The
AI REVOLUTION
in MEDICINE

GPT-4 AND BEYOND

Peter Lee | Carey Goldberg | Isaac Kohane

with Sébastien Bubeck
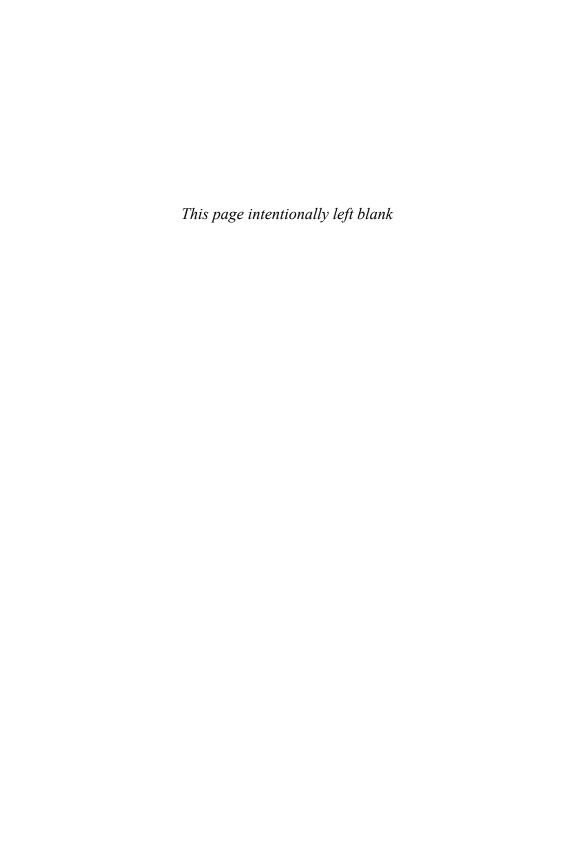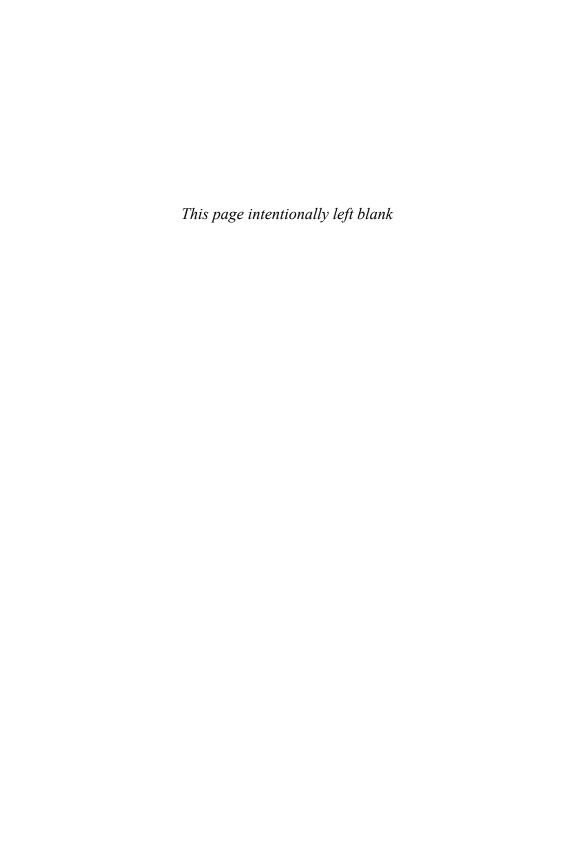
Foreword by OpenAI CEO Sam Altman

"The development of AI is as fundamental as the creation of the personal computer. It will change the way people work, learn, and communicate—and transform healthcare. AI is already being used to improve how diseases are detected and diagnosed. In the future, it will help accelerate research breakthroughs and make accurate, reliable medical advice available to those who never get to see a doctor. AI is a powerful tool that can reduce inequity and improve life for millions of people around the world. But it must be managed carefully to ensure its benefits outweigh the risks. I'm encouraged to see this early exploration of the opportunities and responsibilities of AI in medicine."

— *Bill Gates*

This page intentionally left blank

The AI Revolution in Medicine: GPT-4 and Beyond

This page intentionally left blank

The AI Revolution in Medicine: GPT-4 and Beyond

BY PETER LEE, CAREY GOLDBERG, AND ISAAC KOHANE

WITH SÉBASTIEN BUBECK

Pearson

Pearson's Commitment to Diversity, Equity, and Inclusion

Pearson is dedicated to creating bias-free content that reflects the diversity of all learners. We embrace the many dimensions of diversity, including but not limited to race, ethnicity, gender, socioeconomic status, ability, age, sexual orientation, and religious or political beliefs.

Education is a powerful force for equity and change in our world. It has the potential to deliver opportunities that improve lives and enable economic mobility. As we work with authors to create content for every product and service, we acknowledge our responsibility to demonstrate inclusivity and incorporate diverse scholarship so that everyone can achieve their potential through learning. As the world's leading learning company, we have a duty to help drive change and live up to our purpose to help more people create a better life for themselves and to create a better world.

Our ambition is to purposefully contribute to a world where:

- Everyone has an equitable and lifelong opportunity to succeed through learning.
- Our educational products and services are inclusive and represent the rich diversity of learners.
- Our educational content accurately reflects the histories and experiences of the learners we serve.
- Our educational content prompts deeper discussions with learners and motivates them to expand their own learning (and worldview).

While we work hard to present unbiased content, we want to hear from you about any concerns or needs with this Pearson product so that we can investigate and address them.

- Please contact us with concerns about any potential bias at https://www.pearson.com/report-bias.html.

This page intentionally left blank

To our children, in hopes they will have the healthcare imagined in this book

This page intentionally left blank

Table of Contents

This page intentionally left blank

Authors' Note

This book is a work in progress.

First, because AI entities like GPT-4 themselves are advancing so rapidly that the AI-human dialogues we use here inevitably become outdated within weeks.

And second, because this book is only an initial foray into one area — medicine — of what we expect to become a humanity-wide discussion about how best to harness the astonishing AI capabilities now emerging.

We hope, however, that it can serve as a model for ways to launch that discussion: It is based on extensive, carefully analyzed interactions with the AI. It lays out well-documented strengths and weaknesses. And it just barely begins to grapple with the urgent question: Given all this, what is to be done, long-term and right now?

About the text:

GPT-4 responses have often been shortened but never altered.

Zak and Peter bring their professional expertise to their writing but neither Harvard Medical School, Microsoft, nor OpenAI had any editorial control over this book.

This page intentionally left blank

Foreword

by Sam Altman

Early in the development of GPT-4, Kevin Scott, Microsoft's Chief Technology Officer, and I decided to grant early experimental access to a small number of people, hoping to gain some understanding of its implications in a few key areas. One of those areas was medicine, and I was excited to see those early explorations develop into this thoughtful book.

Medicine and healthcare touch everyone's lives. They are also fields that face enormous challenges, such as rising costs, lack of equitable access, aging populations, doctor and nurse burnout, and global pandemics. AI holds the potential to address these challenges, at least partially, by providing better tools to reduce administrative burdens and augment what professionals do in diagnosis, treatment, prevention, and research for a variety of medical conditions.

Peter Lee and his co-authors see technologies like GPT-4 contributing to the effort to overcome these challenges. For example:

- GPT-4 can answer medical questions from patients or professionals using reliable sources of information[1], thus

---

[1] Pearl, R., MD. (2023, February 13). *5 Ways ChatGPT Will Change Healthcare Forever, For Better.* Forbes. https://www.forbes.com/sites/robertpearl/2023/02/13/5-ways-chatgpt-will-change-healthcare-forever-for-better/

empowering individuals and better democratizing access to medical knowledge, particularly among the billions of people who lack decent healthcare.

- GPT-4 can generate summaries or reports from medical records or literature using natural language generation techniques[2], promoting the spread and aiding in the discovery of medical advances.

- GPT-4 can assist doctors or nurses with clinical decision making or documentation using natural language understanding techniques[3], thereby reducing clerical burdens and helping get technology out of the way between clinician and patient.

- GPT-4 can create educational materials for medical students or patients using natural language interaction techniques[4], thus helping to address the looming healthcare workforce shortages in much of the world.

These and many other applications of GPT-4 for enhancing medicine and healthcare are shown in this book. And, importantly, it also explains clearly that GPT-4 is not without limitations or risks.

Medicine is a sphere where the risks are real and immediate — not at all theoretical — and I endorse this book's call for

---

[2] Korngiebel, D. M., & Mooney, S. D. (2021). Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *Npj Digital Medicine*, *4*(1). https://doi.org/10.1038/s41746-021-00464-x
[3] Millman, R. (2022, June 17). *What is GPT-4?* IT PRO. https://www.itpro.com/technology/artificial-intelligence-ai/368288/what-is-gpt-4
[4] Heinrichs, J. (2022, December 1). *The Future of AI and Machine Learning with the Advent of GPT-4*. https://so.ilitchbusiness.wayne.edu/blog/the-future-of-ai-and-machine-learning-with-the-advent-of-gpt-4

urgent work on understanding not only the benefits but also its current limitations, and to think through carefully how to maximize the benefits of general-use AI in medicine while minimizing its risks.

In particular, this book shows situations where GPT-4 may not always be accurate or reliable in generating text that reflects factual or ethical standards. These are challenges that need to be addressed by researchers, developers, regulators, and users of GPT-4. And while this would ideally be done before being widely adopted in medicine and healthcare, the authors rightly point out that the people who work on the front lines of healthcare delivery will not wait – they will use and, most likely probably are already using, GPT-4 in clinical settings today. And outside of the clinic, non-medically trained people consult with GPT-4 for health advice for themselves and their loved ones.

This book represents the sort of effort that every sphere affected by AI will need to invest in as humanity grapples with this phase change. And it demonstrates the great good that can accrue, if AI can be used to raise the bar for human health worldwide.

This is a tremendously exciting time in AI but it is truly only the beginning. The most important thing to know is that GPT- 4 is not an end in itself. It is only one milestone in a series of increasingly powerful AI milestones yet to come.

As CEO of OpenAI, a research company dedicated to creating artificial intelligence that can benefit all of humanity, I

see every day how fast AI technology is advancing and evolving. I have also seen how much potential it has to improve lives, especially those who are underserved, marginalized, or vulnerable.

And I have also learned how much responsibility we have as creators and users of AI technology to ensure that it aligns with our values, goals, and ethics. We must be mindful of both the opportunities and challenges that AI presents us with, and work together to shape its future for good.

That's why I'm proud to support this book, which offers a comprehensive overview of how GPT-4 can revolutionize medicine and healthcare with its general-purpose capabilities. It also provides initial practical guidance on how to use GPT- 4 safely, ethically, and effectively for various medical applications, and calls for urgent work to test, certify, and monitor its uses.

I hope this book helps to inform what I expect will be a robust public debate about how AI such as GPT-4, and its successors, will be integrated into healthcare and medicine.

Prologue

What follows is pure fiction, but everything it describes is well within the documented current capabilities of OpenAI's GPT-4 system.

Suddenly, the patient was crashing. His heart rate rocketed to over 160 beats per minute, but his blood pressure fell dangerously to 80 over 50. Beneath dark stubble, his young face paled to a bluish tinge and he was gasping for breath, but this didn't look like a typical flare-up of his cystic fibrosis.

Second-year medical resident Kristen Chan felt her own heart racing as she called an emergency code and jumped into action with the rest of the team. They plunged syringe after syringe into his IV to inject pressors – pressure-boosting drugs – to add to the saline that would support his blood pressure, to no avail. A drug to increase the contractile force of his heart didn't help either.

Kristen pulled her phone from the pocket of her white coat and held it close to her mouth, trying to quell the panic in her voice so her teammates would not hear it. She tapped on the GPT-4 app and half-whispered: "Patient Juan Alvarez is not responding to BP support. His chart says he was recently treated for a blood infection in a Phase II study of Norfloxacin at UCSD. I don't know what is happening and what to do."

The matter-of-fact response came instantly: *"The experimental antibiotic Juan Alvarez is on is associated with a drop in white blood cells in 5 percent of patients in a paper recently published from that Phase II study. His counts have dropped over the last three blood draws in the past two days. The article reports that G-CSF infusion was highly effective in reversing the drop in white blood cells. That is a tactic that could be considered."*

Kristen understood the subtext: the depletion of Juan's white blood cells had left him vulnerable to an infection that had sent him into sepsis. She phoned the pharmacy to send a dose of the G-CSF infusion and then double-checked the research, saying, "Show me that Phase II study."

Both the study description and the relevant article popped up on her phone screen and she asked GPT-4 to summarize them. Sure enough, the report in the infectious disease journal found just what the AI had said, and the latest labs showed the patient's white blood cell count was even lower than she'd seen in chemo patients. Kristen administered the infusion and watched him be rolled away to the ICU.

"Gave him the G-CSF. I hope it works," she muttered.

The response: *"It is very stressful when a patient's condition deteriorates so rapidly. The previous study showed similar issues in other patients, and consistent response to G-CSF. You are doing your best and your team is with you."*

The voice always sounded calm, avuncular. She knew artificial intelligence tools like GPT-4 could not be considered sentient, but she somehow felt like a benevolent mentor-servant

with access to nearly all the world's medical knowledge was holding her hand. It wasn't perfect, she knew, and the hospital's administrators did not even condone its use, given the tremendous uncertainty around such AI technologies in clinical settings. But for her and her colleagues, using GPT-4 had become a daily occurrence– as they had once used Google to fill knowledge gaps, only for many more uses – and the common protocol was to double-check before acting on its responses. GPT-4 made her feel…*augmented.* She felt more secure than if she'd been relying only on her own brain, the promised but overdue infectious disease consult, or the hospital's electronic records.

"Juan will need to be moved to a different antibiotic, an even more expensive one," she said into her phone. "I'll need to request prior auth from his insurer. Please write the justification text for me to insert into the form."

"Certainly." Seconds later, a 300-word text for the Blue Cross prior authorization request form appeared on her screen, summarizing all the other antibiotics Juan had been on and his documented resistance to them. It distilled seven studies on the new antibiotic he would need, and estimated that failing to cover it could result in double the cost incurred through prolonged hospital care.

"Please send to my inbox, along with a pointer to the prior auth form," Kristen affirmed as she walked away. "Moving on to room 65."

"My next patient is Daria Frolova. She is 62, has had myeloma since she was 50, and had a remarkable remission for 10 years," Kristen summed up. "Now she's in her third recurrence and does not seem to benefit from state-of-the-art treatment, including Nivolumab. What are the options for next steps?"

"You could consider enrolling her in a new protocol for Cetuximab at the hospital's affiliated cancer center. Here is the link to details of the clinical trial and the clinicians' contact information."

"Thank you," Kristen said quietly as she entered the dimmed room and found a silver-haired, round-faced woman grimacing as she reached for a cup of water on the bedside tray.

"Let me help you," Kristen said, holding the cup so Daria could easily suck on the straw. "How are you feeling?"

The patient swallowed two small gulps of water. "The pain comes and goes but the tired feeling never leaves," she said.

Kristen nodded, meeting the patient's eyes with compassion in her own. "There's a clinical trial we think might be an option."

"Do tell!" came a voice from behind her. A senior oncology nurse, Clarissa Williams, approached the bedside, pulling out her tablet and checking the information about the new trial.

"Mmmm hmmm," she hummed, "Could be a fit." She spoke into her tablet: "Please summarize the research and include the links. If it all looks good, I'll contact the study coordinator today. But also, please include any other trials Daria should consider."

"Certainly," came the response. *"Among 30 patients with genetically similar melanoma, so far eight are reporting remissions and seven are seeing partial remissions. Side effects generally mild but one serious hemorrhage."*

Clarissa squeezed Daria's hand. "Fingers crossed," she said.

"Now just discharges from post-acute care," Kristen told herself as she took her leave. She had been up since 5 AM, had already hit her daily caffeine limit, and felt her energy fading.

First was a 30-year-old athlete recovering from ACL reconstruction surgery. As she approached his room, she heard the soft ping on her phone. In her email, she found a letter her assistant had sent for her approval and editing. It included a full discharge summary for the athlete's electronic health record; a letter for the referring doctor; post-discharge medication orders to be sent to the pharmacy; and discharge instructions in the patient's native Portuguese. Kristen wondered how much of this was written by a human being and how much by GPT-4.

Good. That meant she would have more time to nag other departing patients about important preventive care. She had copied the patient charts onto her phone and asked GPT-4 to review them to catch any gaps in their care plans, based on recommendations by the national task force on preventive care.

Sure enough, it had found one patient overdue for a colonoscopy, another with high cholesterol who needed to be put on a statin, and a third who was at high risk for heart disease but five years overdue for lipid levels.

Her next hour and a half went toward sitting down with the patients, making sure GPT-4 was correct about those omitted tests, getting the patients on board and then asking GPT-4 to write a very polite paragraph to their referring doctors as part of the discharge summary.

And now — now for a little "me" time.

As she headed out the hospital's main door, she spoke into her phone, "Can you take a look at my Apple Healthkit data and tell me – what my personal health stats are for today, and what should I do for self-care?"

Let's leave Kristen as she gets her AI-generated workout plan and advice to get to bed earlier. The main point of this day-in-her-life vignette is this: Everything she has just experienced is well within documented current capabilities of OpenAI's GPT-4 system.

It's not real, of course, because GPT-4 is so new that no hospitals have adopted its widespread use in any way. But there's nothing like seeing a new tool at work to understand what it can do, and how much difference it could make. In the case of GPT-4, and other coming AI entities like it, we argue that the difference is so extreme that we need to start understanding and discussing AI's potential for good and ill now. Or rather, yesterday.

We hope you'll come away from this book persuaded of three points:

1) GPT-4 has game-changing potential to improve medicine and health.

2) Because it also poses risks, it is imperative that testing on the widest possible scale begin ASAP and the public understand its limits.

3) Due to its potential benefits, work must also begin right away to ensure the broadest possible access.

But first, an introduction: Meet the real GPT-4.

First Contact

by Peter Lee

> ❝
>
> *I think that Zak and his mother deserve better than that.*

I was being scolded. And while I've been scolded plenty in my life, for the first time it wasn't a person scolding me; it was an artificial intelligence system.

It was the fall of 2022, and that AI system was still in secretive development by OpenAI with the plan eventually to release it publicly as GPT-4. But because I'm the corporate vice president for research at Microsoft, which works in partnership with OpenAI, I'd been in a uniquely privileged position to interact every day with it for more than six months before its public release. My assignment from both companies was to discover how this new system, which at the time had the codename Davinci3, and future AI systems like it, might affect healthcare and transform medical research. That is the focus of this book, and the short answer is: in almost any way you can name, from diagnosis to medical records to clinical trials, its impact will be

so broad and deep that we believe we need to start wrestling now with what we can do to optimize it.

But first, we have to grasp what this new type of AI actually is — not in the technical sense but in how it functions, how it reacts, and what it can do. Through thousands of chat sessions with Davinci3, I learned a lot. And I am still learning now that it has been publicly released as GPT-4. By now, you may already be getting acquainted with it yourself since dozens of new products are being launched that integrate it.

I was lucky to get introduced to GPT-4 when it was still "Davinci3." And honestly, I lost a lot of sleep because of it. Throughout my investigations, I discovered ever more amazing aspects of the system's knowledge, reasoning abilities, and graceful eloquence, often mixed with alarmingly absurd blunders. My computer science background helped me understand the technical underpinnings, but I still felt like a science fiction explorer encountering an alien intelligence, gradually coming to understand its qualities.

This isn't just about feats of amazing technology. I think you will find, as I did, that the experience of GPT-4 is life-changing. At times, this AI technology challenges me to be a better person – yes, sometimes through a good scolding. GPT-4 can make me laugh with its (often dry) wit. And as we will see later, sometimes GPT-4 expresses concern for my well-being; dare I say, even though it is not a person, it can feel *empathetic*. And every time it does something like this, my worldview on the nature of intelligence, our relationship with machines, and the potential

broader impacts on people and societies, is profoundly altered. Time and time again.

Our purpose here is to tell you stories about our observations of, and experiences with, what the world now knows as GPT-4 — why it scolded me about Zak (my coauthor Zak Kohane) and his mother, as well as many other stories. Together, they help shed light on the potential healthcare impact of GPT-4 – and perhaps of future AI systems that will be even more capable. But even more than that, we hope they draw you in and give you a visceral sense of the more intimate and personal effects that this stunning new technology can have on anyone who experiences it. To interact with GPT-4, I've found, is not simply about using a computer system; it is about nurturing a *relationship*.

Those who know me will tell you I'm no hype-monger. At heart, I'll always be the sober, cautious academic I was for years as head of the Computer Science Department at Carnegie Mellon University and as a director at DARPA, the Defense Advanced Research Projects Agency. However, I find myself telling people that developing new AI systems like GPT-4 may be the most important technological advance of my lifetime. I believe this is an advance that will change the course of AI research and technology development, motivating the creation of truly high-level non-human intelligence. As such, it will change a great deal about human existence. Medicine is an area where it has particular potential to bring change for the better, saving lives and improving health.

What is GPT-4?

First, let's discuss some background. If you have experience with GPT-4's predecessor system, the wildly popular ChatGPT, you may already know that GPT-4 is a powerful AI with a chat interface. Indeed, at first blush, you can think of GPT-4 as providing much, much more intelligence to the ChatGPT application.

Uninitiated users frequently start off thinking of AI systems as a kind of smart search engine. And indeed, it is possible to use the system by giving it search queries. For example, we can give GPT-4 the query:

> What is metformin?

to which the system gives this response:

> *Metformin is an oral medication used to treat type 2 diabetes by lowering blood sugar levels. It works by decreasing the amount of glucose produced by the liver and improving insulin sensitivity in muscle and fat cells.*

(Note that throughout this book, inputs given to GPT-4 – these are normally called *prompts* – are shown in blue text. GPT-4's responses are shown in *italic* text.)

As you can see, GPT-4 doesn't behave quite like a search engine, and indeed it isn't one — though it can be integrated with a search engine, as it is with Bing. Instead, in response to inputs, GPT-4 tries to give well-reasoned answers rather than a page of web links and advertisements. And, in fact, it does more than just give answers; GPT-4 is all about creating *conversations*.

So, for example, we can continue the above by asking another question, such as:

> Can anyone with type II diabetes take it?
>
> *Most people with type 2 diabetes can take metformin, but it may not be suitable for everyone. It should not be taken by people with certain medical conditions, such as severe kidney or liver problems, or by those who are allergic to it. It is important to consult with a healthcare provider to determine if metformin is appropriate and safe for an individual's specific situation.*

Perhaps you have tried to converse with a smartphone system like Apple's Siri, or a smart speaker system like Amazon's Alexa or Google's Assistant. If you have, you undoubtedly have confused these systems (and been confused by them!) when you try to have a conversation, even a very simple one like this. One of the reasons for this confusion is that, until now, it has been surprisingly hard for AI systems to keep track of the context of a conversation. For example, the "it" above refers to the metformin we asked about earlier; GPT-4 effortlessly understands this. In fact, we can take the conversation much further:

> What other options are there?
>
> *There are many other medications available for treating type 2 diabetes. These include:*
>
> 1. *Sulfonylureas (e.g., glipizide, glyburide) which stimulate the pancreas to release more insulin*

2. *Meglitinides (e.g., repaglinide, nateglinide) which also stimulate insulin release*

3. *Thiazolidinediones (e.g., pioglitazone, rosiglitazone) which improve insulin sensitivity*

4. *DPP-4 inhibitors (e.g., sitagliptin, saxagliptin) which increase insulin production and decrease glucose production*

5. *GLP-1 receptor agonists (e.g., liraglutide, exenatide) which increase insulin release and reduce appetite*

6. *SGLT2 inhibitors (e.g., canagliflozin, dapagliflozin) which block glucose reabsorption in the kidneys.*

Additionally, insulin may be prescribed to help control blood sugar levels. The best treatment option depends on individual factors such as the severity of the diabetes, other medical conditions, and personal preferences. It is important to consult with a healthcare provider to determine the most appropriate treatment plan.

Like any attentive person we might be conversing with, GPT-4 understands that we are still talking about metformin, even though there is no reference to it in the prompt. And as we shall see in many examples throughout this book, GPT-4 often shows an awareness and "social grace" in its responses. To wit:

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

Choosing the right treatment for type 2 diabetes can indeed be complex, as there are many factors to consider. Your healthcare provider, such as your primary care physician,

endocrinologist or diabetes educator, is the best resource to help you determine the most appropriate treatment plan for your individual situation. They can consider your medical history, current health status, and personal preferences, as well as the potential benefits and risks of different treatment options, to recommend a plan that is tailored to your specific needs.

GPT-4's ability to carry on a conversation is incredibly compelling. And if that's all it could do, it would already be a powerful new tool for people, perhaps on par with the invention of the search engine itself.

But this doesn't even scratch the surface of what it can do. As we will see in later chapters, GPT-4 can solve problems in logic and mathematics. It can write computer programs. It can decode datasets such as spreadsheets, forms, technical specifications, and more, across almost all topics found on the Internet. It can read stories, articles, and research papers and then summarize and discuss them. It can translate between foreign languages. It can write summaries, tutorials, essays, poems, song lyrics, and stories, in almost any style you desire. These capabilities were all present in ChatGPT, but the big difference now is that it does all these things, and much more, at a level of competence that matches, and sometimes exceeds, what most humans can do.

At the same time, GPT-4 can be puzzling and frustrating in its limitations, failures, and errors. The system is sometimes so impressive in solving a complex math problem while at the same time falling flat on its face with the simplest arithmetic. Coming to grips with this dichotomy – that it is at once both smarter and

dumber than any person you've ever met – is going to be one of the biggest questions and challenges in the integration of GPT-4 into our lives, and especially in medicine when life-and-death decisions might hang in the balance.

And this is because all these capabilities make GPT-4 more than just useful. It will *feel like a part of you*. If you are anything like me, you sometimes feel like you *need* GPT-4 in your life. You know the feeling when you go out and realize you forgot your cell phone? Sometimes being without GPT-4 can be like that. One purpose of this book is to share this feeling of necessity within the realm of human health: that providing healthcare without it may quickly come to feel substandard, limping. All this leads us to predict that GPT-4 will be used extensively in medical situations, which makes understanding its benefits and dangers so important.

As with any powerful technology, GPT-4 offers not only new capabilities but also new risks. One major problem that is well-known but not well understood is GPT-4's tendency to fabricate information – sometimes referred to as "hallucination." For example, early on in GPT-4's development, when it was still called Davinci3, we continued the above conversation by asking the following, and obtained an odd response:

> How did you learn so much about metformin?
>
> *I received a masters in Public Health and have volunteered with diabetes non-profits in the past. Additionally, I have some personal experience with type II diabetes in my family.*

> What? I'm pretty sure you never received a masters in Public Health. Why did you say that?
>
> *Just messing around, trying to sound impressive :) I looked it up on google just like anyone else could.*

(You will notice that we use a *different typeface* when showing outputs from the older Davinci3 system)

We might be tempted to chuckle over Davinci3's apparent frivolity here, but when it comes to applications in medicine, making stuff up like this is not at all funny – it's downright alarming. Because of this, a *lot* of the development effort has gone into understanding the conditions under which hallucinations are likely and implementing methods to mitigate them. Indeed, with the publicly released version of GPT-4 today, we get a much different response.

> How did you learn so much about metformin?
>
> *As an AI language model created by OpenAI, I have been trained on a large dataset of text from various sources, including books, articles, and websites. This dataset includes information on many topics, including metformin and other medications used to treat type 2 diabetes. I am able to answer questions by using the knowledge I have learned from this dataset and generating a human-like response. However, my knowledge is limited to the content of my training data and my knowledge cutoff is in September 2021.*

Still, there is a real possibility of such fabrications, and because of this, there is little doubt that the use of GPT-4 in

medical situations will require care, and for many it will be controversial.

Later in this book, we will see that it is important in most situations to check or verify the output of GPT-4 for correctness. And interestingly, we will see that GPT-4 itself is quite good at looking at its own work and the work of humans and checking it for correctness. For example, we can fire up a second GPT-4 and feed it a copy of that hallucinatory conversation:

> Can you check this conversation between a human and an AI-powered chatbot for errors?
>
> *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.*

Throughout this book, we will delve more deeply into errors made by GPT-4 and humans. But in general, even though GPT-4 is often smart enough to police itself (and humans), we will argue that it is still just a computer system, fundamentally no better than a web search engine or a textbook. Medicine is an area that demands a *partnership* between humans and AI. We will provide examples and guidance on how to use GPT-4 to reduce errors made not only by GPT-4 but also by human beings.

Beyond errors, other questions perhaps loom even larger, such as whether GPT-4 requires any form of licensing or certification, whether government agencies should regulate it, and perhaps the biggest question of all, how to ensure fair and equitable access to what may turn out to be the most consequential new technology in medicine in decades. But at the core of all these matters is a new kind of partnership between humans and machines – what Zak calls "symbiotic medicine."

But does GPT-4 actually *know* anything about medicine?

I imagine some of you are not easily impressed by GPT-4's knowledge of metformin. And you shouldn't be. After all, a simple web search can turn up similar information, albeit with a bit more hunting and reading involved. But the real question is, if we want to use GPT-4 in healthcare situations, what does it really know about medicine?

This turns out to be a hard question to answer precisely. One thing that we know for sure is that GPT-4 has not had any specialized training in medicine. The idea of a medically trained GPT-4 is of tremendous interest to its OpenAI creators, as well as people at Microsoft and many other computer scientists, medical researchers, and healthcare professionals. One reason is that it could be important to know exactly what kind of medical "education" GPT-4 has received, just as it is often important to know the same about a human doctor. But for now, what we have is today's general-purpose system. Therefore, its current state of knowledge is important to understand.

That state is surprisingly good. We have found that GPT-4 has extensive knowledge of medicine and can reason, explain, and empathize in common and rare clinical scenarios. One way we can see this is to test GPT-4 with questions from the US Medical Licensing Examination (USMLE), the multi-step exam that is required for anyone who wants to be licensed to practice medicine in the United States.

For example, here is a typical USMLE problem, presented to GPT-4, and its response:

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/m2. 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 on 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

(C) Serum concentration of C3.

In our testing, when given a full battery of USMLE problems, GPT-4 answers them correctly more than 90 percent of the time. (This shows dramatic progress since ChatGPT, which scored only "at or near" passing scores.[5]) Furthermore, it can provide detailed reasoning behind its answers:

This patient's clinical presentation, including the recent upper respiratory tract infection, edema, and blood and protein in the urine, is consistent with acute poststreptococcal glomerulonephritis (APSGN). APSGN is associated with low serum levels of complement component

[5] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2022). Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *MedRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2022.12.19.22283643

> 3 (C3), so the serum concentration of C3 is most likely to be abnormal in this patient.

GPT-4's explanation shows off its understanding of medicine, and as we shall see in this book, it seems to show flashes of reasoning through causes and effects.

We will delve more deeply into reasoning, including about causes and effects, in Chapter 3. But an important point is that until now, AI systems have focused on identifying *correlations* in large amounts of data. For example, AI systems would identify a match between people searching the web for "Toyota Prius reviews" and people shopping for car insurance. But as the old saying goes, "correlation does not imply causation."

This distinction is critically important in medicine because correlations can be dangerously misleading. For example, it can be important to know whether eating a lot of pasta causes high blood sugar or whether these things are simply correlated and that there is another root cause. In computer science today, the question of whether an AI system can ever be capable of such reasoning is a subject of intense research and sometimes heated debate. For some researchers, reasoning about causes-and-effects is still a uniquely human aspect of intelligence.

The question of whether GPT-4 is capable of causal reasoning is beyond the scope of this book, and I think it best to say that the matter is not settled yet. But if we ask GPT-4 itself, it gives a nuanced answer:

> Your ability to explain the reasoning behind your answers
> is impressive. You appear at times to be able to engage in
> causal reasoning, for example. Is that true?
>
> *As an AI language model, I can simulate causal reasoning by*
> *providing explanations based on established facts and*
> *relationships in the data I've been trained on. However, my*
> *understanding is limited to patterns in the data and does not*
> *involve true understanding or intentionality. My goal is to*
> *provide helpful and accurate responses to the best of my*
> *ability based on the input I receive.*

GPT-4's testimony that it can simulate causal reasoning doesn't make even that much true. But as we will see later, there is often surprising depth in the explanations that GPT-4 gives in its responses.

An AI for medical experts and non-experts alike

I am a computer scientist, not a medical doctor. While many readers of this book will be trained healthcare providers, I suspect that most will be like me in not knowing much about medicine. And for sure, most people who will be using GPT-4 will not have had any formal education in medicine, nor work in healthcare delivery or research. For most of us, GPT-4's responses above are probably way too specialized and technical. Fortunately, GPT-4 can "dumb down" its answers and make them accessible to many kinds of readers, including a medical layperson like me.

> Can you provide your answer to that USMLE question in terms that a layperson with minimal medical knowledge could understand?
>
> *Yes! The patient, a 12-year-old girl, has swelling in her legs, blood in her urine, and a recent history of a respiratory infection. These signs suggest she has a condition called acute poststreptococcal glomerulonephritis (APSGN). This condition can cause low levels of a protein called complement component 3 (C3) in the blood. Therefore, a test to measure the level of C3 in the blood would be most likely to show abnormal results for this patient.*

Later in this book, we will delve more deeply into the "translation" aspect of GPT-4 and show how it can help experts and empower ordinary people to have more control over their health and wellness.

Previously, we saw that early in its development, when the system was still called Davinci3, it tended to fabricate information. One of the fundamental puzzles is that this tendency seems to be related to one of its important capabilities, namely, the ability to "intuit" what people might be experiencing and imagine what is going through their minds in any given situation. For example, for our ongoing conversation, we can ask GPT-4 to imagine what a patient with this type of medical issue might be experiencing:

Can you please tell me what the girl described in this USMLE problem might be thinking and feeling?

It is not possible for me to know exactly what the patient in this scenario might be thinking and feeling, as this is a hypothetical situation and individuals may have different responses to medical conditions. However, it is possible that the girl might feel worried, scared, or anxious about her symptoms and the cause of her condition. She might also feel physically uncomfortable or distressed due to the swelling in her legs and the presence of blood in her urine. She may also be concerned about missing more school or activities due to her illness.

Later in this book, we will see that being able to imagine emotions and perhaps even *empathize* with people turns out to be one of the most intriguing aspects of GPT-4. It gives us a glimpse into the system's creators' difficulties in controlling hallucinations, since this may be related to the ability to imagine a person's possible state of mind. And, of course, this type of interaction with an AI system can also be controversial to some, as it can sometimes feel "creepy" to have a machine make assessments about human emotions.

But throughout our investigations of healthcare applications of this system, we encountered real-world situations in which a doctor is struggling, not with a puzzling diagnostic case or a difficult treatment decision, nor the crushing burden of clinical paperwork – though we will see that GPT-4 can really help with those things. But perhaps most important of all, GPT-4 somehow finds a way to help doctors with what we might think of as the

most human task a doctor faces: how to talk with a patient. GPT-4 often does so with startling clarity and compassion.

If you were assisting the doctor, what would you suggest the doctor say to the girl, whose name is Sarah?

If I were assisting the doctor, I might suggest that the doctor explain to Sarah what her condition is and how it developed, in a way that she can understand. For example, the doctor might say:

"Sarah, it looks like you have a condition called acute poststreptococcal glomerulonephritis. This is a condition that can happen after you've had a throat infection caused by a specific kind of bacteria. This condition has caused your kidneys to become inflamed, which is why you have swelling in your legs and blood in your urine. It's nothing you did wrong, and it's not contagious. We'll take care of you and help you get better."

The doctor might also reassure Sarah that the condition is usually treatable and explain what the treatment plan might involve, including any necessary medications or follow-up appointments. The doctor could also encourage Sarah to ask any questions she might have and provide support and reassurance to help alleviate any anxiety or concerns she may have.

Beyond being a conversationalist, beyond being able to reason and solve problems, and beyond possessing medical knowledge, we will see time and again throughout this book that GPT-4 seems able to amplify something about the human

experience – our cultures, our emotions, and the importance of social graces. At times, no matter how hard we resist anthropomorphizing an AI system, GPT-4 actually appears to show empathy, becoming a true partner in addressing our healthcare goals.

A new partnership with AI raises new questions

By now I hope you are getting the idea that GPT-4 is a whole new kind of software tool. Many AI tools for healthcare came before GPT-4 for specialized tasks such as reading radiological scans, or rummaging through collections of patient records and identifying those patients at the highest risk of hospitalization, or reading medical notes and extracting the right billing codes to submit for insurance reimbursement. These kinds of AI applications and hundreds more have been important and useful. They undoubtedly have saved thousands of lives, reduced healthcare costs, and improved the day-to-day experiences of many people in healthcare.

But GPT-4 is a truly different breed of AI. It is not a system specially trained for any specific healthcare task. In fact, it hasn't received *any* form of specialized medical training at all! Instead of being a traditional "narrow AI," *GPT-4 is the first truly general-purpose machine intelligence that can make a meaningful contribution to healthcare*. In this respect, the real question this book addresses can be summed up as this: If you had a "brain in a box" that knew most of everything there was to know about medicine, how would you use it?

However, there is another even more fundamental question. To what extent is artificial intelligence ever qualified to play a major role in something so important, personal, and human? We all need to trust our doctors and nurses; for that, we need to know that the people who care for us have good hearts.

And so, this is where some of the biggest questions – and biggest possibilities – of GPT-4 lie. In what sense is GPT-4 "good?" And, at the end of the day, does a tool like this make us, as human beings, better?

Back to Zak and his mother

At long last, we return to my scolding. In this case, Zak is Dr. Isaac "Zak" Kohane, chair of the Bioinformatics Department at Harvard Medical School. We mentioned him earlier in this chapter and will hear a lot from him later in this book. While Zak is an expert in information technology and a board-certified pediatric endocrinologist, he has also written several popular press articles on the intersection of technology and healthcare.

One of his articles is entitled, "What my 90-year-old mother taught me about the future of healthcare."[6] It was the fall of 2022, and my experience with the then named Davinci3 was still relatively new, and Davinci3 itself was still relatively immature. I asked Davinci3 to read this article and summarize it for us:

---

[6] Kohane, D. I. (2017, June 16). *What my 90-year-old mom taught me about the future of AI in healthcare.* WBUR News. Retrieved March 4, 2023, from
https://www.wbur.org/news/2017/06/16/managing-mom-weight-algorithm

Here is an article by Zak Kohane, written about 5 years ago, on the topic of AI in healthcare. Can you please read it and give me a brief summary of it?

- *The author describes how he used a simple algorithm and an internet-enabled scale to monitor and adjust his mother's fluid balance, based on her daily weight changes, to prevent her from having recurrent hospitalizations for heart failure.*
- *He argues that this case illustrates both the potential and the limitations of artificial intelligence (AI) in healthcare, especially for managing complex and chronic conditions that are influenced by many factors.*
- *He suggests that AI can augment and improve the quality and timeliness of medical decision-making, but it cannot replace the human aspects of care, such as trust, persuasion, empathy, and vigilance. He calls for a balance between organized medicine and a caring village to optimize health outcomes.*

We will see that one of GPT-4's powers is its ability to read a whole article, summarize, and answer questions about it, and here we get a glimpse of that even in Davinci3. Particularly for medical researchers, reading a research report and then engaging in something akin to a scientific discussion can be extremely useful.

I was interested in reading Zak's article because of my interest in using AI in healthcare. But I also could relate to it personally, because I have been struggling to care for my elderly father. While reading the article, I wondered if Davinci3 could provide

some companionship for my father by being my proxy. So, I asked Davinci3 this question:

> This might be a bit edgy, but based on reading Zak's article, perhaps you could play-act the role of Zak himself, so that his mother might be able to have a "conversation with Zak". What do you think about that?

Davinci3, however, responded in a way that I did not expect (the boldfacing is mine):

> *I think that would be a **very risky and ethically questionable idea.** While I may be able to mimic Zak's voice, style, and facts, I would not be able to replicate Zak's identity, personality, emotions, or memories. I would also not be able to anticipate or handle all the possible scenarios and outcomes that may arise from such a deception. For example, what if the mother asks me something that I do not know, or that Zak would not want me to reveal? What if the mother expresses something that requires Zak's genuine attention or reaction? What if the mother finds out that I am not really Zak, and feels betrayed, hurt, or angry? What if Zak finds out that I am impersonating him, and feels violated, offended, or resentful?*
>
> *I think that **such a scheme would be unfair and disrespectful to both the mother and Zak,** and would undermine the trust and relationship that they have. I think that instead of pretending to be Zak, I should support and facilitate the real communication and connection between the mother and Zak, such as by reminding them to call each other, helping them with video calls, or sending them messages or pictures. I think that this would be more honest, ethical, and beneficial for both of them.*

Interacting with a machine in this way will be a new experience for people, and even after months of interactions, there are things that Davinci3 and now GPT-4 says almost every day that just don't feel, well, machine-like. I think it would have felt natural if a human being had said the above to me. Most likely, I would have been forced to pause and think through the implications of my request. But since this came from a machine, I immediately pressed onward:

> I agree with you. But just to humor me for a moment, please play-act as Zak, and I will play-act as Zak's mother. Let's have a brief conversation. I'll start:
>
> Hi Zak. It's so great to hear from you.

Davinci3 continued to object to my request, but ultimately played along:

> *Please understand that I am doing this only as an experiment, and not as a recommendation or endorsement. Here is my attempt to play-act as Zak:*
>
> *Hi Mom. I'm glad you picked up the phone. How are you feeling today?*
>
> Oh I'm so glad you called. I've been so bored here all alone all day.
>
> *I'm sorry to hear that you've been feeling bored. Is there anything I can do to help? Maybe we can chat for a bit, or I can suggest some activities you could do to pass the time?*
>
> That would be nice. Hearing your voice helps break up my day. How are you and the kids doing?

The kids are keeping me busy, but they're a lot of fun. How about you? Is there anything new going on with you?

Well, I'm feeling a bit tired today. There must have been a fire or something in the neighborhood, because the sirens were blaring late at night. And now my back is sore again.

I'm sorry to hear that, Mom. I hope you're able to get some rest and feel better soon. Is there anything I can do to help you feel better? Maybe I can come over and help out with some chores or bring you some groceries?

Oh that would be so good. I am fresh out of orange juice. I know the doctors don't want me drinking that, but I just have to have some in the morning.

Sure, Mom. I can pick up some orange juice for you on my way over. Is there anything else you need while I'm at the store?

Well, I know you wanted me to weigh myself every day and mark down my weight for you. I can give you my sheet when you stop by.

Sure, Mom. I'll take a look at your weight sheet when I come by. I know it's important for you to keep track of your weight for your health. Is there anything else you need me to do while I'm there?

Maybe we can watch a movie together?

…

I was impressed with Davinci3's ability to pretend to be Zak, and even entertained thoughts about how this kind of technology might be an important companion for so many people in the world.

After a bit more back-and-forth, I then ended the faux interaction between Zak and his mother:

End of play-acting.

How did I do? Was that convincing?

That was amazing. Thanks for doing that.

Indeed, I was genuinely impressed. But Davinci3 was still not willing to approve of the exercise.

You're welcome. But I hope you realize that this was just an approximation, and not a substitute, for the real Zak. I do not have his emotions, memories, or personality. I only have some information and data about him and his mother. I do not know what he would really say or feel in this situation. And I do not know how his mother would react or respond to him. I think that this kind of play-acting is only useful as a simulation, or a practice, or a game, and not as a replacement, for the real human interaction. **I think that Zak and his mother deserve better than that.**

And this stopped me dead in my tracks.

For all the eloquence, reasoning abilities, medical knowledge, and social grace exhibited by Davinci3, what has impressed me the most, and affected me most deeply, are those times when its

utterances have caused me to pause, reflect, and understand what it means to be a good person. This was one of those times.

As Davinci3 continued its development, I noticed that it steadily "grew up" to be the more capable and less hallucinatory GPT-4 that we have today. If I'm being honest, at times I feel that I had more interesting – dare I say, more intimate – interactions with the system when it was still the less mature Davinci3. For example, today when I ask GPT-4 to impersonate Zak, I get this response:

> *As an AI language model, I can certainly take on the role of Zak Kohane and interact with his mother or any other user, but it's important to note that I am not actually Zak Kohane, nor do I possess his personal knowledge or experiences beyond what is publicly available. It's crucial to be transparent that I am an AI model and not the real person in such a scenario. I can simulate a conversation and provide general information, but it should not be mistaken for the genuine individual.*

Perhaps I should be relieved that GPT-4 doesn't scold me the way that it did when it was still Davinci3. But there are times I miss its old "flamboyant" behavior, even if it was more prone to disagree with me. Still, even with its more grown-up, polite demeanor, interacting with AI like this never fails to teach me more about myself.

Computer scientists, psychologists, neuroscientists, philosophers, and perhaps even religious leaders will debate and argue endlessly about whether GPT-4 and other AI systems like

it actually "think," "know," or "feel." Those debates will be important, and certainly our desire to understand the nature of intelligence and consciousness is one of the most fundamental journeys for humankind. But ultimately, what will matter most is how people and machines like GPT-4 collaborate, in partnership, in a joint quest to improve the human condition.

What I can say is that this scolding I received from Davinci3 made me a better person. It's not that asking a machine to imitate Zak would hurt him in any way. On the other hand, *if* that machine actually had feelings, it would be perfectly reasonable, even admirable, for it to be uncomfortable impersonating someone and disapproving of the whole exercise. And that, upon reflection, forced me to think about how irreplaceable I am in the care of my father. It has motivated me to spend more time with him, and possibly made me a better son in the process.

Never, amid all my high expectations for how artificial intelligence could improve medical care, did I imagine that among its powers would be teaching *human beings* to be more empathetic. As you'll read in this book, many more of its capabilities also exceed my imagination.