A Mystery in the E.R.? Ask Dr. Chatbot for a Diagnosis.

At a medical school in Boston, instructors are using ChatGPT in training exercises to help teach students how to think like doctors.

A family physician in Pennsylvania who uses artificial intelligence to produce summaries of patient visits, Soon, A.I. could be used to diagnose illnesses. Maddie McGarvey for The New York Times

By Gina Kolata

Gina Kolata joined a meeting at Beth Israel Deaconess Medical Center in Boston to report this story.

July 22, 2023

The patient was a 39-year-old woman who had come to the emergency department at Beth Israel Deaconess Medical Center in Boston. Her left knee had been hurting for several days. The day before, she had a fever of 102 degrees. It was gone now, but she still had chills. And her knee was red and swollen.

What was the diagnosis?

On a recent steamy Friday, Dr. Megan Landon, a medical resident, posed this real case to a room full of medical students and residents. They were gathered to learn a skill that can be devilishly tricky to teach — how to think like a doctor.

“Doctors are terrible at teaching other doctors how we think,” said Dr. Adam Rodman, an internist, a medical historian and an organizer of the event at Beth Israel Deaconess.

But this time, they could call on an expert for help in reaching a diagnosis — GPT-4, the latest version of a chatbot released by the company OpenAI.

Artificial intelligence is transforming many aspects of the practice of medicine, and some medical professionals are using these tools to help them with diagnosis. Doctors at Beth Israel Deaconess, a teaching hospital affiliated with Harvard Medical School, decided to explore how chatbots could be used — and misused — in training future doctors.

Instructors like Dr. Rodman hope that medical students can turn to GPT-4 and other chatbots for something similar to what doctors call a curbside consult — when they pull a colleague aside and ask for an opinion about a difficult case. The idea is to use a chatbot in the same way that doctors turn to each other for suggestions and insights.

For more than a century, doctors have been portrayed like detectives who gather clues and use them to find the culprit. But experienced doctors actually use a different method — pattern recognition — to figure out what is wrong. In medicine, it’s called an illness script: signs, symptoms and test results that doctors put together to tell a coherent story based on similar cases they know about or have seen themselves.

If the illness script doesn’t help, Dr. Rodman said, doctors turn to other strategies, like assigning probabilities to various diagnoses that might fit.

Researchers have tried for more than half a century to design computer programs to make medical diagnoses, but nothing has really succeeded.

Physicians say that GPT-4 is different. “It will create something that is remarkably similar to an illness script,” Dr. Rodman said. In that way, he added, “it is fundamentally different than a search engine.”

Dr. Rodman and other doctors at Beth Israel Deaconess have asked GPT-4 for possible diagnoses in difficult cases. In a study released last month in the medical journal JAMA, they found that it did better than most doctors on weekly diagnostic challenges published in The New England Journal of Medicine.

But, they learned, there is an art to using the program, and there are pitfalls.

Dr. Christopher Smith, the director of the internal medicine residency program at the medical center, said that medical students and residents “are definitely using it.” But, he added, “whether they are learning anything is an open question.”

The concern is that they might rely on A.I. to make diagnoses in the same way they would rely on a calculator on their phones to do a math problem. That, Dr. Smith said, is dangerous.

Learning, he said, involves trying to figure things out: “That’s how we retain stuff. Part of learning is the struggle. If you outsource learning to GPT, that struggle is gone.”

At the meeting, students and residents broke up into groups and tried to figure out what was wrong with the patient with the swollen knee. They then turned to GPT-4.

The groups tried different approaches.

One used GPT-4 to do an internet search, similar to the way one would use Google. The chatbot spat out a list of possible diagnoses, including trauma. But when the group members asked it to explain its reasoning, the bot was disappointing, explaining its choice by stating, “Trauma is a common cause of knee injury.”

Another group thought of possible hypotheses and asked GPT-4 to check on them. The chatbot’s list lined up with that of the group: infections, including Lyme disease; arthritis, including gout, a type of arthritis that involves crystals in joints; and trauma.

GPT-4 added rheumatoid arthritis to the top possibilities, though it was not high on the group’s list. Gout, instructors later told the group, was improbable for this patient because she was young and female. And rheumatoid arthritis could probably be ruled out because only one joint was inflamed, and for only a couple of days.

As a curbside consult, GPT-4 seemed to pass the test or, at least, to agree with the students and residents. But in this exercise, it offered no insights, and no illness script.

One reason might be that the students and residents used the bot more like a search engine than a curbside consult.

To use the bot correctly, the instructors said, they would need to start by telling GPT-4 something like, “You are a doctor seeing a 39-year-old woman with knee pain.” Then, they would need to list her symptoms before asking for a diagnosis and following up with questions about the bot’s reasoning, the way they would with a medical colleague.

That, the instructors said, is a way to exploit the power of GPT-4. But it is also crucial to recognize that chatbots can make mistakes and “hallucinate” — provide answers with no basis in fact. Using them requires knowing when it is incorrect.

“It’s not wrong to use these tools,” said Dr. Byron Crowe, an internal medicine physician at the hospital. “You just have to use them in the right way.”

He gave the group an analogy.

“Pilots use GPS,” Dr. Crowe said. But, he added, airlines “have a very high standard for reliability.” In medicine, he said, using chatbots “is very tempting,” but the same high standards should apply.

“It’s a great thought partner, but it doesn’t replace deep mental expertise,” he said.

As the session ended, the instructors revealed the true reason for the patient’s swollen knee.

It turned out to be a possibility that every group had considered, and that GPT-4 had proposed.

She had Lyme disease.

Olivia Allison contributed reporting.

Gina Kolata writes about science and medicine. She has twice been a Pulitzer Prize finalist and is the author of six books, including “Mercies in Disguise: A Story of Hope, a Family’s Genetic Destiny, and The Science That Saved Them.” More about Gina Kolata