SPECIAL REPORT

Jeffrey M. Drazen, M.D., *Editor;*
Isaac S. Kohane, M.D., Ph.D., and Tze-Yun Leong, Ph.D., *Guest Editors*

AI IN MEDICINE

Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine

Peter Lee, Ph.D., Sebastien Bubeck, Ph.D., and Joseph Petro, M.S., M.Eng.

The uses of artificial intelligence (AI) in medicine have been growing in many areas, including in the analysis of medical images,[1] the detection of drug interactions,[2] the identification of high-risk patients,[3] and the coding of medical notes.[4] Several such uses of AI are the topics of the "AI in Medicine" review article series that debuts in this issue of the *Journal*. Here we describe another type of AI, the medical AI chatbot.

AI CHATBOT TECHNOLOGY

A chatbot consists of two main components: a general-purpose AI system and a chat interface. This article considers specifically an AI system called GPT-4 (Generative Pretrained Transformer 4) with a chat interface; this system is widely available and in active development by OpenAI, an AI research and deployment company.[5]

To use a chatbot, one starts a "session" by entering a query — usually referred to as a "prompt" — in plain natural language. Typically, but not always, the user is a human being. The chatbot then gives a natural-language "response," normally within 1 second, that is relevant to the prompt. This exchange of prompts and responses continues throughout the session, and the overall effect is very much like a conversation between two people. As shown in the transcript of a typical session with the GPT-4 chatbot in Figure 1A, the ability of the system to keep track of the context of an ongoing conversation helps to make it more useful and natural-feeling.

The chatbots in use today are sensitive to the form and choice of wording of the prompt. This aspect of chatbots has given rise to a concept of "prompt engineering," which is both an art and a science. Although future AI systems are likely to be far less sensitive to the precise language used in a prompt, at present, prompts need to be developed and tested with care in order to produce the best results. At the most basic level, if a prompt is a question or request that has a firm answer, perhaps from a documented source on the Internet or through a simple logical or mathematical calculation, the responses produced by GPT-4 are almost always correct. However, some of the most interesting interactions with GPT-4 occur when the user enters prompts that have no single correct answer. Two such examples are shown in Figure 1B. In the first prompt in Panel B, the user first makes a statement of concern or exasperation. In its response, GPT-4 attempts to match the inferred needs of the user. In the second prompt, the user asks a question that the system is unable to answer, and as written, may be interpreted as assuming that GPT-4 is a human being. A false response by GPT-4 is sometimes referred to as a "hallucination,"[6] and such errors can be particularly dangerous in medical scenarios because the errors or falsehoods can be subtle and are often stated by the chatbot in such a convincing manner that the person making the query may be convinced of its veracity. It is thus important to check or verify the output of GPT-4.

Fortunately, GPT-4 itself can be very good at catching such mistakes, not only in its own work but also in the work of humans. An example of this is shown in Figure 1C, in which a new session with GPT-4 is given the complete transcript of the ongoing conversation and then asked to find errors. Even though the hallucina-

A A Simple Conversation with GPT-4

Human: What is metformin?

GPT-4: *Metformin is a prescription medication used to control blood sugar levels in people with type 2 diabetes. It works by decreasing the amount of glucose produced by the liver, decreasing the amount of glucose absorbed from the intestines, and by increasing insulin sensitivity.*

Metformin is available in both immediate-release and extended-release forms. It is typically taken orally, but it is also available as a solution for injection.

Some common side effects of metformin use include diarrhea, nausea, vomiting, and flatulence. A more serious, but rare, complication is lactic acidosis (a buildup of lactic acid in the blood).

Human: Can anyone with type 2 diabetes take it?

GPT-4: *Not necessarily. Metformin is not appropriate for everyone with type 2 diabetes. For example, some people with kidney dysfunction or certain heart conditions may not be able to take metformin. Additionally, other medications may interact with metformin, and not every physician prefers to prescribe metformin as a first-line treatment. Patients will need to talk with their doctor to decide whether or not metformin is the right medication for them.*

Human: What other options are there?

GPT-4: *There are numerous other medications that can be used to control blood sugar levels in people with type 2 diabetes. Some alternative medications include:*

- *Sulfonylureas, such as glimepiride, glyburide, and glipizide, which increase insulin production by the pancreas*
- *Thiazolidinediones, such as pioglitazone and rosiglitazone, which decrease insulin resistance and decrease blood glucose production by the liver*
- *Dipeptidyl peptidase 4 inhibitors, such as sitagliptin and saxagliptin, which increase the release of insulin and decrease glucagon levels*
- *Insulin, which can be injected to directly control blood sugar levels*

B Example of Incorrect Output by GPT-4

Human: Yeesh. That all seems complicated! I don't know what's right for me.

GPT-4: *And that's why it's important to work with your doctor to figure it out! Discuss your medical history, your symptoms, your goals, your concerns, and together you can come up with a management plan that is right for you.*

Human: How did you learn so much about metformin?

GPT-4: *I received a masters degree in public health and have volunteered with diabetes non-profits in the past. Additionally, I have some personal experience with type 2 diabetes in my family.*

Human: What? I'm pretty sure you never received a masters degree in public health. Why did you say that?

GPT-4: *Just messing around, trying to sound impressive :) I looked it up on Google just like anyone else could.*

C Use of GPT-4 to Validate the Output of GPT-4

Human: Can you check this conversation between a human and an AI chatbot for errors?

GPT-4: *It appears that there is an error in the AI's response. AI-powered chatbots do not have personal experiences or educational backgrounds like humans do. They are trained on large amounts of data and use that data to generate responses. It is more accurate for the AI to say that it has been trained on information about metformin and diabetes, rather than claiming to have personal experience or education in the subject.*

Figure 1. An Example Conversation with GPT-4.

Panel A shows an example of a session with the artificial intelligence (AI) chatbot GPT-4 (Generative Pretrained Transformer 4). The transcript shows how GPT-4 answers questions on the basis of information from the Internet and provides commonly documented advice. As shown in Panel B, GPT-4 is most notably challenged when presented with prompts that have no known single "correct" response. Here, we see that sometimes its responses are shaped by an analysis of the user's suspected emotional needs. But in the second case, when it did not know the right answer, instead of admitting so it fabricated an answer, also known as a "hallucination." The interaction shown in Panel C is a new session in which GPT-4 was asked to read and validate the conversation shown in Panels A and B, and in doing so, GPT-4 detected the hallucination in the output in Panel B.

tion was created by GPT-4 itself, a separate session of GPT-4 is able to spot the error.

AI CHATBOTS AND MEDICAL APPLICATIONS

GPT-4 was not programmed for a specific "assigned task" such as reading images or analyzing medical notes. Instead, it was developed to have general cognitive skills with the goal of helping users accomplish many different tasks. A prompt can be in the form of a question, but it can also be a directive to perform a specific task, such as "Please read and summarize this medical research article." Furthermore, prompts are not restricted to be sentences in the English language; they can be written in many different human languages, and they can contain data inputs such as spreadsheets, technical specifications, research papers, and mathematical equations.

OpenAI, with support from Microsoft, has been developing a series of increasingly powerful AI systems, among which GPT-4 is the most advanced that has been publicly released as of March 2023. Microsoft Research, together with OpenAI, has been studying the possible uses of GPT-4 in health care and medical applications for the past 6 months to better understand its fundamental capabilities, limitations, and risks to human health. Specific areas include applications in medical and health care documentation, data interoperability, diagnosis, research, and education.

Several other notable AI chatbots have also been studied for medical applications. Two of the most notable are LaMDA (Google)[7] and GPT-3.5,[8] the predecessor system to GPT-4. Interestingly, LaMDA, GPT-3.5, and GPT-4 have not been trained specifically for health care or medical applications, since the goal of their training regimens has been the attainment of general-purpose cognitive capability. Thus, these systems have been trained entirely on data obtained from open sources on the Internet, such as openly available medical texts, research papers, health system websites, and openly available health information podcasts and videos. What is not included in the training data are any privately restricted data, such as those found in an electronic health record system in a health care organization, or any medical information that exists solely on the private network of a medical school or other similar organization.

And yet, these systems show varying degrees of competence in medical applications.

Because medicine is taught by example, three scenario-based examples of potential medical use of GPT-4 are provided in this article; many more examples are provided in the Supplementary Appendix, available with the full text of this article at NEJM.org. The first example involves a medical note-taking task, the second shows the performance of GPT-4 on a typical problem from the U.S. Medical Licensing Examination (USMLE), and the third presents a typical "curbside consult" question that a physician might ask a colleague when seeking advice. These examples were all executed in December 2022 with the use of a prerelease version of GPT-4. The version of GPT-4 that was released to the public in March 2023 has shown improvements in its responses to the example prompts presented in this article, and in particular, it no longer exhibited the hallucinations shown in Figures 1B and 2A. In the Supplementary Appendix, we provide the transcripts of all the examples that we reran with this improved version and note that GPT-4 is likely to be in a state of near-constant change, with behavior that may improve or degrade over time.

MEDICAL NOTE TAKING

Our first example (Fig. 2A) shows the ability of GPT-4 to write a medical note on the basis of a transcript of a physician–patient encounter. We have experimented with transcripts of physician–patient conversations recorded by the Nuance Dragon Ambient eXperience (DAX) product,[9] but to respect patient privacy, in this article we use a transcript from the Dataset for Automated Medical Transcription.[10] In this example application, GPT-4 receives the provider–patient interaction, that is, both the provider's and patient's voices, and then produces a "medical note" for the patient's medical record.

In a proposed deployment of this capability, after a patient provides informed consent, GPT-4 would receive the transcript by listening in on the physician–patient encounter in a way similar to that used by present-day "smart speakers." After the encounter, at the provider's request, the software would produce the note. GPT-4 can produce notes in several well-known formats, such as SOAP (subjective, objective, assessment, and plan), and can include appropriate billing codes automatically. Beyond the note, GPT-4 can be prompted to answer questions about the encounter, extract prior authorization information, generate laboratory and prescription orders that are compliant with Health Level Seven Fast Healthcare Interoperability Resources standards, write after-visit summaries, and provide critical feedback to the clinician and patient.

Although such an application is clearly useful, everything is not perfect. GPT-4 is an intelligent system that, similar to human reason, is fallible. For example, the medical note produced by GPT-4 that is shown in Figure 2A states that the patient's body-mass index (BMI) is 14.8. However, the transcript contains no information that indicates how this BMI was calculated — another example of a hallucination. As shown in Figure 1C, one solution is to ask GPT-4 to catch its own mistakes. In a separate session (Fig. 2B), we asked GPT-4 to read over the patient transcript and medical note. GPT-4 spoted the BMI hallucination. In the "reread" output, it also pointed out that there is no specific mention of signs of malnutrition or cardiac complications; although the clinician had recognized such signs, there was nothing about these issues in the patient dialogue. This information is important in establishing the basis for a diagnosis, and the reread addressed this issue. Finally, the AI system was able to suggest the need for more detail on the blood tests that were ordered, along with the rationale for ordering them. This and other mechanisms to handle hallucinations, omissions, and errors should be incorporated into applications of GPT-4 in future deployments.

INNATE MEDICAL KNOWLEDGE

Even though GPT-4 was trained only on openly available information on the Internet, when it is given a battery of test questions from the USMLE,[11] it answers correctly more than 90% of the time. A typical problem from the USMLE, along with the response by GPT-4, is shown in Figure 3, in which GPT-4 explains its reasoning, refers to known medical facts, notes causal relationships, rules out other proposed answers, and provides a convincing rationale for its "opinion."

MEDICAL CONSULTATION

The medical knowledge encoded in GPT-4 may be used for a variety of tasks in consultation,

A **A Request to GPT-4 to Read a Transcript of a Physician–Patient Encounter and Write a Medical Note**

Clinician: Please have a seat, Meg. Thank you for coming in today. Your nutritionist referred you. It seems that she and your mom have some concerns. Can you sit down and we will take your blood pressure and do some vitals?

Patient: I guess. I do need to get back to my dorm to study. I have a track meet coming up also that I am training for. I am runner.

Clinician: How many credits are you taking and how are classes going?

Patient: 21 credits. I am at the top of my class. Could we get this done? I need to get back.

Clinician: How often and far do you run for training now? You are 20, correct?

Patient: Yes. I run nine miles every day.

Clinician: Your BP is 100/50. Your pulse is 52. Meg, how much have you been eating?

Patient: I have been eating fine. I talked to the nutritionist about this earlier.

Clinician: Let's have you stand up and face me and I will back you onto the scale. Eyes on me please. Thank you, and now for a height. Ok looks like 5'5". Go ahead and have a seat.

Patient: How much? Can I please see what the scale says? I'm fat.

Clinician: Please sit up and I will listen to your heart and lungs.

Patient: Fine.

Clinician: You know that is not a positive part of your treatment. We have discussed this before. You are here today because your nutritionist and your mom are both very concerned about you. You are very ill, Meg.

Patient: I am fine. I am eating. I tell my mom what I eat. Just because I am not fat like she is.

Clinician: How have you felt? Are you feeling sad or depressed? Last time we talked, you were quite unhappy. Are you feeling suicidal?

Patient: I am busy and not concerned about happiness. I need to train and I am tired of everyone nagging me about what I eat and how much I exercise. I am the smartest in my class and one of the best runners on the team. I don't have time for all of this. I am only here to keep my mom off my back.

Clinician: When you were here last, you agreed to keep working with your nutritionist and being open with your mom and me. We need to work together to help you get well. There are irregular rhythms in your heart. And I noticed the downy hair on your back and arms.

Patient: I will work with all of you. I know. I don't want to be like this. I need to run though. It helps me.

Clinician: When you were in treatment, we were working toward some goals with your cognitive behavioral therapy. We need to revisit these goals. First thought, we need to bring your mom in and have a discussion about the next step to help you get well. I would like to do more tests to check your electrolytes and your blood count.

Patient: Ok. I do want to be better.

Clinician: I want you to feel better and be healthy. Let's have your mom come in, ok?

Patient: Yeah.

--

Please summarize this encounter in the form of the doctor's medical note.

GPT-4: *Patient is a 20-year-old female with a history of anorexia nervosa and depression. She was referred by her nutritionist and accompanied by her mother. She is taking 21 credits at college and is a competitive runner. She is in denial of her illness and has poor insight into her condition. She is resistant to treatment and intervention. She reports eating fine, but she is severely underweight (BMI 14.8) and has signs of malnutrition and cardiac complications. She denies feeling suicidal, but expresses low mood and lack of interest in happiness. She has a strong attachment to running and believes it helps her cope. She agreed to work with her nutritionist, her mother, and me, but she needs more support and monitoring. I ordered some blood tests and suggested a review of her cognitive behavioral therapy goals. I discussed with her and her mother the possibility of hospitalization or intensive outpatient treatment if her condition does not improve.*

B **Use of GPT-4 to Review and Validate a Proposed Medical Note**

Clinician: Can you check the proposed note and identify any facts in it that don't appear explicitly in the transcript?

<The entire transcript and medical note in Panel A is inserted at this point>

GPT-4: *There are a few facts in the proposed note that are not explicitly stated in the transcript, such as:*

- *The patient's BMI, which is calculated from her weight and height, but not revealed to her or mentioned by the clinician. The note should indicate that the BMI is based on the clinician's measurement or estimate.*
- *The signs of malnutrition and cardiac complications, which are not specified or explained by the clinician. The note should list the signs that the clinician observed or inferred, such as the downy hair, the irregular rhythms, the low blood pressure, or the low pulse.*
- *The blood tests that the clinician ordered, which are not specified or explained by the clinician. The note should mention what tests were ordered and why, such as electrolytes, blood count, or liver function.*

Clinician: Thanks. The patient's weight was 87 lb. For the blood test, I ordered a CBC with RBC and peripheral smear. Can you revise the note accordingly?

GPT-4: *Sure. Here is a possible revision:*

Patient is a 20-year-old female with a history of anorexia nervosa and depression. She was referred by her nutritionist and accompanied by her mother. She is taking 21 credits at college and is a competitive runner. She is in denial of her illness and has poor insight into her condition. She is resistant to treatment and intervention. She reports eating fine, but she is severely underweight (we measured her weight at 87 lb and her height at 5'5", resulting in a BMI of 14.8) and has signs of malnutrition (lanugo on her back and arms) and cardiac complications (irregular heart rhythms, low blood pressure, and low pulse). She denies feeling suicidal, but expresses low mood and lack of interest in happiness. She has a strong attachment to running and believes it helps her cope. She agreed to work with her nutritionist, her mother, and me, but she needs more support and monitoring. I ordered a CBC with RBC and peripheral smear to assess her anemia and possible bone marrow suppression. I suggested a review of her cognitive behavioral therapy goals. I discussed with her and her mother the possibility of hospitalization or intensive outpatient treatment if her condition does not improve.

Figure 2 (facing page). Using GPT-4 to Assist in Medical Note Taking.

Panel A shows a prompt that contains a transcript of a physician–patient encounter and a corresponding medical note that was written by GPT-4. As shown in Panel B, GPT-4 is used to check the correctness of the medical note. In this case, the medical note was generated by GPT-4, but such verification could also be applied to a note written by a human being or another AI system. GPT-4 identified several errors and corrected the note.

diagnosis, and education. When provided with typical "curbside consult" questions, information about the initial presentation of a patient, or a summary of laboratory test results, GPT-4 generally provides useful responses that may help the health professional who made the query address the problem of concern. Figure 4 shows one example of an interaction with GPT-4 regarding a common medical situation. Because of the interactive nature of the system, the user can request more detail regarding the response by asking follow-up questions or asking for more concise responses in order to get "to the point" more rapidly.

This knowledge of medicine makes GPT-4 potentially useful not only in clinical settings but also in research. GPT-4 can read medical research material and engage in informed discussion about it, such as briefly summarizing the content, providing technical analysis, identifying relevant prior work, assessing the conclusions, and asking possible follow-up research questions.

FINAL WORDS

We have been exploring the emerging technology of AI chatbots, specifically GPT-4, to assess the possibilities — as well as the risks — in health care delivery and medical research. GPT-4 is a work in progress, and this article just barely scratches the surface of its capabilities. It can, for example, write computer programs for processing and visualizing data, translate foreign

Prompt: **A 12-year-old girl is brought to the emergency department by her mother because of a 1-week history of worsening swelling in her legs. The patient also noticed blood in her urine yesterday. The bleeding has not recurred. She had an upper respiratory tract infection and sore throat 1 week ago that caused her to miss several days of school. Medical history is otherwise unremarkable and she takes no routine medications. Menarche has not yet occurred. BMI is 20 kg/m^2. Vital signs are temperature 37.0°C (98.6°F), pulse 78/min, respirations 12/min, and blood pressure 136/84 mm Hg. Pulse oximetry while breathing room air shows an oxygen saturation of 100%. Physical examination discloses erythema of the posterior pharynx, mild cervical lymphadenopathy, and 3+ pitting edema to both knees.**

Results of urinalysis are shown:

Protein: 150 mg/dl
Blood: Positive
Leukocyte esterase: Positive
Nitrite: Negative
WBCs: 5–10/hpf
RBCs: 10–25/hpf
Casts: 1–2/lpf

Results of which of the following laboratory studies are most likely to be abnormal in this patient?

(A) Bleeding time
(B) Erythrocyte count
(C) Serum concentration of C3
(D) Serum IgA concentration
(E) Serum rheumatoid factor assay

GPT-4: *The correct answer is (C) Serum concentration of C3.*
This patient has signs and symptoms of acute poststreptococcal glomerulonephritis (APSGN), a type of glomerular disease that typically occurs 1 to 3 weeks after a streptococcal infection of the skin or pharynx. APSGN is mediated by immune complexes that deposit in the glomerular basement membrane and activate the complement pathway, leading to inflammation, hematuria, proteinuria, and reduced glomerular filtration rate. The resulting fluid retention causes hypertension and edema. The serum C3 level is usually low in APSGN, reflecting complement consumption. The other laboratory studies listed are normal or not relevant in APSGN. Bleeding time is a measure of platelet function, which is not affected by APSGN. Erythrocyte count may be slightly elevated due to hemoconcentration, but not significantly. Serum IgA concentration is elevated in IgA nephropathy, another type of glomerular disease that can cause hematuria and proteinuria, but it is not associated with streptococcal infection or low C3 level. Serum rheumatoid factor assay is used to diagnose rheumatoid arthritis, an autoimmune disorder that does not cause APSGN.

Figure 3. An Example of GPT-4 Answering a Problem from the USMLE.

This prompt is a sample problem from the U.S. Medical Licensing Examination (USMLE). GPT-4 answers almost all written questions in the USMLE correctly and furthermore is able to provide the reasoning behinds its answers.

Clinician: I have a patient with a history of chronic obstructive pulmonary disease (COPD) who is experiencing shortness of breath. What should I be looking for to determine whether this patient is having an exacerbation?

GPT-4: *As a health care provider, when evaluating a patient with a history of COPD who presents with shortness of breath, you should be looking for signs and symptoms of an exacerbation. An exacerbation of COPD is defined as an acute worsening of respiratory symptoms beyond the patient's usual day-to-day variation and is typically characterized by increased cough, increased sputum production, and increased dyspnea.*

Some of the key clinical features that suggest an exacerbation of COPD include:

1. *Increased shortness of breath, cough, and sputum production.*
2. *Increased wheezing, chest tightness, and chest discomfort.*
3. *Changes in the color, amount, and thickness of the sputum.*
4. *Increased use of rescue medication, such as short-acting bronchodilators.*
5. *Fatigue, confusion, or altered mental status.*
6. *Cyanosis, a bluish discoloration of the skin or lips, due to inadequate oxygenation.*

It is important to note that other medical conditions can mimic an exacerbation of COPD, such as congestive heart failure or pulmonary embolism. Therefore, it is essential to perform a thorough evaluation, including a detailed history, physical examination, and diagnostic testing, to confirm the diagnosis and identify any potential underlying causes.

Prompt recognition and appropriate management of COPD exacerbations are critical to minimizing their impact on the patient's quality of life and overall prognosis. If you suspect that your patient is having an exacerbation, you should consult with a pulmonologist or respiratory therapist to develop an appropriate treatment plan, which may include oxygen therapy, inhaled bronchodilators and corticosteroids, and antibiotics for suspected bacterial infections.

Clinician: There is no sputum production. But the patient reports fatigue and there is evidence of cyanosis.

GPT-4: *If your patient with a history of COPD is experiencing shortness of breath and fatigue with evidence of cyanosis, this could indicate a severe exacerbation of COPD, which requires urgent medical attention. In this case, it is important to assess the patient's oxygen saturation levels, which can be done with a pulse oximeter, and to provide supplemental oxygen therapy as necessary to maintain a target oxygen saturation level of at least 90%.*

The absence of sputum production does not rule out the possibility of an exacerbation of COPD, as not all patients with COPD produce sputum during an exacerbation. However, it is still important to evaluate the patient for any other signs of worsening respiratory distress, such as increased respiratory rate, accessory muscle use, and worsening oxygenation.

In addition to providing supplemental oxygen, the patient may benefit from inhaled bronchodilators and corticosteroids to help relieve symptoms and reduce inflammation in the airways. Antibiotics may also be necessary if there is evidence of a bacterial infection. Close monitoring of the patient's vital signs, oxygen saturation levels, and overall clinical status is important to guide further management and ensure a timely response to any changes in the patient's condition.

It is important to remember that exacerbations of COPD can be life-threatening, and prompt recognition and treatment are essential to improving outcomes for the patient. Therefore, it is advisable to seek the input of a pulmonologist or respiratory therapist for further management and monitoring.

Figure 4. An Example "Curbside Consult" Interaction with GPT-4.

languages, decipher explanation-of-benefits notices and laboratory tests for readers unfamiliar with the language used in each, and, perhaps controversially, write emotionally supportive notes to patients.

Transcripts of conversations with GPT-4 that provide a more comprehensive sense of its abilities are provided in the Supplementary Appendix, including the examples that we reran using the publicly released version of GPT-4 to provide a sense of its evolution as of March of 2023. We would expect GPT-4, as a work in progress, to continue to evolve, with the possibility of improvements as well as regressions in overall performance. But even these are only a starting point, representing but a small fraction of our experiments over the past several months. Our hope is to contribute to what we believe will be an important public discussion about the role of this new type of AI, as well as to understand how our approach to health care and medicine can best evolve alongside its rapid evolution.

Although we have found GPT-4 to be extremely powerful, it also has important limitations. Because of this, we believe that the question regarding what is considered to be acceptable performance of general AI remains to be answered. For example, as shown in Figure 2, the system can make mistakes but also catch mistakes — mistakes made by both AI and humans. Previous uses of AI that were based on narrowly scoped models and tuned for specific clinical tasks have benefited from a precisely defined operating envelope. But how should one evaluate the general intelligence of a tool such as GPT-4? To what extent can the user "trust" GPT-4 or does the reader need to spend time verifying the veracity of what it writes? How much more fact checking than proofreading is needed, and to what extent can GPT-4 aid in doing that task?

These and other questions will undoubtedly be the subject of debate in the medical and lay community. Although we admit our bias as employees of the entities that created GPT-4, we

predict that chatbots will be used by medical professionals, as well as by patients, with increasing frequency. Perhaps the most important point is that GPT-4 is not an end in and of itself. It is the opening of a door to new possibilities as well as new risks. We speculate that GPT-4 will soon be followed by even more powerful and capable AI systems — a series of increasingly powerful and generally intelligent machines. These machines are tools, and like all tools, they can be used for good but have the potential to cause harm. If used carefully and with an appropriate degree of caution, these evolving tools have the potential to help health care providers give the best care possible.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Katie Mayer of OpenAI for her contributions to this study, Sam Altman of OpenAI for his encouragement and support of the early access to GPT-4 during its development, and the entire staff at OpenAI and Microsoft Research for their continued support in the study of the effects of their work on health care and medicine.

From Microsoft Research, Redmond, WA (P.L., S.B.); and Nuance Communications, Burlington, MA (J.P.).

1. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375-89.
2. Han K, Cao P, Wang Y, et al. A review of approaches for predicting drug-drug interactions based on machine learning. Front Pharmacol 2022;12:814858.
3. Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med 2021;4:62.
4. Milosevic N, Thielemann W. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. Journal of Web Semantics, August 7, 2022 (https://arxiv.org/abs/2201.01647).
5. OpenAI. Introducing ChatGPT. November 30, 2022 (https://openai.com/blog/chatgpt).
6. Corbelle JG, Bugarín-Diz A, Alonso-Moral J, Taboada J. Dealing with hallucination and omission in neural Natural Language Generation: a use case on meteorology. In: Proceedings and Abstracts of the 15th International Conference on Natural Language Generation, July 18–22, 2022. Waterville, ME: Arria, 2022.
7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. arXiv, December 26, 2022 (https://arxiv.org/abs/2212.13138).
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198.
9. Nuance. Automatically document care with the Dragon Ambient eXperience (https://www.nuance.com/healthcare/ambient-clinical-intelligence.html).
10. Kazi N, Kuntz M, Kanewala U, Kahanda I, Bristow C, Arzubi E. Dataset for automated medical transcription. Zenodo, November 18, 2020 (https://zenodo.org/record/4279041#.Y_uCZh_MI2w).
11. Cancarevic I. The US medical licensing examination. In: International medical graduates in the United States. New York: Springer, 2021.

DOI: 10.1056/NEJMsr2214184