Clinical Chemistry

Generative Artificial Intelligence: More of the Same or Off the Control Chart?

Article:
Christopher Snyder, Mark A Zaydman, Thomas Chong, Jason Baron, Jonathan H Chen, and Brian Jackson.
Generative Artificial Intelligence: More of the Same or Off the Control Chart?
Clin Chem 2023; 69(10): 1101–6. https://doi.org/10.1093/clinchem/hvad129

Guest: Dr. Mark Zaydman from Washington University School of Medicine in Saint Louis, Missouri.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett. ChatGPT. Many of us have heard of it and some of us may have even interacted with it but what exactly is it? We are mostly familiar with the term "artificial intelligence" or using a computer to execute tasks that we typically consider to be performed by a human but what about "generative artificial intelligence?" These new large language models, including ChatGPT, represent the latest advance in AI and have attracted substantial interest due to their remarkable ability to generate human-like responses. So how reliable are these new tools? Is the content they generate factually correct? Can their power be harnessed to make a positive impact on health care delivery, or do they pose too great a risk to data security and patient privacy? Are these models likely to establish themselves as indispensable tools that we rely on heavily to complete our daily tasks or simply just another technological novelty likely to fall by the wayside? A new Q&A session appearing in the October 2023 issue of *Clinical Chemistry* discusses generative artificial intelligence, explains strengths and weaknesses of current models, and proposes potential future applications in the clinical laboratory.

In this podcast, we are pleased to welcome one of the moderators of that Q&A session. Dr. Mark Zaydman is an Assistant Professor of Pathology and Immunology at Washington University School of Medicine in St. Louis. Missouri. His research interests center on the applications of data science to improve the quality and efficiency of laboratory services. Dr. Zaydman, let's get basic first. What are generative large language models and how do they work?

Mark Zaydman: So, to answer that unfortunately there's a bit of terminology we should cover. So, these are a form of what's called artificial intelligence. And what that means is, we have a computer that's doing something that a human is doing, in this case, processing language and generating text and dialogue. Now specifically, these are a subtype of AI, artificial intelligence, called machine learning, and this is where the AI learns how to do its task without being explicitly taught how

© 2023 Association for Diagnostics & Laboratory Medicine

to do so. And how that works is the model is going to be learned from data instead of from a human instructor.

Okay, so we're talking about machine learning and specifically these are what are called neural networks. So these are a type of machine learning algorithm that has been modeled after layers of neurons like we have in the human visual cortex. And it's important because the type of algorithms can learn really complex, high-level patterns in the training data and that's what enables these models to learn complex patterns such as context within language.

Okay. So then we get to large language model. As the name suggests, we have a large model because it has up to hundreds of billions of parameters that was trained on a very large data set. Think internet scale data. And then language model because the task that these models are trained at is processing language. And then the generative part refers to the fact that these are going to actually generate new written or spoken language in response to user prompt. So that's how we get generative large language model. ChatGPT is one famous example of a generative large language model produced by OpenAI. Technically speaking, GPT is the model. Chat refers to a user interface with that model but there are others such as Meta AI's Llama and Google's Bard. So that's what they are now. Now how do they work? In general, these models are going to use their knowledge, the different patterns of language and context, that it learned from the training data to predict what text should follow a user's prompt. So in our Q&A article, our experts, several of them at least evoked this analogy of auto complete on steroids. So, think about when you start typing into Google and it predicts, you know, what's the end of your word, what's the end of your sentence, well, these models are going to go even beyond that. They're on steroids because they're not going to stop at a word or sentence, they could even do a paragraph, an essay, a book, you name it. So our experts remarked on how incredible that this kind of dumb sounding process of 'predict the next word, predict the next word, predict the next word' leads to these amazing high-level capabilities like the ability to synthesize a document, even to create a poem and add emotion and styling. And these outputs can be so remarkable that it's often difficult to remember that the model actually has no intrinsic conceptual or topical understanding of the content matter.

And so, how they work, you know, honestly you can't predict for any specific output exactly why the model made that output, but at least in terms of the key ingredients for making this work, it appears to be the sheer scale of the model in terms of number of parameters, the amount of training data, and then also this model architecture that allows it to learn

© 2023 Association for Diagnostics & Laboratory Medicine

not just the statistical association between individual pairs of words but higher-order associations that capture context.

Bob Barrett: So doctor, how are these different from conventional machine learning?

Mark Zaydman: So they're quite different. Conventional machine learning, you typically are developing a model that's going to map a set of inputs to a set of outputs. So, think the inputs could be all the clinical data for a patient and then the output could be sepsis/not sepsis. Now this is very different. So generative AI is actually creating novel content in response to a prompt. One of our expert said it's like you have an output and now you create the inputs. My analogy for this is conventional AI is going to say, take all your emails and say spam/not spam. Generative AI is going to actually write you an email, okay? So very different tasks. Another difference is that conventional machine learning models, you often train these models for a specific task by giving it labeled data. In contrast, these large language models are more general, they were trained on these massive troves of unlabeled data without a specific task in mind. And so as a result, these models can be used for language tests across all different application domains with essentially zero or just a small amount of additional training.

So you might hear the term "foundation model" and think of it that that this is a general-purpose language model that could serve as the foundation for all different types of AI applications. So there's a lot of jargon in there but hopefully that will clarify for the listener what a generative large language model is.

Bob Barrett: And most of the explanation has been about language in this large language model. So how can this be used in laboratory medicine?

Mark Zaydman: Yes. So our experts did a great job of pointing out that everywhere in lab medicine, there are language tasks that are quite tedious, and that these large language models could be a great productivity tool. And so think about language tests that involve review by some domain expert. So drafting and formatting documents, text processing such as data mining unstructured data from the electronic health record. Think about chatbots replacing some of the work of customer service and training, think of even possibly templating test interpretation as long as we're going to have some expert review. So, those were ways that these foundation models, without any additional training, might be used in lab medicine. We could think more into the future if we could take these foundation models and make them become slightly more expert in laboratory medicine and sciences. Then we might get to a state where the model itself could participate

© 2023 Association for Diagnostics & Laboratory Medicine

in direct question and answer with the patient, and we could even then entertain perhaps a model providing a primary interpretation or diagnosis. These are more future-looking applications as we're not quite there yet.

Bob Barrett: So are there any risks or safety concerns related to using generative large language models in laboratory medicine?

Mark Zaydman: Absolutely, absolutely there are. And these were again brought up by our experts. So I'll tell you about a couple. The first issue is related to accuracy. So these models are trained to produce convincing novel language content. They're not trained to produce accurate, necessarily, dialogue or responses. And this has been studied. So there was a nice publication by Dr. Sarina Yang that was published previously in *Clin Chem* where they took ChatGPT for a spin. And so they provided it some lab medicine-specific questions and what they found was that ChatGPT's responses were often inaccurate. This is not surprising because again these were general models that weren't trained specifically on lab medicine as an expertise domain. And actually they don't even have access directly to the data in the training corpus. And so they can't tell you necessarily the correct answer or why they produce an answer. And this can be made even more problematic because the answers are so convincing and there's no output from the model that says how certain it is.

All right, so that's clearly not acceptable for clinical care because inaccurate answers could potentially harm patients.

Another problem is reproducibility. So these models are stochastic, which means they're going to produce different outputs if asked the same question over and over. And there's actually a little knob you can turn in terms of how unpredictable or how predictable the model should be. Now, that could be really good for some users. For example, if an author wants some really creative ideas for a new novel, then turning up that unpredictability knob might lead to some really interesting and creative outputs. That's not typically what we want in medicine. In medicine, we want reproducibility, and we want predictability. Another concern is privacy, and I cannot emphasize enough that users should not be putting protected health information into ChatGPT. That is not a HIPAA compliant tool and by doing so you are exposing information to a third party. There are ways to access GPT and other large language models in a HIPPA compliant manner and I would refer that interested listeners should really talk to their local IT and informatics specialists and should check with their institutional guidelines for appropriate uses.

A final concern that was brought up is that of fairness. As you can imagine, when you review all of the language on the

© 2023 Association for Diagnostics & Laboratory Medicine

internet, you come across stereotypes and biases and hate speech, and those ideas can be integrated into the model. And then when we apply them, they could propagate those stereotypes and biases in ways that may harm abused and under privileged and marginalized groups within our society. This is even worse because often it's difficult to detect those unfair decisions since the models are somewhat opaque. So I think those are the big concerns that we have right now, accuracy reproducibility, privacy and fairness.

Bob Barrett: Thank you so much there. This is a brave new world. Tell us a little more about the Q&A article. Who were the experts and were there any highlights of agreement or probably even more fun, disagreement?

Mark Zaydman: Yes, we were lucky to have four experts who are all actively practicing and publishing at this interface between laboratory medicine, informatics, and applied digital technologies. Specifically, we have Thomas Chong from Baylor College of Medicine and Texas Children's Hospital; Jason Baron from Mass Gen, Harvard Medical School, and Roche Diagnostics; Jonathan Chen from Stanford University; and Brian Jackson from ARUP Laboratories. I'd also like to just briefly mention and acknowledge my co-moderator, Dr. Christopher Snyder, a great CP resident at Washington University in St. Louis, with a background in machine learning.

You know, Bob there wasn't a terrible amount of disagreement. There appeared to be some varying opinions with respect to how disruptive this technology really is going to be both in the near term and down the line. But there were really interesting recurrent themes that illustrated a general consensus and emphasize some very important points. I think everyone in the group agreed that this is an exciting advance that's worth paying attention to and not just a parlor trick that used to play you know, fun games. They all thought that this could be a useful productivity tool and help us be more efficient with our work. They all emphasize the risks and that appropriate use case selection is really key as we look to move forward with large language models and they also emphasize that there are currently critical gaps that need to be addressed with respect to the regulatory ethical and validation frameworks that are missing. So that was the Q&A expert group and kind of the general themes that came up in our piece.

Bob Barrett: Well, finally doctor, let's get personal, are you currently using generative large language models in your own work?

Mark Zaydman: Actually, I am. I found a couple use cases, thus far. So first, I'm faculty in the section of pathology informatics at Washington University and we do a lot of coding, we're developing novel software, and we're doing active data

© 2023 Association for Diagnostics & Laboratory Medicine

analytics. And so, I found that ChatGPT is a very useful coding assistant if I want to bring up a new web app or use a package I'm not familiar with. I can go right to ChatGPT and ask it to template what that code would look like. And then I can dialogue back and forth and have it modified the parts I don't like or add in details. I think this is really nice use case. We have kind of a narrow constrained subject domain where the model seems to do very well. Errors are very detectable. The code is either going to run or not run. It's either going to pass our rigorous testing or not pass testing.

And then we have experts that are reviewing the model's work. And so I think that highlights a lot of the aspects that were brought up by the experts in terms of what defines a really good use case and critically, I find that I'm more productive when I use the model compared to if I don't, even though I have to review its work. It kind of provides this rich input-output interface for me where previously I'd have to go to Google and figure out the magical right set of keywords that would provide me the answer I'm looking for, versus with ChatGPT, I can just speak very plainly to it, express my problem, and refine and converge upon a solution.

So coding assistant. The other use case that I have is a research project that I'm very excited about. So the problem we're trying to address is that information about laboratory testing is typically presented at a reading level that is far above the typical reading level of our patients. So we know that patients in the US are typically at about the sixth grade reading level and when we don't provide them accessible descriptions, that leads to misunderstanding and difficulties as they navigate their health journey. And so our hypothesis is that we can fine-tune a generative large language model to produce a description of laboratory tests that are both accurate but also more accessible to the patient. What really drove this is I was told I should present information in sixth grade English and I sat down and I tried and I discovered that I don't know how to speak to a sixth grader. And that's where a language model is a perfect way to transform my information in my language into a more appropriate reading level.

So these are the use cases we've defined so far but I'm really excited to see what other folks come up with in the field.

Bob Barrett: Well, we can definitively say this, that this is actually you speaking now. So this is not any ChatGPT or anything.

Mark Zaydman: Yes, I can confirm, although that's exactly what a GPT would say, right?

© 2023 Association for Diagnostics & Laboratory Medicine

Bob Barrett: That's true. That's true. All right, well thank you so much. We appreciate the time. I mean, this is just the beginning of this, isn't it?

Mark Zaydman: I think it is. It will be interesting to see how things evolve as people define new use cases and identify new challenges with this technology. But I'm sure this won't be the last time we have new technology that takes us by surprise and we have to re-evaluate how we move forward.

Bob Barrett: That was Dr. Mark Zaydman from Washington University School of Medicine in St. Louis Missouri. He served as moderator of a Q&A session discussing generative artificial intelligence and its application to laboratory medicine. It appeared in the October 2023 issue of *Clinical Chemistry* and he's been our guest for this podcast on that topic. I'm Bob Barrett, thanks for listening.

© 2023 Association for Diagnostics & Laboratory Medicine