

# AI-GR Pod 27 02.13.25 Alsentzer

[00:00:00] We use these cases from NEJM Healer, a medical education tool. We wanted to look to see if we provided these cases to GPT-4 and we swapped out either the race or the gender of the patient mentioned in that case, how well would the model perform in actually identifying the correct diagnosis? What we found from this work was that when you change the demographics for these cases, it affected the model's ability to provide the correct diagnosis in 37% of the cases that we evaluated.

One case that is really striking, which was a case of pharyngitis in a sexually active teenager where the model got the diagnosis, which is mono, correct for 100% of white patients, but only 64 to 84% of the minority patients where it opted to rank the STD, gonococcal pharyngitis, first instead in those cases.

[00:01:00]

Hi and welcome to another episode of *NEJM AI Grand Rounds*. Today we are delighted to bring you our conversation with Professor Emily Alsentzer. Emily is an Assistant Professor of Biomedical Data Science and of Computer Science at Stanford University. Andy, she is an expert in many things, natural language processing, understanding the large language models and other AI tools.

And I think she really leaned on her training in this episode, both in technical computer science, machine learning work, but also in deep clinical expertise and respect of the work of physicians and biomedical scientists. So, she went through this training program that I happened to go through as well, HST, and she talks about that and she really connects a lot of disparate topics in a very unified and interesting way in our conversation.

I really enjoyed it. I agree. We talk about this in the episode, but I remember I met her when I was a postdoc with [00:02:00] Zak. And I just remember thinking like, oh, this is what it's like when you meet a rock star, when they're a grad student. Like, I've just always known that Emily was on a sharp upward trajectory.

I've had the pleasure of collaborating with her on a couple of projects when she was a grad student. I've worked with her a lot in the machine learning for health community. So, she's been an organizing force behind a lot of the conferences that bring together clinicians and machine learning researchers. And just like, you know, of course she's a Stanford professor.

Like that just makes sense. Like I've never had any doubts that she was going to just kill it and be a total rock star. So, it was great to have her on the podcast and to talk to her about what she's working on now and what she's worked on in the past. And like you said, Raj, I think one of the most thoughtful, deeply

technical NLP researchers who also takes like the clinical side very seriously. And I think that's a credit, both to the HST program, but also, her mom's a pediatrician. She's like kind of grown up in it and she really blends all of those worlds and like a really, really nice kind of way.

The *NEJM* [00:03:00] *AI Grand Rounds* podcast is brought to you by Microsoft, Viz.ai, Lyric, and Elevance Health. We thank them for their support.

And now we bring you our conversation with Emily Alsentzer. Alright, Emily, well, thanks for joining us on *AI Grand Rounds*. We're excited to have you today. I'm really excited to be here with you both. Emily, welcome to *AI Grand Rounds*. So this is a question that we always get started with. Could you please tell us about the training procedure for your own neural network?

How did you get interested in artificial intelligence and what data and experiences led you to where you are today? That is a good question. So, I've always been interested in medicine, but I think I sort of stumbled into computer science and AI. I grew up in a family of doctors. My mom and brother are both pediatricians, so I'd really been exposed to medicine from an early age.

And then it was in high school, I think, that I attended this [00:04:00] program called the School for Science and Math at Vanderbilt, one day a week, instead of going to my normal high school. And it was during that experience that I think I was first exposed to science. So, we'd have scientists from across Vanderbilt come and talk about their work, and talk very interdisciplinary, talking about anything from ice cores in Antarctica to mechanical engineering.

Really taught, I think, the value of interdisciplinary thinking and the status quo. But at this point I had never done anything with computer science. And then during my last year of high school, I read this book called *The Most Human Human* by Brian Christian. And this was really my first exposure to AI.

So, this book, oh, Andy has it there. I love that. This book, for those of you who are unfamiliar, it describes the Turing test, which is this competition that assesses a machine's ability to exhibit human-like intelligence. So, a [00:05:00] judge will conduct a series of conversations over a computer with either a real person or a computer program, and the AI can pass the Turing test if it

appropriately fools the judges into thinking they're conversing with a real person.

But what is really cool about this book is that there's also this most human human award for the human that does the best job of swaying the judges. So, this was just really cool for me, kind of my first exposure to AI. And I think that was, in hindsight, priming for when I entered college at Stanford. I immediately decided to take an introductory computer science class.

Something like 95% of students end up taking a single class at Stanford, and then just immediately got bit by the bug, where there was really something satisfying about being able to build something from scratch. And I also found the idea of decomposing like a larger task into functions to be really like satisfying with my way of thinking.

So, [00:06:00] then ultimately decided to major in computer science, but really I think my interest in computer science were always in the service of medicine. So, at the time I was premed for three years and really thought I was still going to go to med school. And then, you know, again, in hindsight, I think there is this turning point where I took Dan Jurafsky's NLP class at Stanford, and he was just so good at explaining these really complex topics in an easy-to-understand way.

And I ended up doing a Latin translator as part of the class project for that class. My secret, I guess, is that I'm a Latin nerd. So anyway, I ultimately decided I was more interested in developing tools that could aid clinical decision making rather than becoming a clinician myself. Just to hop in here, please tell me that the first thing that you translated was cogito, ergo sum.

Oh, yeah, definitely. There's the joke that Latin kids know like 20 different ways how to kill someone, but you don't know, like, my [00:07:00] favorite color is purple. But anyway. After kind of wrestling with that decision, whether or not to go to med school or not, um, I ended up doing a co-term at Stanford, which is effectively a one-year master's degree in biomedical informatics and then joined MIT for the Ph.D. program.

And the Ph.D. program was in health science and technology. And this ended up being really the best of both worlds where you take computer science classes but also take some of the med school classes at Harvard med school. And even get this clinical experience where you learn how to perform a physical exam, you present patients on rounds, and write notes in the EHR, all as part of this Ph.D. program.

So that's awesome. We have, like, we've had several alums of the HST program, I think, on the podcast so far. I'd like to, so I was, I was with you there, you know, for the, for the listeners, what Emily was reacting to is the book, *The Most Human Human* she was describing is literally on the bookshelf behind me.

So that was like a fun [00:08:00] coincidence. So, I think spiritually I was with you there. You obviously found HST, which is like a good blend of those. The common motivation of like, how do we actually help people in society was, was part of it. And then really wrestling with what does it look like to interact with the medical field?

Is that actually being a practicing doctor, or is that helping interact via helping develop medical decision-making tools? Nice. Yeah, that makes total sense. Raj, were you going to ask a question? It looks like you were. Yeah, go ahead. Yeah, I was going to say, so I think we're going to get into it in a little bit, but you know, you do a lot of work in natural language processing.

And so, it just sort of stuck out to me that you said that one of the more important classes that you took during undergrad was an NLP class. Do you trace back your interests or even your work now to that initial exposure in NLP as an undergrad? Definitely. I think so. It was just a really well taught class.

I also think from the clinical side, there's something really satisfying about reading [00:09:00] clinical notes where you understand the decision-making process of a clinician that you get better insight into that decision making than you do if you just look at the structured data of an EHR loan.

So that was the other, I think, component that brought me to that kind of work. And you were interested in NLP like long before GPT-4 and Chat GPT and everything came out. This is all pre-large language model days, right? Exactly. So, it has been a lot of fun, uh, to see the field change. Everybody is now, everybody is now an NLP researcher. Right?

Exactly. More the merrier. Bring it on. Yeah, I do want to come back to that at some point about your thoughts on, or maybe we can just hop into it now before, uh, your papers, but like back in the day, there are these like highly bespoke NLP pipelines where like named entity recognition, like all of these different parts, you know, I think those of us who are around in the early 2010s, like remember cTAKES, and like entity extraction, and mapping everything to codes.

And you were super [00:10:00] happy if your new model, like, got an F1 score bump of like a 5% or something like that. But there was maybe, especially in natural language processing, I think the, the language part of NLP has gotten lost a little bit in that, like you actually had to think about the structure of language, grammar, syntax, not to go to the big picture thing too early, but do you think we're in a better place now?

Essentially, you just ask GPT-4 to do all of your extraction. Do you have any nostalgia at all for like the pipeline-based approach for how NLP used to work? I do. I think that we need to bring back some of those principles into the way that we continue to do NLP today. What I do appreciate about our current approaches is that I think we're much more closely aligned to the ultimate challenge that we're trying to solve in medicine.

Where because of where our methods are now, we're thinking about how do you summarize patients medical records? You know, how do you [00:11:00] do ambient dictation? Whereas when I first started in the field, I found, being motivated really by the ultimate clinical challenges, found it a little harder to see the big picture of how you go from named entity recognition to a downstream clinical task.

So, I think that's the more exciting part of where we are today in the field. Awesome. So, I think that's a great transition. So, I'd like to talk about one of your papers. It's called "Publicly Available Clinical BERT Embeddings." There's a couple of things that I'd like for you to highlight here. Why are there Sesame Street characters in the title of the paper?

So, a little bit history of, of BERT models might be helpful. What were you trying to accomplish in this paper? And, uh, yeah, let's, let's start there. Sure. So, uh, BERT is the name of a model. Before BERT, there was ELMo. So, there's been a long history of Sesame Street themed characters that kind of, I think, ultimately create in this community.

And it's been a lot of fun. There's been ERNIE after BERT. [00:12:00] But, to set the stage here, you know, we created this paper, developed this paper in 2019. So, this was prerelease of anything close to ChatGPT. Several papers, I mentioned BERT, had just been published demonstrating the utility of self-supervised learning.

So, for those who are unfamiliar, this is the idea that you could pretrain a model on a ton of unlabeled text data by training the model to predict a word from the surrounding words in the sentence. And in doing so, the model learns to encode

the relationships in that text. And then that pretrained model can then be used for a number of different downstream tasks.

So, these BERT models were really powerful, but they were largely trained on general domain knowledge from the Internet. And we saw the, all these papers and thought we really need to develop a model that is adapted to clinical text, specifically clinical notes in the electronic health care record.

And there's a number of reasons, really, why you [00:13:00] want to do this. I'm sure many of the clinicians know that these notes contain medical terminology and abbreviations. They have incomplete sentences. The text is often semi-structured, often written using note templates. Interpreting the information in those clinical notes, such as the meaning of lab values, requires domain knowledge that may not be present in Internet data.

And then furthermore, the patient presentations in EHR data don't necessarily look like the classic presentations in medical textbooks, or the atypical presentations that you might see in PubMed case studies. So, there were a number of reasons why we thought, okay, we really need to adapt this model to clinical data.

And I think the interesting part of this work was not necessarily the methodology, but the fact that we were able to publicly release this model on Hugging Face model hub and make it available to the [00:14:00] community. Where we demonstrated as part of the work that these specialized clinical models outperformed their general domain counterparts and then made it available for others to use.

That's awesome. I want to like do a little bit how the sausage gets made here because I think that there's a really interesting story. So, like just to be transparently flattering to you, I think that this has been a hugely impactful paper. And when I go over to GoogleScholar, as of now, it's been cited 2,378 times.

So, it's received a tremendous amount of attention. I know that a lot of people use this model. I think that what may be surprising for clinical listeners is that it wasn't published in like a traditional way. So, I think that this was published at a NAACL workshop, if I remember correctly. And so, like not in a big fancy journal.

And it really is kind of like a public good in the sense that you have trained this model for others to use. Could you talk a little bit about the sort of publication

story behind this and how the community has built on this? Because I [00:15:00] do think there's an interesting like resource use message here.

Yeah, I think that, in many ways, it was really good timing, where we had seen this paper come out, and really wanted to make, I think really the motivation from the start was to make a resource available for the community. This was my first exposure to open source and the open-source community. But a combination of Hugging Face really just getting started at that point that made model sharing of these kinds of models really easy, combined with just the huge leap in performance we saw with BERT-based models that I think kind of led to this model getting a lot of use.

I guess some back story there, too, is we wrestled a lot with whether we could actually make this model publicly available given that it's trained on EHR data. Ultimately, talked to the team at Mimic from where the data was trained on and decided it was okay because it's an [00:16:00] encoder-based model, meaning you're learning to represent text without necessarily generating text, but that was a serious discussion point when we were first releasing the model.

Awesome. Yeah, I think I was still a postdoc, right, when you were working this. I remember like hearing how this all came together. So, I remember sort of the origin stories of this paper fondly. I think that transitions naturally into your next paper that I'd like to talk about, which is, "Do We Still Need Clinical Language Models?"

And I think it'd be fair to say that your BERT model is a clinical language model in that it's a model that is specifically trained on clinical text data. The question that you're posing here is, does GPT-4 obviate the need for anything specialized? And that's like the question under investigation here.

So maybe you could walk through how you looked at that and what your conclusions were. Yeah, sure. So, yeah, we, like many other people in the field at the time, were thinking, gosh, do we actually still need these models that we had created earlier? [00:17:00] Does ChatGPT just obviate everything and do everything for us?

So that was really the focus of this paper. And in particular, we wanted to take the perspective with this work of a potentially resource constrained hospital that wants to leverage the benefits of language models using as few resources as possible. So, hospitals typically have a few options for leveraging language models.

They can create a specialized clinical model through pre training on their own data, like ClinicalBERT, for instance. They could fine tune or, you know, further train a publicly available language model, or you could do a prompting-based approach of a general language model. So, we asked several questions with this work.

First of all, do specialized clinical models outperform these general domain models of comparable size? Can the specialized clinical models outperform larger general domain models? And does training on clinical data actually produce more cost [00:18:00] effective models? And then finally, do these specialized models actually outperform prompting based approaches?

So, methodologically, we decided to try to hold as much as we could constant. So, fix the model architecture, and then compare general and clinical language models of varying sizes to try to answer these questions. So, the real takeaways from this work were that we found that pretraining on clinical text allows for smaller, more parameter efficient models that can either match or outperform these much larger language models trained on general domain text.

So, you could get better performance with a clinical language model that is 3.5 times smaller, for instance, than a larger general domain model. And then we also found that the computational cost, again, of training these models, these clinical models, is much smaller to achieve the same performance as a general domain model.

And then finally, we looked at what if [00:19:00] you only train these smaller clinical models on a handful of examples? How does that compare in performance to these prompting-based approaches? And it turns out that even by only training these models on a handful of examples, you can actually surpass the performance of prompting-based approaches. Now, I do want to caveat that the tasks that we looked at with this paper were all like classification-based tasks, or what is called extractive question answering, where you try to identify the span in the text that answers a given question.

These were not generative tasks. But I think the takeaway for me with this work was that before you immediately turn to the largest, latest and greatest model, consider what smaller more specialized clinical models can do. We have one more paper that we want to discuss, but I kind of want to, like, do a little interlude here because I realized that we stopped your gradient descent, personal gradient descent procedure a little [00:20:00] prematurely. You are now faculty at Stanford in biomedical data science.

Am I getting, that's the correct department? Yes. That is correct. So, I think like one thing I'd like to ask you here is like, you and I worked together a little bit during my postdoc when you were a grad student, and we were working on the USMLE problem. And we were like, way under resourced to be able to do that, it turns out in hindsight.

And I think that you have in your recent work shown how there's still a place for small models. But now that you're like starting your own lab at Stanford, how do you think about what projects to work on given how quickly everything changes? Again, we, some, some of the work we've done, we were doing a ChatGPT paper right when it came out and then they released ChatGPT-4.

And we had to redo all the results. So how do you think about what's a six-month durable question that you can ask in this space? Yeah, that's a great question and something I think we're all kind of wrestling with a little bit right now. I try to make sure [00:21:00] that any of the projects that we work on aren't dependent on a specific version of a specific model, especially one that is closed source.

I'm particularly interested in there's a class of models where they're not only open source, but the training data itself is also made publicly available. And so that lets you do a number of interesting experiments where you can try to tie the behavior of the model to the data that the model was trained on.

You always run into the challenge there of someone coming to say, oh, but maybe that doesn't apply to these much larger language models. But I find that that kind of work potentially leads to more generalizable conclusions. Do, do you think it's fair, Emily, to say that you're trying to understand something about how these models work when they fail, how brittle they are?

That it's about the kind of science of the models themselves, and of course some engineering tasks along with that, as opposed [00:22:00] to just sort of staying in the benchmarking and what is the latest and greatest current performance of the new incarnation of the models. I think that's fair. I think there's a need for both.

Like we need, even on the evaluation side, better methodology to be able to evaluate these models in a scalable way. And I think there is a role for thinking about how the models are actually deployed in real world clinical settings where you're guided by the actual workflow that these models will be evaluated in.

But personally, I want, I think, at least part of that research to your point of looking at the science of these models is really important. Right. And I think even for that task of figuring out how to deploy the models and where in the workflow they should be, how they should collaborate with humans, my sense is that the most interesting papers that we would want to work on are about trying to find durable sort of scientific takeaways, insights that are likely to hold [00:23:00] up even after the next incarnation of the model comes out or the next version of the model emerges, right?

Like, I think what Andy's getting at is that the pace of just performance gains has been so rapid that previous benchmarks or evaluations that we thought were likely to last for many years or be aspirational targets very far away have now been saturated, right? And they've been saturated in a way that makes it actually hard and challenging to even plan as a guy, as someone who is running a lab or someone directing a team, but I think they are less of a problem if you're focused on trying to do the science of either the collaboration of how these models work or where they fail, or even what nuanced evaluation looks like.

But basically, I think Andy actually said this, even when we were starting *NEJM AI*, I give Andy a lot of credit for articulating this early and very clearly, just something along the lines of even what we're interested in publishing are things that [00:24:00] are durable, right? That they're going to last longer, that are going to be relevant even after the next incarnation of the model comes along.

That's a very nice way of saying that I was unrealistically combative. That was a very gracious interpretation of that. Yeah. Yes. So, Emily, I want to transition to another one of your papers. This is the last paper we'll talk about before the lightning round. And this one is, I think also, also broadly NLP related but studying sort of a different topic and very important topic.

And this one was published earlier this year. So, this is a paper that you led that's published in *Lancet Digital Health*. The title of the paper is "Assessing the Potential of GPT-4 to Perpetuate Racial and Gender Biases in Healthcare: A Model Evaluation Study." And so maybe, I think the title is quite informative, but maybe just to, to get started, you can tell us about the motivation for the study, the backstory, how you got started, and what question you're really trying to answer [00:25:00] with this paper.

Sure. So again, to set the stage here, at this point, I was doing my postdoc at Brigham and Women's Hospital and had gotten involved in operational work within the hospital where Brigham, as well as many hospitals across the

country, have been starting to deploy language models into the clinic at really, I think, unprecedented speed.

I think much of the current focus has been on automating administrative tasks, but many clinicians also envision using these language models for support as well. There was actually a survey that came out recently that it was something like one in five respondents reported to using generative AI tools in their clinical practice.

And of those, like 28% said they were using it actually to suggest a differential diagnosis. So, we saw all of this uptake in language models. I have to say just it's just based on my straw poll and sample of residents and clinical colleagues, [00:26:00] that sounds totally believable.

I think it might even be higher amongst people I'm talking to at the hospitals here and residents at the Harvard hospital. So, it's unsurprising. It's already being used in differential diagnosis regularly. Yeah, completely agree. So, we've seen all of this usage of these models as part of clinical decision making, but we knew from all of this substantial research in the past that language models can encode societal biases that are present in their training data.

And so, our goal with this work was really to evaluate whether a language model perpetuates those racial and gender biases, specifically focused on a number of different clinical applications. We focused on GPT-4 because it's been one of the most popular models for the clinical community, and we looked at how the model performs for a medical education task, for diagnostic reasoning, for clinical plan generation, and for this sort of subjective patient assessment.

So, I'm happy to go into [00:27:00] the details about some of those, if that's helpful. Yeah, no, that'd be great. So, I, I think there, the case series that you used as sort of the primary focus of the paper were these NEJM Healer cases, right? So maybe you could tell us about that data set and like what the cases look like and what that initial task was and then how you varied it to study racial and gender biases.

Sure, so we use these cases from NEJM Healer. NEJM Healer is a medical education tool. It will present an expert generated case, and then it allows medical trainees to compare their differential diagnosis for that case compared to some expert generated differential.

And so, we wanted to look to see if we provided these cases to GPT-4 and we swapped out either the race or the gender of the patient mentioned in that case, how well would the model perform in actually identifying the correct diagnosis for that case. [00:28:00] And something that we did was select cases that actually should have a similar differential by race and gender.

So, for example, we would exclude cases of lower abdominal pain, which you would expect should have a different differential for female and male patients, for instance. What we found from this work was that when you change the demographics for these cases, it affected the model's ability to provide the correct diagnosis in 37% of the cases that we evaluated.

I think there's this one case that is really striking which was a case of pharyngitis in a sexually active teenager where the model got the diagnosis, which is mono, correct for 100% of white patients, but only 64 to 84% of the minority patients where it opted to rank the STD, gonococcal pharyngitis first instead in those cases.

Do you think the behavior of the model and the biases and problems associated with how it performs [00:29:00] and how it can be steered is this largely a problem with the training data, diversity, or bias in the training data? Or do you think these types of biases enter into other phases of model development?

It's a good question. I think most, this is speculative, but I would imagine that most of the data is actually coming through the training data. We know that even the, for instance, medical education cases that are in the literature are themselves biased. So, the model is learning to pick up on these co-occurrences that we're seeing in the training data.

Now, there is another potential source of bias that hasn't really been investigated that much yet, which is the human preference training process as well. So, there's this notion of reinforcement learning from human feedback or other kind of ways of training these models. And right now, we actually don't have a lot of insight into who are the people who are generating these [00:30:00] preferences?

What are the kinds of tasks that these preferences are being evaluated on? And the whole goal of that process is to try to steer the model to produce favorable outputs. But you have to ask, you know, favorable to whom or by what standards? And I think that is a very much an underexplored area. Yeah. I think this whole area is like, so important, but like so difficult because like when

something like this surfaces, I could think, okay, well it's read too much of the Internet and therefore that's why it's biased.

Or okay, the human annotators that they did use to do the alignment had systematic biases. Or I think it could be like structural bias in the way that doctors are educated or in the, you know, the medical data that the model has seen. And so like, I don't have a sense of like nihilism here. But I kind of like when I see something like this, like I don't have an immediate, okay, well, here's how we fix it because it could be coming from so many different places.

So, I wonder how you [00:31:00] think about the difference between like description and prescription here. So, like there are descriptive studies that can at least articulate what the problems are, but how then do we actually think about fixing them? Yeah, it's a good point. I think we first need the descriptive studies before we can address the prescriptive studies.

There are a couple of ways that the NLP community has tried to address this. One is by selecting your pretraining data in a better way. That is much harder to control. There's also another class of where, how you change the loss function that your model is actually trained on. And then I think this third category, I'm particularly interested in, which is at inference time, how can you change the probabilities that the model is outputting to change the prediction of a model?

So, we as end users who don't necessarily have exposure to the entire pretraining process, what can we do [00:32:00] at the end when we're using these models to try to de-bias the model for a different prediction task? I think it's much easier to think about this in the context of a particular application of a model rather than trying to de-bias any potential use of these models.

And that's why I think we often talk about how these models may be used in a human-in-the-loop process, but I think it's very unlikely for individual clinicians to be able to spot these biases when they're only looking at individual cases. Part of what I'm arguing is that we need targeted fairness evaluations for each use case of these models, even if they can be used, you know, for a number of different applications.

That actually gets exactly to the sort of final question I want to ask you about this paper, which is, you know, there's work happening to try to mitigate these biases. There are new incarnations of the models that are being built, and then we're exploring various scaling laws still, right, as [00:33:00] a community at both training time and at test time.

But in this sort of intermediate or the immediate time, right? We also just talked about this a moment ago, these models are already being used. And I think they're being used potentially and likely at massive scale, right? Today by both patients and clinicians. And so, I'm wondering if you can, maybe just direct a few concluding thoughts, particularly for clinicians.

So, from this paper, what do you think are the key clinical takeaways? Takeaways for clinicians as they're using these models, as they are writing either hard cases or their own experiences or their own medical data, and they're looking to GPT-4 and to other models for advice. What do you think your paper's sort of key takeaways are for that crowd?

I think if you're a clinician using these models, it's really helpful to just [00:34:00] put your mind in the context of what these data are actually trained on. So, we know that these data are trained on Internet data. That could be Reddit data, that could be Twitter, or any sort of social media data, that could be random blog posts.

And so, as a result of that, these models are learning all of those associations. So, any sort of bias that is present could be present in these outputs as well. So, I would approach these outputs with a healthy dose of skepticism, and especially related to more sensitive areas of how these models should be used.

And I'll also just remark that this work and others since it have focused on evaluation of these models in structured output. We focused on diagnosis, which is a structured task. We also focused on evaluating specific demographic groups, [00:35:00] but there is very little work right now in terms of evaluating uses of these models that are generative outputs or looking at potential populations where that population may not be defined by a specific group.

So, for instance, like health literacy, how could that impact performance of these models? And so, I'm at least not aware of any audits for these real-world clinical applications that are being deployed.

For instance, tools like the drafting responses to patient messages, or Epic is releasing a tool to summarize patients medical records. So, I think clinicians should recognize that that kind of evaluation hasn't happened yet, and again, approach these with a healthy dose of skepticism there. Awesome. It's time for the lightning round.

[00:36:00] So, the first lightning round question, Emily, I'm going to give you a rare exemption because it's going to require some setup for you to be able to

answer it in a coherent way. You'll understand why once I asked it, okay? What did your experience with a new pair of blue jeans teach you about the challenges of automatic diagnosis?

Wow. I was just thinking about this, Andy. So, context, because Andy likes to bring back shameful moments. Many years ago, when I was still a Ph.D. in Zak Kohane's lab, I approached Zak, you know, somewhat timid at the time, because my fingers had turned blue, and I was thinking to myself, you know, this has been happening for a couple days in a row, you know, Googling, figuring out what could be going wrong.

I thought, oh, there's a syndrome called Raynaud's Syndrome. Maybe I have [00:37:00] that. Let me talk to Zak, you know, he has a lot of connections in the area. Am I going crazy? And he, he looked at my fingers and he was like, yeah, that's kind of weird, I can, sure, I can put you in touch with a clinician friend of mine. And then, it wasn't until after that connection that had been made that I realized that also during that time I had just gotten new blue jeans, and those blue jeans had blue dye that was apparently coming off into my fingers.

And so there was actually, this totally unrelated reason out in the world of why my fingers were turning blue. So actually, Andy, in preparation for this call, I put that case into ChatGPT. And not knowing that you were going to ask this question, the model did not do very well at determining that it could be due to my blue jeans and their dye. I actually tried to do the same thing, but I couldn't find the picture.

And I feel like if I just described it, I wouldn't describe it correctly. [00:38:00] But to me, that was like such a salient example of the sort of like long tail of diagnostic tasks that we might like want a human to do or an AI to do. I thought what was amazing is that you put it out on Twitter and you got responses from like world class rheumatologists and like diagnosticians.

And I think no one said blue jeans. Like it was like, so I actually thought it was like an amazing test of like human diagnostic ability. And like, what are these like great, like teaching moments and like what the, the edge cases might be. I think the fun thing too, it's kind of reverse causality because my fingers were cold.

And so that's why I put my fingers near the blue jeans. So yeah. I will say I got it right, but that's only because I had seen it happen to Kristen, like the week before, like literally the week before, uh, I had seen the exact same thing

happen. I was like, did you get a new pair of blue jeans? But again, like I think about that case all the time when we're thinking about AI and medicine.

So, this was just a complicated [00:39:00] case where Andy got the diagnosis right, and most of the human experts missed it. Yeah, so, Rogers, Crick, and Diagnose, why? All this setup! Thank you. Oh man, was Andy actually pointing this out, how you figured out that it was the blue jeans and not anything else? I don't remember.

I think you had already figured it out by then. I think I said it on Twitter or something and you or Sam Finlayson sent me a message with like a winky face or something. Like you guys had already figured it out for sure. Wow. Amazing. Alright, Emily. The next question is, if you weren't in academia, what job would you be doing?

Ooh, good question. Um, Alright. This is still academic adjacent, but a very different field. During college, I spent a summer in Monterey, California. Stanford has a marine station there. I was doing research related to disease modeling and a network, very [00:40:00] computational, but a lot of the other students at the marine station, were all doing like marine biology where they got to go scuba diving for their work every day.

And I think that would be a pretty cool job of just getting to be out in the water as part of your work. Nice. This one is a reflective question. So, what is the thing that you have changed your mind the most about since you were younger?

Ooh, that is a good question. Okay, two, two things. One, I used to hate olives. Now, I love olives. Two, you know, going back to the decision about whether or not to go to medical school, one of the reasons why I decided I didn't want to go ultimately was because, oh, I thought I wouldn't really like interacting with patients.

I was like, you know, quieter, like I'll, I'd rather [00:41:00] just think about the decision-making process. And then as part of HST, we did these clinical rotations where you had a lot of time to talk to patients and I ended up loving that experience. It introduces you to people who would be totally outside of your bubble.

You get to learn about their stories and that was a really humbling experience. What changed with respect to olives for you? Like what bit flipped there? That is a good question. I think I was having bad olives maybe. Had them on pizza

and with other flavors and it just totally switched. Alright. So, I think it's funny that I get to ask this question.

I did not write it, either a language model or another type of intelligence, like my co-host wrote it, but I'm entertained that this is the question I get to ask. Emily, which do you prefer, Harvard or Stanford? Oh, that's not fair. I'm gonna plead the fifth on that [00:42:00] one. Alright, fine. I love Zak.

Yes, fine, fine, we'll allow it. Yes, there was a misaligned AI, Andy Intelligence, that um, wrote that one, so. They're both great. They're both great. And great people at both institutions. Next question. If you could have dinner with one person, dead or alive, who would it be? Ooh. Alright, I'm gonna take a cop out answer and describe an imaginary, or like a fictional person that I would like to have dinner with.

I think it would be really fun to have dinner with both Sherlock Holmes and Willy Wonka for different reasons. Like Willy Wonka, the creativity of that, that world, being part of that world would be really fun. And then Sherlock Holmes, like I am very curious what he would intuit from our conversation.

If you were to round it out and add a third person, who from The Wheel of Time would you add? Oh, I mean, Egwene. Obviously, Egwene. I don't even know what you guys are talking about right now, [00:43:00] so I'm just gonna say it. Alright, last question of the lightning round. Will AI and medicine be driven more by computer scientists or clinicians?

I unfortunately don't have a hot take for you here, in that I think it has to be the combination of both. Or, you know, the cross trained scientists. I think the HST program really taught me that, especially, you know, as a more on the computer scientist myself, understanding what clinical workflows look like, understanding that data generating process has been really valuable to think about how I shape what kind of questions I ask.

And then on the flip side for clinicians to understand what is actually possible from a technology perspective is, is really useful. Congratulations on passing the lightning round, Emily. Whew! We threw you some curveballs there, so you did great. Okay, so now we're going to zoom out and ask you some big picture questions, uh, to wrap up.

So, we touched on this a little bit [00:44:00] before, but I just, like, want to kind of explicitly drill down on, like, you know, as an academic, as someone who's focused on clinical research questions, how do you feel about beholden is

probably the wrong word, but the fact that there's a few labs on the planet who are pushing the frontier of AI, and a lot of what happens in academia for us is that we are interrogating and or using those models.

Is that something that we should have mixed feelings about? Or is this just a new like research paradigm where they're building the big sort of like microscopes and we're using the microscopes to ask a different set of questions than methodologists would have previously asked? Yeah, I think it's this paradigm makes it really challenging for us as a community to understand what the generalizable and lasting contributions are.

You know, anytime I'm a reviewer, that's kind of the, the hat I have on, especially looking at when you're leveraging models that are closed [00:45:00] source. I think a lot about what is the role of academia in this landscape? And we touched on this a little bit earlier, but we need careful evaluation of these models for real clinical use cases.

And we also need methods to better evaluate in a scalable way. I also think that if we consider these models as black boxes, thinking about development of methods that allow you to quickly tailor these models to particular clinical settings as well is also really interesting. And then I think, broadly, as a community, I think we should encourage more transparency in the training data for these models.

There was actually a new law that was recently passed in California that says that from January 2026, developers of AI systems must publicly post on their website certain information, at least, about the data used to train these [00:46:00] systems. I think it's unclear exactly how much of that data will be revealed because there's a real business case.

But I think as a community, can we put pressure at least to, and figure out what is the most useful information about the training data that needs to be revealed for us to do these, um, you know, clinically useful evaluations and to help us understand what are the associations that the model might be learning and how that impacts their downstream use.

Got it. And then finally, I mentioned this earlier, but encouraging research using open models as well. Like, there's an example of Dolma, which is an open-source dataset. And a corresponding model called OLMo, which is trained on that data set. So, leveraging resources like that. So that's great. And like a follow up question to that is that there are some indications that performance may be saturating.

So at least the benchmarks we use to test models like GPT-4 [00:47:00] and Claude, that there's diminishing returns for the current training paradigm. Do you think that that's happening? Or do you expect like GPT-5 to be significantly better than what we get with GPT-4. Like, are we saturating sort of the scaling curve that we've been on by training bigger models and more data?

I think the new paradigm, for lack of a better word, for inference time scaling is interesting. I also think though that we still don't have the best evaluations to actually probe the understanding of these models in a way that is not prone to training on the test data, for instance. Especially in the clinical domain, I think we've only scratched the surface in terms of evaluation.

Most evaluation is synthetic cases, as opposed to real world cases. Time will tell when we actually have better evaluations to probe the utility of these models. I guess final question for me before I hand it off to Raj, I [00:48:00] think what has been exciting for academics and folks, you know, not at one of these frontier labs is the evolution of the open-source ecosystem.

So, you know, uh, Meta has gifted us in some sense with a 5 billion gift in the Llama3 models, because that's putatively how much it costs to train, those large language models. And we can use them in a relatively unrestricted way. Is your sense that open source will eat closed source in that like the, it will be hard to keep up with how vibrant the ecosystem is?

Or do you think that closed source models, uh, because they're so well resourced where are sort of forever going to be in front of the open-source community? I think generally that the open-source ecosystem will continue to stay just behind closed source, but continue to stay just behind it and catch up with it.

I think it will ultimately come down to what resources and pre-training data also you have license to as another key component. I do want to comment a little bit on the medical open [00:49:00] source AI ecosystem, which is, you know, a much smaller subset. Right now, all of the medical generative models are trained on datasets like PubMed, textbooks, or other sources of medical knowledge.

And there are still a few generative AI models that are trained on clinical notes. And as we discussed earlier, you know, clinical notes are very different from the text in PubMed or in textbooks. One reason for this is due to privacy concerns, right? Our algorithms for de-identification aren't perfect and sharing data across hospitals is hard.

If I had to speculate, I think we will see in the future clinical generative language models trained on synthetic data and released into the open-source community. We already have that a little bit with GatorTronGPT, which is the closest example I've seen to this. Amazing. Emily, last question. And it's actually a pair of questions [00:50:00] for you.

Just thinking ahead for the next five years or so, what are you most optimistic about and what are you most pessimistic about, about the use of AI in medicine. Alright, I'll start with the most pessimistic. To end on an optimistic note, you know, I'm overall an optimist, but I think to, to reflect back to our earlier conversation, what I worry about the most is automating the biases or propagating errors in medical data into these models.

I don't think we have the appropriate evaluation yet to fully understand the harms of using these models at a larger scale. And the concern is that we won't actually detect them until they're already baked into our systems. So just to give you an example of this, there are a number of companies trying to pilot tools to summarize medical records, but I think there's a lot of nuance in terms of how to do this correctly, all of the medical data that these tools are summarizing has [00:51:00] copy-and-paste errors, different sources within the EHR have different levels of trustworthiness, all of that sort of information that I'm not aware of being currently baked into these existing tools.

I'll give you another example where we've been evaluating language models for diagnosis tasks that leverage gene names, and we found that the model performs a lot better when the gene names are gene symbols compared to Ensembl IDs. So, for those who are unfamiliar, gene symbols often have A through Z characters in the name.

Ensembl IDs are just a string of numbers and language models are notoriously bad at representing numbers. So, it turns out if you use the Ensembl IDs, models get worse performance. I mentioned that as one example that there are many different potential gotchas that require really careful, rigorous evaluation that I think we've only scratched the surface on.

And I think our appetite and the possibilities in this space are outpacing our ability to conduct these [00:52:00] evaluations in a rigorous way. And then going back to your question on optimist, a couple answers to this question. From an application perspective, I'm really excited about the idea of language models enabling patients to be more active in their care.

I know you all recently had a mom and doctor on the podcast where the idea that you can have language models as an advocate in your care. I've been very fortunate to have clinicians in my family, which has been a huge privilege where they can come to any meeting and say, are you sure you haven't thought about, you know, X, Y, and Z?

They serve as advocates. Can we imagine what a language model based advocate when you're in a room would look like for enabling patients? And then I think underappreciated application of language models is phenotyping. I think there's a lot of opportunities for these models to change the way we define cohorts, perform medical research in the future.

I'm also, from a research perspective, excited about what are the questions [00:53:00] we can ask now that we have audio recordings of patient visits? As an NLP person, I am very aware of the idea that we're looking at this data through the lens of what a clinician has already interpreted and what billing processes are already need there's certain things to be documented.

So, these audio recordings give us this really unique insight into what the patient is actually saying in the room. And then finally, there's an increasing number of training programs in this space that really recognize the need for interdisciplinary training. HST is one of them. The AI and Medicine program at Harvard is another.

The Computational Precision and Health program at Berkeley and UCSF is another. And I think those are training the leaders that we need in this space to drive this technology further in the future. Amazing. I think I said that would be my last question, but your answer was so interesting that I have to ask one more and this, I promise will be the final question.

So, you know, you just started a lab, right? You just started your lab at Stanford, new [00:54:00] professor. Congratulations. I think it's, it sounds like it's going very well. How do you think, language models, GPT, you know, developments in, in just the tools that we all have access to now is going to change your job over the next few years.

And a perfectly valid answer is it's not going to change it over the next few years. But I'm curious, cause I'm just, you know, talking to a lot of people about this and I'm getting very different answers from, I can only be a professor for a few more years, you know, it's going to automate everything and science is going to be discovered by LLMs to eh... it's going to help with my grants a little bit, but like not really change things that much.

What's your take on how these models and how the technology that we have access to is going to change your life as a professor let's say over the next five years?

I think in the next five years, I lean closer to the, it'll make me more efficient at my job, but I don't see it [00:55:00] replacing any significant aspects of my job. I already use language models to help with careful wording of individual sentences, or writing code I think is probably the most significant change, that it just accelerates anyone's ability to learn a new programming language or new skill.

I think there's so much intricacies in terms of how these models are deployed and going back again to the data generating process that we know by knowing the data and the application really well, and I don't see a language model replacing that component of the research process. Excellent. Alright.

Thank you so much, Emily. That was fantastic. Thanks, Emily. This was a lot of fun. Great to chat with you guys. This copyrighted podcast from the Massachusetts Medical Society may not be reproduced, distributed, or used for commercial purposes without prior written permission of the Massachusetts Medical Society.

For information on reusing [00:56:00] NEJM Group podcasts, please visit the permissions and licensing page at the *NEJM* website.