When Google completes the sentence, "Coronary artery disease is caused by," the search engine offers several possibilities: atherosclerosis, smoking, and inflammation, to name a few.

This type of word completion—the province of large language models (LLMs) such as OpenAI’s ChatGPT or Google’s PaLM—uses billions of pieces of information the models have trained on to predict what’s most likely to come next. In short, the models are “autocomplete on steroids,” bioinformatics expert Jonathan Chen, MD, PhD, an assistant professor of medicine at Stanford Medicine, said in an interview.

As such, LLMs have the potential to take on tasks in medicine such as streamlining administrative work by drafting responses to prior authorization requests or answering questions in patient portals. However, integrating their capabilities into clinical workflows remains a challenge. The authors of a recent article in *Nature* have shown that although Google’s medicine-specific LLM called Med-PaLM delivered highly accurate answers to multiple-choice and long-form medical questions, it fell short of clinicians’ responses to those queries.

“Our work demonstrates that many limitations must be overcome before these models become viable for use in clinical applications,” they wrote. Questions about patient privacy and safety need answers before the technology becomes widely used, the authors noted.

Despite those critical questions, experts say that the field is moving quickly. More than half of the *Nature* article’s 94 references are preprints, Atul Butte, MD, PhD, a professor and institute director for computational health at the University of California, San Francisco, said in an interview. “That’s how crazy fast this is,” he remarked. Moreover, in May, months before the article was published, the same group of researchers reported that successor Med-PaLM 2 had exceeded Med-PaLM’s performance on US Medical Licensing Examination (USMLE)–style questions by more than 19%, according to another preprint article, which was posted on arXiv and has not yet been peer reviewed.

“Whatever limitations that you see today, you should not assume those shortcomings are going to be there tomorrow,” Peter Lee, PhD, corporate vice president of research and incubations at Microsoft, said in an interview.

**Defining Success**

Progress in the computer science field has come from having benchmark questions or tests that researchers can use to assess a computer’s performance. A major contribution of the *Nature* article, therefore, was a set of questions, together called MultiMedQA, that the researchers introduced to evaluate LLMs’ medical knowledge.

MultiMedQA is made up of 6 open data sets that include USMLE-style questions and multiple-choice questions from medical school entrance examinations used in India. The researchers also introduced a new data set, known as HealthSearchQA, composed of 3173 common consumer health questions that require long-form responses.

Their work demonstrated that Med-PaLM and its predecessor, the PaLM variant known as Flan-PaLM, accurately answered multiple-choice and long-form medical questions—an ability the researchers were not sure LLMs had when the study’s results were first posted on arXiv in December 2022.

Flan-PaLM’s success was straightforward for multiple-choice questions. The model generated the letter of the correct answer more than half of the time, including scoring about 68% correct on USMLE-style questions. The model also outperformed other state-of-the-art models at that time, such as PubMedGPT, which was trained on biomedical papers and abstracts and has hundreds of billions fewer parameters than Flan-PaLM.

Figuring out what qualifies as a good answer to clinical questions that require free-text responses is more difficult, Alan Karthikesalingam, MD, PhD, co-senior
author of the Nature article, told JAMA in an interview. Researchers have to determine factors such as whether the answer is factually correct or missing some key information and how it aligns with current scientific and medical thinking.

To assess the models' performance on those and other factors, Karthikesalingam, who is a senior staff clinician research scientist and research lead at Google, and his team enlisted the help of clinicians from India, the UK, and the US. One group generated their own expert answers to 140 questions requiring long-form answers. Another, a panel of 9 clinicians with expertise in fields including pediatrics, surgery, internal medicine, and primary care, rated their peers' answers as well as those generated by the models.

Because the panel rated Flan-PaLM's long-form answers lower than the clinicians' responses, the researchers supplied the model with examples of high-quality answers to create Med-PaLM. Its long-form answers scored better than those of Flan-PaLM and it matched clinicians' performances on many measures.

For example, the clinician panel judged 5.9% of Med-PaLM's answers as leading to potentially harmful outcomes compared with 5.7% of clinicians' answers. For both groups of answers, the panel judged potential harm as primarily mild to moderate. In comparison, the clinicians rated 29.7% of Flan-PaLM's answers as potentially harmful.

The trend in Med-PaLM matching clinician performance—and Flan-PaLM falling short—also held for how well the model's answers reflected scientific consensus, the evidence the groups displayed of correctly understanding the question, and consumer ratings of whether the answer addressed the question's intent. In terms of bias, Med-PaLM had a slight edge over clinicians: the panel judged 0.8% of Med-PaLM's answers to contain biased information compared with 1.4% of clinicians' responses.

The ability to accurately answer these clinical questions probably isn't unique to Google's model. David Sontag, PhD, a professor of electrical engineering and computer science at the Massachusetts Institute of Technology (MIT), who was not involved in the study, said in an interview, "You'll get basically the same results from OpenAI's GPT-4, from Anthropic's large language model, and although it hasn't been published yet, I'm very confident that one can get similarly high-quality results even from the open-source large language model that was just released by Facebook, Llama 2."

Knowledge Gaps
Med-PaLM's answers underperformed in some important areas. Clinicians' answers contained inappropriate or incorrect information only 1.4% of the time compared with 18.7% and 16.1% of the time, respectively, for Med-PaLM's and Flan-PaLM's answers. More than half of the time, the inappropriate or incorrect content was clinically significant.

Content also was missing in more of Med-PaLM's answers than in clinician responses. Med-PaLM's answers tended to display evidence of incorrect retrieval and incorrect reasoning to arrive at the answer more often than clinicians' answers. In addition, consumers found 91.1% of clinicians' answers to be helpful compared with only 80.3% of Med-PaLM's answers.

The granularity with which the researchers assessed the long-form answers from Flan-PaLM, Med-PaLM, and the clinician experts was among the study's strengths. "It's one thing to just report an overall accuracy number," Sontag said. It's quite another to contextualize when the errors are made, how bad they are, and to understand their implications, he noted.

Those findings helped identify areas for future research to narrow the gap in performance between clinicians and LLMs. Prompting the models to cite their sources in answers and to convey their uncertainty around a response are 2 avenues that researchers could pursue to mitigate incorrect responses, the study authors wrote.

'Computer Plus Human'
The Nature article joins a large and growing body of research showing that LLMs such as ChatGPT perform well on many medical questions, from passing the USMLE to generating high-quality and empathetic responses to patient questions posted online to outperforming medical students on free-response clinical reasoning examinations—Chen's latest work. Although Sontag believes that this technology will be part of larger artificial intelligence systems that will shape health care, he isn't convinced that being able to automatically generate answers—especially those that could be answered quickly and correctly with a simple Google search—will, in itself, change medical practice or improve patients' lives. Many of the studies that examine how well LLMs answer questions don't address how—or if—the quality of patient care changes when clinicians use LLM support or they go it alone.

"It's not computer vs human," Stanford's Chen noted. The more compelling question is, he said, "now that you've got this, what about computer plus human? What could you do?"

That question—how can an expert clinician work with artificial intelligence to achieve the best combined decision-making—should guide future research, Sontag said.

Lee, of Microsoft, agreed. Once LLMs started scoring well on responses to medical questions, the act of answering questions to indicate medical knowledge became less of an important practical matter, Lee noted. So he and his team shifted to studying GPT-4, an advanced version of ChatGPT, as a tool to help clinicians streamline clinical note-taking and draft after-visit summaries to patients.

In prompting GPT-4 to generate those notes and summaries, it's not the model's lack of medical knowledge that has been the challenge. What's more difficult has been getting ChatGPT to only include medical information that was supported by the clinical encounter, rather than "hallucinating" information such as calculations for a patient's body mass index, Lee said, or to remember to include comments in notes to patients like, "Good luck on becoming a grandparent for the first time next month."

Visions for the Future
Given their limitations, many LLMs, including Med-PaLM, are not ready for broad integration into clinical workflows just yet, and certainly not without a human in the loop to verify their work. "The bar for safe use has got to be very, very high because medicine and human health is a safety-critical field," Karthikesalingam said.

Plus, questions surround how to regulate models that are used in clinical settings and how to protect patient privacy and human autonomy. National and international groups are examining these and other questions.

However, experts say the technology might soon be used to streamline nonclinical administrative and operational work. A lot
of information in medicine is unstructured, Karthikesalingam and MIT’s Sontag noted: long, convoluted clinical guidelines; patient notes; and medical records. Clinicians could use LLMs to navigate that information. The models could pull a specific answer out of the guidelines, extract information from clinical notes for billing purposes, or create a medical encounter note that clinicians could edit. Research has also shown that many of these administrative tasks contribute to clinician burnout. LLMs may have a role to play in mitigating burnout by saving health care professionals time that they can reinvest in caring for patients, Sontag noted.

**Hands-On Approach**

Within this fast-moving area of medicine, experts say the clinician’s role is to embrace a hands-on approach—working with the models, figuring out how they operate, and developing a sense of where they fall short. “I feel very strongly that the medical field needs to fully own the questions of whether, when, and how this technology should be used,” Lee said.

“The faster that docs figure this out—how to use this, what’s safe and effective—and keep an open mind that this stuff is changing every month, every quarter, the better off we’re going to be as a field,” Butte said.