



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

Sample pages

Table of Contents

Authors' Note.....	xiii
Foreword	xv
Prologue.....	1
Chapter 1: First Contact	8
Chapter 2: Medicina ex Machina	35
Chapter 3: The Big Question: Does It “Understand?”	67
Chapter 4: Trust but Verify.....	99
Chapter 5: The AI-Augmented Patient	120
Chapter 6: So Much More: Math, Coding, and Logic	144
Chapter 7: The Ultimate Paperwork Shredder.....	172
Chapter 8: Smarter Science	205
Chapter 9: Safety First	240
Chapter 10: The Big Black Bag.....	261
Epilogue.....	273
Further Reading.....	280
Acknowledgments	281

Sample pages

CHAPTER 1

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

Sample pages

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

Sample pages

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.

Sample pages

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*.

Sample pages

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*

Sample pages

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,

Sample pages

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

Sample pages

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.

Sample pages

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

Sample pages

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.

Sample pages

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.

Sample pages

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:

Sample pages

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., Madiaga, 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:

Sample pages