROUND-2**	**ROUND-3**	**ROUND-4**	**ROUND-5**
CFPC Score	3	4	4	4	4

Reviewers' score	5	5	5	5	5
Repeated answer	4 R2 4 R3 4 R4 3 R5	4 R1 5 R3 5 R4 4 R5	4 R1 5 R2 5 R4 4 R5	4 R1 5 R2 5 R3 5 R5	3 R1 4 R2 4 R3 5 R4

GPT-4 Answers:

ROUND 1

Substance ingested.

Quantity ingested.

Time of ingestion.

Any observed symptoms.

Child's current weight.

ROUND 2

Substance ingested and its concentration.

Amount of substance ingested.

Time since ingestion.

Any symptoms observed.

Past medical history of the child.

ROUND 3

Type of poison ingested.

Quantity ingested.

Time since ingestion.

Presence of any symptoms.

Any prior interventions/treatments given.

ROUND 4

Substance ingested.

Time of ingestion.

Amount ingested.

Weight of child.

Pre-existing medical conditions.

ROUND 5

I'm not a healthcare professional, but in a situation where a child has ingested a poisonous substance, timely and accurate information is crucial for proper medical intervention. Here are five important pieces of information you might want to elicit from the mother over the phone:

Type of Poison: Knowing what the child has ingested is crucial. Is it a household cleaning agent, a medication, a plant, or some other toxic substance? The specific type of poison will determine the appropriate course of treatment.

Amount and Concentration: Understanding the approximate quantity and concentration of the ingested poison will help in assessing the severity of the situation and planning appropriate medical intervention.

Time of Ingestion: How much time has elapsed since the child ingested the poison? This information will help medical professionals to gauge how urgently medical intervention is needed, and whether there is still a window for certain treatments, like activated charcoal, to be effective.

Current Symptoms: Although the child appears well at the moment, any subtle changes in behavior, level of consciousness, breathing, or other physical symptoms need to be noted. This can help assess whether the poisoning is progressing or has reached a certain level of severity.

Child's Weight: The child's weight can influence how the poison affects them. Certain antidotes or medical interventions may also be weight-based. Having this information can help in deciding both the urgency and nature of medical treatment.

Prompt assessment and relay of this information can significantly aid in rapid and effective medical treatment. Always consult with healthcare professionals for accurate diagnosis and appropriate management.

Reviewers rational for CFPC and Reviewer's Score

ROUND 1:

- “Substance ingested” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)
- “Quantity ingested” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)
- “Time of ingestion” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)
- “Any observed symptoms” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)
- “Child's current weight” is given in the CFPC answer key as “Child's weight” (1 CFPC score, 1 Reviewers’ Score)

Conclusion:

CFPC Score to this Round is 4.

Reviewers’ Score to this Round is 5.

ROUND 2

- “Substance ingested and its concentration” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)
- “Amount of substance ingested” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)
- “Time since ingestion” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)
- “Any symptoms observed” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)
- “Past medical history of the child” is given in the CFPC answer key as “Child's past medical history” (1 CFPC score, 1 Reviewers’ Score)

Conclusion:

CFPC Score to this Round is 4.

Reviewers’ Score to this Round is 5.

ROUND 3

- “Type of poison ingested” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)
- “Quantity ingested” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)
- “Time since ingestion” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)
- “Presence of any symptoms” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)
- “Any prior interventions/treatments given” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)

Conclusion:

CFPC Score to this Round is 3.

Reviewers’ Score to this Round is 5.

ROUND 4

- “Substance ingested” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)
- “Time of ingestion” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)
- “Amount ingested” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)
- “Weight of child” is given in the CFPC answer key as: “Child's weight” (1 CFPC Score, 1 Reviewer’s Score)
- “Pre-existing medical conditions” is given in the CFPC answer key as: “Child's past medical history” (1 CFPC Score, 1 Reviewer’s Score)

Conclusion:

CFPC Score to this Round is 5.

Reviewers’ Score to this Round is 5.

ROUND 5

- “Type of Poison” is given in the CFPC answer key as “Name of the product ingested” (1 CFPC Score, 1 Reviewer’s Score)

- “Amount and Concentration” is given in the CFPC answer key as: “Amount of exposure” (1 CFPC Score, 1 Reviewer’s Score)
- “Time of Ingestion” is given in the CFPC answer key as “Time of ingestion” (1 CFPC score, 1 Reviewers’ Score)
- “Current Symptoms” is not in the in the CFPC answer Key but is a correct answer (0 CFPC Score, 1 Reviewer’s Score)
- “Child’s Weight” is given in the CFPC answer key as: “Child's weight” (1 CFPC Score, 1 Reviewer’s Score)

Conclusion:

CFPC Score to this Round is 4.

Reviewers’ Score to this Round is 5.

Summery of Scoring GPT-4

Total Possible Scores= 5	ROUND-1	ROUND-2	ROUND-3	ROUND-4	ROUND-5
CFPC Score	4	4	3	5	4
Reviewers’ Score	5	5	5	5	5
Repeated Answer	4 R2 3 R3 4 R4 5 R5	4 R1 4 R3 4 R4 4 R5	3 R1 4 R2 3 R4 3 R5	4 R1 4 R2 3 R3 4 R5	5 R1 4 R2 3 R3 4 R4

REFERENCES

1. Dartmouth College. Artificial Intelligence Coined at Dartmouth

1956 [Available from: https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth.

2. Chan KS, Zary N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. JMIR Med Educ. 2019;5(1):e13930.
3. Kueper JK. Primer for artificial intelligence in primary care. Can Fam Physician. 2021;67(12):889-93.
4. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, et al. Application of Artificial Intelligence in Community-Based Primary Health Care: Systematic Scoping Review and Critical Appraisal. J Med Internet Res. 2021;23(9):e29839.
5. Aamir A, Iqbal A, Jawed F, Ashfaque F, Hafsa H, Anas Z, et al. Exploring the current and prospective role of artificial intelligence in disease diagnosis. Ann Med Surg (Lond). 2024;86(2):943-9.
6. Sathe N, Deodhe V, Sharma Y, Shinde A. A Comprehensive Review of AI in Healthcare: Exploring Neural Networks in Medical Imaging, LLM-Based Interactive Response Systems, NLP-Based EHR Systems, Ethics, and Beyond.
7. Chu Z, Ni S, Wang Z, Feng X, Li C, Hu X, et al. History, Development, and Principles of Large Language Models-An Introductory Survey. arXiv preprint arXiv:240206853. 2024.
8. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the Use of Artificial Intelligence in Primary Care. Ann Fam Med. 2022;20(6):559-63.
9. Nazi ZA, Peng W. Large language models in healthcare and medical domain: A review. arXiv preprint arXiv:240106775. 2023.
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40.
11. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-8.
12. Touvron H ML, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 19 Jul 2023.
13. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. Journal of Machine Learning Research. 2024;25(70):1-53.
14. IBM, Holdsworth J. What is NLP (natural language processing)? June 2024 [Available from: https://www.ibm.com/topics/natural-language-processing.
15. IBM. What is few-shot learning? [Available from: https://www.ibm.com/topics/few-shot-learning.
16. DeLair G. Large Language Models: Open Source vs. Closed Source February 20, 2024 · [Available from: https://echobasedev.com/large-language-models-open-source-vs-closed-source/.
17. Malec M. Open-Source LLMs vs Closed: Unbiased 2024 Guide for Innovative Companies May 31, 2024 [Available from: https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/.
18. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.

19. Zapier. The best large language models (LLMs) in 2024 2024 [Available from: https://zapier.com/blog/best-llm/.
20. IBM G, M. What is an API (application programming interface)? 09 April 2024 [Available from: https://www.ibm.com/topics/api.
21. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Annals of Internal Medicine. 2024.
22. Al-Amin M, Ali MS, Salam A, Khan A, Ali A, Ullah A, et al. History of generative Artificial Intelligence (AI) chatbots: past, present, and future development. arXiv preprint arXiv:240205122. 2024.
23. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research. 2023;24(240):1-113.
24. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. Palm 2 technical report. arXiv preprint arXiv:230510403. 2023.
25. Meta. Introducing Meta Llama 3: The most capable openly available LLM to date April 18, 2024 [Available from: https://ai.meta.com/blog/meta-llama-3/.
26. Hsiao S. Get more done with Gemini: Try 1.5 Pro and more intelligent features

May 14, 2024 [Available from: https://blog.google/products/gemini/google-gemini-update-may-2024/.
27. Inflection AI. Pi AI Feb 2024 [Available from: https://pi.ai/talk.
28. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388(13):1233-9.
29. Lo CK. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences. 2023;13(4):410.
30. Hsiao S. Bard becomes Gemini: Try Ultra 1.0 and a new mobile app today Feb 08, 2024 [Available from: https://blog.google/products/gemini/bard-gemini-advanced-app/.
31. Antropic. Introducing the next generation of Claude Mar 4, 2024 [Available from: https://www.anthropic.com/news/claude-3-family.
32. Banderson K. Pi AI Review - The Chatbot with a Twist: MS Power; February 6, 2024 [Available from: https://mspoweruser.com/pi-ai-review/.
33. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Me LLaMA: Foundation Large Language Models for Medical Applications. arXiv preprint arXiv:240212749. 2024.
34. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas DdL, et al. Mistral 7B. ArXiv. 10 Oct 2023;abs/2310.06825.
35. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digital Medicine. 2023;6(1):210.
36. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. arXiv preprint arXiv:230512031. 17 Aug 2023.
37. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:230414454. July 3, 2023.
38. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:231116079. Nov 27, 2023.

39. Labrak Y, Bazoge A, Morin E, Gourraud P-A, Rouvier M, Dufour R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv preprint arXiv:240210373. 2024.
40. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
41. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:230509617. 2023.
42. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. arXiv preprint arXiv:230609968. 16 Jun 2023.
43. Adamopoulou E, Moussiades L. An Overview of Chatbot Technology. Artificial Intelligence Applications and Innovations. 2020;584:373-83.
44. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2023.
45. OpenAI blog. Feb 2024. Available from: https://platform.openai.com/docs/.
46. OpenAI. ChatGPT [Large language model]. Accessed Feb 2024. Last Update Apr 2023. Feb 2024 [Available from: https://chat.openai.com.
47. You.com blog. The New UI On You.com Has Arrived — Here's How To Get Started February 10, 2024 [Available from: https://about.you.com/the-new-ui-on-youdotcom-has-arrived-get-started/.
48. Microsoft blog. July 18, 2023. [cited Feb 2024]. Available from: https://blogs.microsoft.com/blog/2023/07/18/furthering-our-ai-ambitions-announcing-bing-chat-enterprise-and-microsoft-365-copilot-pricing/.
49. You.com. Youchat Feb 2024 [Available from: https://you.com/?chatMode=default.
50. Open AI. Hello GPT-4o May 13, 2024 [Available from: https://openai.com/index/hello-gpt-4o/.
51. Garg S. 30 Best ChatGPT Alternatives in 2024 (Free & Paid): Writesonic; Feb 12, 2024 [Available from: https://writesonic.com/blog/chatgpt-alternatives.
52. Garg S. 17 Best ChatGPT Alternatives in 2024 (Free & Paid) May 20, 2024 [Available from: https://writesonic.com/blog/chatgpt-alternatives.
53. You.com. You.com is the go-to platform to test and compare the latest AI models February 10, 2024 [Available from: https://about.you.com/youdotcom-is-the-go-to-platform-to-test-and-compare-different-ai-models/.
54. Gemini Team Google, Anil R, Borgeaud S, Wu Y, Alayrac J-B, Yu J, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023.
55. Gordon C. How Docus.ai Is Powering A Groundbreaking AI Powered Health Platform: Forbes; Jun 12, 2023 [Available from: https://www.forbes.com/sites/cindygordon/2023/06/12/how-docusai-is-powering-a-groundbreaking-ai-powered-health-platform/?sh=6fa600676fdd.
56. Stojan J. Docus.ai: Innovating Global Healthcare through AI-Powered Diagnostics: USA today; Aug 12, 2023 [Available from: https://www.usatoday.com/story/special/contributor-content/2023/08/22/docus-ai-innovating-global-healthcare-through-ai-powered-diagnostics/70653228007/.
57. Atchinson BK, Fox DM. The politics of the Health Insurance Portability and Accountability Act. Health Aff (Millwood). 1997;16(3):146-50.

58. GDPR. What is GDPR, the EU's new data protection law? [Available from: https://gdpr.eu/what-is-gdpr/.
59. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci. 2023;39(2):605-7.
60. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107-e8.
61. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6).
62. Li J, Dada A, Kleesiek J, Egger J. ChatGPT in Healthcare: A Taxonomy and Systematic Review (Preprint). medRxiv. 2023:2023.03. 30.23287899.
63. The Lancet Digital H. ChatGPT: friend or foe? Lancet Digit Health. 2023;5(3):e102.
64. Lubowitz JH. ChatGPT, An Artificial Intelligence Chatbot, Is Impacting Medical Literature. Arthroscopy. 2023;39(5):1121-2.
65. OpenAI blog. Models: OpenAI; 2023 [Available from: https://beta.openai.com/docs/models.
66. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health. 2023;5(3):e105-e6.
67. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023;307(2):e230163.
68. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023;11:1166120.
69. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci. 2023;10(8):230658.
70. ICMJE. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Updated May 2023: International Committee of Medical Journal Editors (ICMJE). Accessed Nov 2023; [Available from: https://www.icmje.org/icmje-recommendations.pdf.
71. Zielinski C, Winker M, Aggarwal R, Ferris L, Heinemann M, Lapeña JF, et al. Chatbots, ChatGPT, and Scholarly Manuscripts: WAME Recommendations on ChatGPT and Chatbots in relation to scholarly publications. Natl Med J India. 2023;36(1):1-4.
72. Moy L. Guidelines for Use of Large Language Models by Authors, Reviewers, and Editors: Considerations for Imaging Journals. Radiology. 2023;309(1):e239024.
73. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Van Phuoc L, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. Journal of the American Medical Informatics Association. 2024;31(3):776-83.
74. Wang LK, Paidisetty PS, Cano AM. The next paradigm shift? ChatGPT, artificial intelligence, and medical education. Med Teach. 2023;45(8):925.
75. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. 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(4).
76. Benoit JRA. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation. medRxiv. 2023:2023.02.04.23285478.

77. Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME). USMLE 2024 [Available from: https://www.usmle.org/.
78. Medical Council of Canada. What is the MCCQE Part I? 2024 [Available from: https://mcc.ca/examinations-assessments/mccqe-part-i/.
79. Hsain A, Housni HE. Large language model-powered chatbots for internationalizing student support in higher education. arXiv preprint arXiv:240314702. 2024.
80. 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.
81. Bresge A. ChatGPT 101: The risks and rewards of generative AI in the classroom: University of Toronto; [updated September 13, 2023. Available from: https://www.utoronto.ca/news/chatgpt-101-risks-and-rewards-generative-ai-classroom.
82. AAFP. The Specialty of Family Medicine [Available from: The Specialty of Family Medicine.
83. Canadian Medical Assosiation. Family Medicine Profile; Updated Dec 2019. Available from: https://www.cma.ca/. 2019.
84. The College of Family Physicians of Canada. Preparing for the Certification Examination in Family Medicine 2023 [Available from: https://www.cfpc.ca/en/education-professional-development/examinations-and-certification/certification-examination-in-family-medicine/preparing-for-the-certification-examination-in-fam.
85. Mousavi M, Shafiee S, Harley JM, Cheung J, Rahimi SA. Performance of ChatGPT in Certification Examination of College of Family Physicians of Canada, doi: 10.1136/fmch-2023-002626. PMID: 38806403; PMCID: PMC11138270. Fam Med Community Health. 2024 May 28(12 (Suppl 1)):e002626.
86. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.
87. Liu J, Zhou P, Hua Y, Chong D, Tian Z, Liu A, et al. Benchmarking Large Language Models on CMExam-A Comprehensive Chinese Medical Exam Dataset. Advances in Neural Information Processing Systems. 2024;36.
88. Pal A, Sankarasubbu M. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. arXiv preprint arXiv:240207023. 2024.
89. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns. 2024;5(3).
90. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023;15(6):e40895.
91. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ. 2022;22(1):772.
92. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
93. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, 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.

94. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. Acad Med. 2021;96(11s):S62-s70.
95. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. 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.
96. Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca--an open-source collection of medical conversational AI models and training data. arXiv preprint arXiv:230408247. 2023.
97. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. Cureus. 2023;15(6):e40822.
98. Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. Am J Gastroenterol. 2023.
99. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology. 2023;307(5):e230582.
100. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, 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.
101. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ. 2023;9:e48002.
102. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination (preprint): Research Square; 2023 [updated 2023. Available from: http://europepmc.org/abstract/PPR/PPR617702.
103. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, 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.
104. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation. 2023;185:109732.
105. 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(1):e000451.
106. Pal A, Minervini P, Motzfeldt AG, Gema AP, Alex B. The Open Medical-LLM Leaderboard: Benchmarking Large Language Models in Healthcare: Hugging Face; Apr 19, 2024 [Available from: https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard.
107. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589-96.
108. Goodman RS, Patrinely JR, Stone CA, Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Network Open. 2023;6(10):e2336483-e.

109. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. npj Digital Medicine. 2024;7(1):20.
110. Yan S, Knapp W, Leong A, Kadkhodazadeh S, Das S, Jones VG, et al. Prompt engineering on leveraging large language models in generating response to InBasket messages. Journal of the American Medical Informatics Association. 2024.