AI-GR P18 05.09.24 Nigam Shah

[00:00:00] The point we made in the article is that all of these existing models are not trained on medical data and are not instruction tuned to follow up or respond to what we ask it to do. And then we just say, summarize this patient record for me and feed it a patient record. It has never seen a patient record and no human taught it what a good summary looks like.

[00:00:25] And then we expect it to work. Maybe we should be using our medical textbooks and curated material like UpToDate, ClinicalKey, or what have you, and then further tuning them on EHR data. It is possible to curate the EHR data going into that. So, just like, for example, clinical guidelines are based on results of RCTs or observational studies, which are itself based on some cleaned up version of EHR data, we could conceive feeding in a diet of good quality records to learn from. [00:01:00]

[00:01:00] Care as it should be, not care as it is.

[00:01:06] Hi and welcome to a new episode of *NEJM AI Grand Rounds*. I'm Raj Manrai and I'm here with my co-host Andy Beam. And today we are delighted to bring you our conversation with Nigam Shah. Nigam is a Professor of Medicine at Stanford University and the Chief Data Scientist for Stanford Health Care.

[00:01:23] Andy, this was a really fun conversation. I thought that Nigam had a lot of insights, both about working with data that comes out of the health care system, as well as how to navigate collaborations between clinicians and machine learning scientists. All-in-all, this was a really fun and insightful conversation.

[00:01:40] Yeah. First and foremost, it's always fun to talk to Nigam. So, it was great to finally have him on the podcast. Obviously, he's a world class biomedical informatics researcher. He's published lots of really impactful translational papers, and I think actually really takes translation very seriously in a way that a lot of academics don't.

[00:01:57] For example, he also says things like [00:02:00] CapEx, which you don't normally hear a professor talking about, but he has this whole other side of his professional identity on the business of health care. And so, he helps Stanford's health care system think about how to deliver health care more effectively to better serve their patient population.

[00:02:15] So again, I consider him kind of like a full stack health care data scientist. And not only does he write the papers, he also cares about how the health care actually gets delivered. So, it was super fun to learn from him in this conversation. And yeah, it was, it was great to sit down with Nigam and hear what he's been up to.

[00:02:34] The *NEJM AI Grand Rounds* podcast is brought to you by Microsoft, Viz.ai, Lyric, and Elevance Health. We thank them for their support.

[00:02:48] And now we bring you our conversation with Nigam Shah. Nigam, welcome to *AI Grand Rounds*. We're thrilled to have you today. Well, thank you for having me, guys. Pleasure to be here. Nigam, so, this is a question [00:03:00] we always like to get started with. Could you please tell us about the training procedure for your own neural network?

[00:03:05] How did you get interested in AI? What data and experiences led you to where you are today? Well, that is a great question. Uh, there was a, not a great straight forward descent. Uh, there was a couple of spikes in the loss function along the way, so to speak, which basically means it was a meandering path.

[00:03:23] So I started out as a doctor in India, got my M.B.B.S. degree and I was going to be an orthopedic surgeon, believe it or not. And we have a family friend who got his Ph.D. in the U.S. in the seventies and knew me since I was a little baby. And he convinced me to try my hand at research and said, apply for a Ph.D. program.

[00:03:43] And if you don't like it, come back and you can be a carpenter. Like, what's your problem? You know, like, okay, he has a point. So, I applied, got into a bunch of places and that was year 2000. And the Human Genome Project kind of hit the mainstream news, and I got really hooked into [00:04:00] the notion of using computation to understand biology and medicine.

[00:04:05] And I was in a molecular medicine program, so I lobbied my committee to let me switch to the bioinformatics major. And when I finished, people on my committee basically said, people who have medical degrees and who do reasoning engines, so my thesis was a reasoning engine on yeast biology, said you should go to Stanford.

[00:04:26] So I came there as a postdoc and never left. That's an interesting fork where you were on a trajectory to be an orthopedic surgeon. What was it about that moment in your life that made you want to take a hard left and go get a Ph.D.? Mostly the influence of this gentleman, Dilip Desai, who's been a family friend for decades.

[00:04:47] And his argument was that he thought, I'm good at logical thinking, and he said, you can be a carpenter, or you can try this thing, and if you don't like it, in one year you can come back and be a [00:05:00] carpenter. Is carpenter shorthand for being an ortho bro? Yeah, pretty much. Okay. So, you were too smart for orthopedic surgery was his main concern it sounded like.

[00:05:13] Or, you know, we could spin it many different ways. There's a bunch of ortho jokes in there. But overall, the point was that try this new thing which I had no exposure to in India. And they said, if you don't like it, you can always come back. And the experimentation cost is not that high. So, I think that the general advice was to, if you don't know, give it a try.

[00:05:34] Looking back, did you make the right decision? Oh, like I am thrilled. Maybe I would have been happy fixing hips and knees. Like, I don't know, but I'm having a blast. OK. So, at Stanford then, you did a postdoc. Could you tell us a little bit about who you worked with and what you did during your postdoc?

[00:05:50] Absolutely. So, it was with Dr. Mark Musen, who several of our listeners might know. And my thesis project was about building a reasoning engine [00:06:00] for yeast. And at that time in 2005, I honestly believed that we would have figured out how to structure medical data into knowledge bases and everybody would publish facts instead of just prose.

[00:06:12] And Mark was leading the center called the National Center for Biomedical Ontology, which was one of the 10 or 12 NCBCs. And that center's mandate was to create a portal that would catalog all the different standard terminologies and ontologies and ways of representing structured biological knowledge.

[00:06:34] And I was like, if I solve this one, you know, we're going to have knowledge basis on which we can reason over like in two years. And so, I was like, yep, that sounds like the best thing to do as a postdoc. And that's why I joined. So, this might be a question that's better posed at the end, but we think about different eras of AI.

[00:06:53] We think about the expert era and like ontologies and reasoning systems are definitely part of that clade of AI. [00:07:00] How do you think about the role of ontologies and knowledge representation in the current LLM moment? That is a great, great question. I think ontologies will make a comeback. There are already things like graph neural networks.

[00:07:16] And if we're able to store prior knowledge in a structured format in a tinier footprint, we don't have to spend the compute to relearn it from unstructured prose. So far, I would say it's still a little bit of a research area as to how do we correctly inject the prior knowledge that might be in medical terminologies like SNOMED and RxNorm and whatnot.

[00:07:43] Into the learning process of a language model. Well, I guess before we, we hop over just one more thing is that like, this is one of the big open questions in AI right now is how much world knowledge do you need to explicitly model in the system? So, it seems to me that you're one of the people who's well positioned to [00:08:00] unify the expert era of AI with a large language model.

[00:08:03] So I'll be interested to see what you do with that going forward. Yeah, well, I'll keep you guys posted. So, I think that's a great transition point to your work now, Nigam. So, we want to talk about both threads of what you do, both your research arc, how you run your lab, as well as a new role that we understand that you have, or maybe not, you know, maybe a year or two now at this point.

[00:08:25] And so in addition to being a professor at Stanford, we understand that you are the inaugural Chief Data Scientist of Stanford Health Care. We thought maybe we could start there before we dive into your research. And so maybe you could tell us, you could just begin by telling us what the Chief Data Scientist at Stanford Health Care does and why you signed up for the job.

[00:08:45] So the one-line version of the job is to bring AI into clinical use safely, ethically, and cost effectively. And I mean, we hear so much about AI these days, right? I mean, [00:09:00] language models and image processors and whatnot. And I encourage our listeners to sort of do this thought experiment. If you look at 100 companies or 500 companies that are selling some AI solution and add up what they believe is their total addressable market, I would not be surprised if it's larger than the total cost of health care in this country.

[00:09:24] The point being that yes, these new technologies are amazing, but we have to find a way to make them sustainable. And sustainable not just from a financial standpoint, but from a burden on our physician's standpoint, from value to our patient's standpoint, and just the cognitive complexity of managing everything.

[00:09:43] So that's the reason it's a big, heady problem, as one of my other mentors, Larry Hunter likes to say, if after tenure you don't take on something that would have otherwise gotten you fired, then, you know, you're wasting tenure. Can I just hop in here? You say words that I don't hear [00:10:00] academics say often, like TAM, Total Addressable Market, CapEx for Capital Expenditure.

[00:10:04] In your role as Chief Data Scientist, how did you learn, that part of, what is essentially the business of health care? It's a great question. So, what did you do, you know, for academics who are interested, actually, in making a similar move? How did you learn that vocabulary and that skill set?

[00:10:19] So mostly by experimentation and not on the job. Prior to the job. So, the joke on our campus at Stanford is that you need to spin out a company as a part of getting tenure. And I'd spun out two companies before I took the job, actually three companies before I took the job. And, uh, as part of the founding of those companies is how I learned the business of how to think about problems, both health care and outside.

[00:10:47] And, you know, we might make fun of businesspeople, especially when we're scientists and doctors, but the discipline that they have to follow. is worth learning as we approach scalable [00:11:00] solutions or seek scalable solutions to problems. I also think the concept of product market fit holds for academics, even if you aren't actually starting a business, that you actually have to have a solution to a problem that someone actually has.

[00:11:13] And so you probably learn, even outside of the economics and business models, lots of reusable skills in spinning out companies. Absolutely. 100%. Yeah. Nigam, could you talk about your approach for working with clinicians on these projects and bringing AI into health care at Stanford? So, I have a very bimodal sort of schizophrenic view on that, and I will give both versions to our listeners. One view is that you partner with them, and you understand their problems and attempt to build a solution that solves the problem that they articulate or that you observe. You know, this notion of product market fit that, uh, Andrew was just mentioning.

[00:11:53] But the other view is why don't we look at the overall structure of [00:12:00] what we're trying to do and imagine a solution that would not be visible to somebody who's in the midst of it at the frontline. And I try to do both. It’s cognitively quite challenging, but if we don’t, then, you know, as the cliche goes, if you ask people what they need, they’ll tell you faster horses that poop less.

[00:12:20] And so I think as in others in my, in my role, uh, we, we sort of have to try to balance this. Yes, let’s listen to the people on the front lines, clinicians, nurses, pharmacists, but at the same time, let’s listen to the problem behind the problem they’re telling us. So, we can come up with the solution that they really need but cannot articulate.

[00:12:41] So a way to think about that, would it be fair to say you need to listen to the problems, but maybe ignore the proposed solutions? Yes, yeah, I think that’s a great way to put it. Okay, great and where are you in piloting AI in clinical use now at Stanford Health Care? So, is this something that you [00:13:00] see, and maybe the answer is both, maybe you can give us some examples, but is this something that you see more as an application on the kind of back end or administrative work that’s involved in the delivery of health care?

[00:13:12] And I’ll include in that even things like writing notes and documenting what’s happening during the patient encounter. Or are you piloting and excited about technology to assist with diagnosis and the actual provision of care? And tools to equip your physicians to help them. Second opinion or another take on, on a given case.

[00:13:36] So, I bucket things into the use of AI to advance the science of medicine, the practice of medicine, or the delivery of medicine, health care basically. And the allocation tends to be heavier on the health care side, mostly given the way our legal system and risk tolerance [00:14:00] plays out. So, it’s a lot easier to deploy a solution that is producing a bill in an automated fashion than there is to deploy a solution that is doing automated diagnosis.

[00:14:13] So as a field, I think as we start, we go after the low risk, quote unquote, safer cases, so to speak. But that doesn’t mean we don’t experiment with the hard ones. Because if we have to deliver on the promise of AI to improve access and improve equity in care, we have to arrive at solutions that automate parts of things that physicians and nurses and pharmacists currently do.

[00:14:43] But I wouldn’t start there. And so, you know, we, we got to nibble at the edges first, but in the research, we got to dive in, into the complex problems. And so again, it, it’s a little bit bimodal in that sense that when I walk over to what I jokingly call my [00:15:00] .edu self, the, you know, jeans and t-shirt self. We want to take on the problems that are five-to-10 years out in terms of their solution.

[00:15:07] And then when I go to my .org self, which is like formal pants and a shirt with black shoes, then we're talking about things that have a one-to-three-year horizon in terms of their impact. And having that both perspectives is super fun, actually. Can you tell us, for your .org self, uh, about some of the pilot AI studies that you are running, where you see the frontier, where you are in implementation?

[00:15:34] So, there's a couple of pilots that are publicly announced. I'm not running them, like my team would be in a supportive role there. There's one that, you know, everyone's heard about using GPT-4 to produce responses to my health messages or patient messages. So, this is basically the patient portal messages come in and using GPT-4 to draft a response that a human would then review.

[00:15:57] And, you know, without sharing too many [00:16:00] details, it's under review right now, the amount of work doesn't change as much, but the providers feel happier. They feel supported. And their cognitive load on their brain goes down. So, it is helping in that manner, just not in the manner we would have anticipated up front.

[00:16:18] Staying in the sort of health care delivery realm, I saw you give a talk once where you made a distinction that once I heard it, I was like, ah, of course. And it was the distinction between efficiency and productivity. And that hospitals actually probably only care about productivity. Could you, if you remember the talk, could you walk us through that distinction and why it's an important one for AI?

[00:16:42] Absolutely. So, this is the one that often gets me into trouble on occasion. And let's just work with the MyHealth, the patient portal example, right? Let's say I as a physician, or you know, you're working as a doctor, and you're getting 200 messages a day. And right now, you [00:17:00] work between 9 to 10 p.m. or 8 to 10 p.m.

[00:17:03] to respond to those messages. So-called pajama time, so to speak. We can have an AI solution that works beautifully and completely takes away that work. Okay, let's just imagine that the 9 to 10 work pajama time is gone. From an access to patients standpoint, nothing has changed. They're still getting the exact same responses and no new patients are getting care.

[00:17:31] And so from a productivity lens, have we using the same resources produced more care? The answer is no. In fact, you could argue in a very perverted sense that productivity has gone down because that AI solution is an additional expense we have to pay for. And then you can do some math and say, well, you have a happier physician, less burnout, less turnover, and that has some value, and hence you can show like marginal productivity. [00:18:00]

[00:18:01] So that is the example where it is an amazing efficiency gain, like infinite efficiency gain if it just worked out of the box. Negative or marginal productivity gain. Now this is also an example of, you know, you ask the people at the front lines, that is their pressing need, and there's good product market fit, and it is a solution that people are asking for.

[00:18:27] But if we take a step back and say we have to manufacture good health care for 8 billion people on the planet, is this the most important problem to solve? And then the answer is probably not. And so, when you get local, Palo Alto, Stanford Health Care, our 15,000 employees and our patients, this problem becomes top of mind.

[00:18:53] But if you're a new Ph.D. student, trying to impact the care of a billion people, [00:19:00] there's probably way other problems that are more exciting, maybe harder, but worth solving. And so, coming back to the exact question, efficiency versus productivity, if we just follow the media and the companies, they're all going after efficiency gains.

[00:19:21] Assuming they can be parlayed into productivity gains. And my main point is that that is not, that assumption is not always correct. Does this set us up for, so if you go after productivity gains, meaning you can produce more RVUs per physician if they're using AI, to a significantly less happy workforce though?

[00:19:43] Because now, you're doing the same job. You're producing more. Um, so what's the flip side of that? Yeah. So, it depends how we define productivity, right? I mean, if productivity is defined as in a very perverse way as more RVUs, we could destroy morale in six months if we [00:20:00] wanted to using AI. Uh, but if we define productivity as how many primary care needs were served per unit dollar spent.

[00:20:11] It's a different way to think about productivity. Why does a primary care, does it cost 150 bucks or 250 bucks or whatever it costs? Can we do it for five bucks? But we're not thinking that way. Like I don't see any AI company out there that says, I’m going to deliver a world class primary care visit experience for five bucks or less.

[00:20:32] Why do you think that is? Just the way the payments are made, right? I mean, they got to chase the payment and the existing budgets. And right now, there is no budget in any health system that says, we will deliver primary care visit for five bucks. Now there are companies out there that are trying to be completely disruptive.

[00:20:50] There’s one that we recently found out and I have no affiliation with them, which is like a pod that you walk in, and it delivers like the five or 10 basic primary care services. [00:21:00] Uh, and there’s no human anywhere in there. And it’s like 99 bucks a month subscription, and you get unlimited primary care for the five or 10 things they cover.

[00:21:12] I agree that access is important, and why should a primary care visit cost 150? If I follow that argument to its conclusion, it seems like the primary care provider as a position, seems to be destined for extinction because if you’re charging five dollars a pop, there’s no way there’s a doctor behind that.

[00:21:33] Maybe there’s an NP or a different type of provider attached to that. Are we heading, if we can actually provide primary care at scale like that, to a world where we don’t actually have primary care MDs anymore? I don’t think so. I don’t think so. So, if you look at history of technology, like ATMs were supposed to put bank tellers out of business.

[00:21:52] But after ATMs came around, the ranks of bank tellers went up. Cars were supposed to put people who drive other people [00:22:00] around out of business, but the number of humans who drive other humans around went up after automobiles came around. So, I think what will happen is that the things people do will change.

[00:22:10] Like, if a computer can manage somebody’s insulin dosage better than me, like, by all means, go to the computer. Use the human for tasks that a computer can’t do. Like, on the one hand, we have this massive physician shortage. On the other hand, we require a primary care physician to sign off on a statin prescription.

[00:22:30] Like, why? Makes no sense. I mean, nobody ever got addicted to statins. Raj – Do you have any follow ups? Uh, no, I, I think we can go to large language models. Okay. So, we, we’ve tried to forestall the large language model discussion for as long as possible, but we always, it was inevitable. We always end up here. So, I'd like to anchor on your recent perspective in *JAMA*.

[00:22:52] It covered a lot of ground. For our listeners, it's titled "Creation and adoption of large language models in medicine." I think for [00:23:00] one, it provides one of the clearest descriptions of technically what's going on with large language models. Wnd we've discussed them before, but we've probably done a disservice to our listeners and not really defined a lot of the key terms.

[00:23:12] So since you've done such a good job at this in print. Could you first walk us through the basics of how LLMs work and then maybe touch on uh, some of the advanced topics like instruction fine tuning and things like that? Yeah, happy to, happy to. Thanks for bringing that up. So, I like to explain language models with this example.

[00:23:32] Imagine there were only two sentences in the world. Where are we going? And where are we at? And then if I asked our listeners to calculate what is the probability of seeing the word going? After I've seen the three words, where are we? And most people can say, well, you know, one out of two, so 50%, 0.5.

[00:23:55] That's basically the intuition behind language models. Now just imagine playing this [00:24:00] game over billions of documents and trillions of sentences and calculating the probability, not just of seeing the word going, having seen three words, the probability of seeing an entire paragraph, having seen 20 pages prior to that.

[00:24:17] So those 20 pages are the context, and what you're producing is the generation part. And then imagine learning an equation that instead of having 3 or 4 probabilities in it, has a billion probabilities in it, or a hundred billion probabilities in it. That's what a language model is intuitively. Now, the beauty of this is that for a computer, language just doesn't mean words in English, Spanish, German, or anything of that nature.

[00:24:45] They could be any sequence of symbols. Could be amino acids, could be DNA, nucleic acids, could be sounds, could be codes, CPT, ICD codes. And so [00:25:00] anything that has the sequential structure where one symbol comes after the other, and there's some rhyme or reason, AKA a grammar to that. You can throw this technology at it and learn a language model.

[00:25:15] And when you've learned a language model, you get two things. Thing number one, which everybody's now familiar with, is you poke the language model with some words, and it tells you a few words. The chat interface, where the model's being used to generate new content. There's another way to use language models, which is you feed in a sequence of tokens, and it gives you back a vector of numbers, which is a representation, or embedding, of what you fed in into numerical form.

[00:25:47] Essentially putting your data as a line in a table. And both of these have value. We can use the generation mechanisms to have a [00:26:00] conversation with the language models. And we can use this embedding business in order to represent things, protein structures, documents, patient records, images, EKGs, in a pneumatic representation on which we can do computation, such as find similar patients.

[00:26:20] Predict what is going to happen next. Tell me how many days until a heart attack might happen. And so that's sort of the intuition behind writing that, like to explain to our clinician colleagues that what are these technologies capable of beyond just chat on the Internet. Thanks. Oh, go ahead. Yeah. And then the sort of the second part of that is that, alright, so now we have these technologies.

[00:26:42] Uh, there's lots of people building language models. What are they feeding to the model as it's learning? And turns out the majority of the things out there have not trained on EHR data. They've trained on something else. So, the [00:27:00] things that they produce also sound like, or read like, the something else that they've learned from.

[00:27:06] So that's one item. What are the inputs going in? And then second, as we're chatting, these things are just producing the words based on probabilities. To give a GPT-based example, if two years ago, if we had told GPT, explain the moon landing to a six-year-old, it would say, explain gravity to a six-year-old.

[00:27:27] So a bunch of humans had to sit down and say, like, no, no, no, GPT, that's not true. Uh, that's, I mean, it's true, but that's not correct. What we need you to do is to say, people went to the moon, and collected some rocks, and sampled, and took some pictures, and came back. Like, that is the right answer. So, we either show the right answer, or we ask it to produce five, 10 answers, and we pick the best one.

[00:27:50] All of this called instruction tuning or reinforcement learning with human feedback and whatnot. And the point we made in the article is that all of these existing models are not [00:28:00] trained on medical data. And are not instruction tuned to follow up or respond to what we ask it to do. And then we just say, summarize this patient record for me and feed it a patient record.

[00:28:12] It has never seen a patient record. And no human taught it what a good summary looks like. And then we expect it to work. So, if I was going to say that back to you. We currently use models that are trained on text data from the Internet. And you're arguing that maybe that's not the right set of signals or the right set of symbols to train a language model for medicine and we should, we should be using EHR data.

[00:28:35] Is that fair? That is one way to think about it, but maybe we should be using the general Internet or maybe we should be using our medical textbooks and curated material like UpToDate, ClinicalKey or what have you, and then further tuning them on EHR data, just like a med student. Like the med student doesn't educate him or herself on Reddit.

[00:28:57] Well, there, there are, uh, [00:29:00] subreddits where they prepare for the board exams and there are, like, communities like that where they actually do learn a lot from Reddit. But just the, so I, Nigam, I think the explanation is fantastic and I really liked the article that, Andy referenced that you wrote, but just to maybe push back a little bit on

[00:29:17] what it's trained on. Do we really know what GPT-4, for example, is trained on? Do we know that it doesn't have third party medical data, electronic health records from some locale, some country, as part of its corpus of training data? Uh, yeah, we don't. You're absolutely right. We don't know. They won't tell us.

[00:29:37] Uh, at least for the Llama models, the public domain models, they're not trained on EHR data. But for GPT, who knows? I guess too, one of the things that comes to mind is once you start training a model on EHR data, you're not training it to do medicine, you're training it to do health care. And so, as I'm sure you know better than anyone else, [00:30:00] the EHR data primarily exists to facilitate billing and reimbursement and is at times only loosely correlated with the actual clinical state of the patient.

[00:30:10] So how do you think about introducing all of the warts that we know exist in EHR data to something that only has sort of a platonic understanding of medicine? Yeah, no, that's a great question. But if we want to train a model to do billing, isn't that the best data to train off of? Right. You know, so exactly.

[00:30:28] So if you want to make GPT biller, that seems obvious to me, but if you want it to actually practice medicine, it seems less clear to me that a language model trained on EHR data is actually what you want. Yeah. And in that case, you know, we'd probably go towards training on medical textbooks and society guidelines and so on.

[00:30:47] And it is possible to curate the EHR data going into that. So just like, for example, clinical guidelines are based on results of RCTs or observational studies [00:31:00] which are itself based on some cleaned up version of EHR data, we could conceive feeding in a diet of good quality records to learn from. Care as it should be, not care as it is.

[00:31:13] I think it's also just like worth pointing out that most doctors only learn to bill after they become attendings and are therefore done with the training, that that actually is the last thing that they learn how to do after they've learned all of medicine. Absolutely. I mean, we go to the school of medicine, not to the school of health care.

[00:31:30] Andy, are you suggesting we should train our models the same way? There's at least maybe a non-commutative order of operations there for people. So, another thing that you mentioned, you know, again, putting on, I think this would be your .org hat is the value prop for LLMs in health care. You discussed that in this article.

[00:31:48] So how does that play in here? I mean, we had Mark Cuban on the podcast earlier and he said that the shine came off of LLMs very quickly for him and that essentially, he's just using it as autocomplete. And [00:32:00] so have we overestimated the value prop of LLMs in health care, or have we not fully understood what they're capable of yet?

[00:32:06] I think we're overestimating the value proposition. We're, we're kind of, yeah, I would agree with that assessment. And I think what the hard questions we're not asking is, what are the systems that we need to build? That can be driven by a language model and would have value. And the second question we're not asking is, do we really need to always use one giant language model for everything?

[00:32:31] And why do we not build specialist models? And then the third one I'd interject, and you brought up Mark Cuban, I mean, he's revolutionized genetics and drugs. Why do we want to be hostage, like we as health care and doctors, be hostage to technology companies with these closed source models? Why can we not pool our textbooks and our EHR data, which I agree, there's lots of egos and, you know, vested interest and so on, but just as a thought experiment, you know, if 10 [00:33:00] health systems came together and partnered with the top five medical publishers to first learn from the medical literature and then from their curated EHR data and put that model in the public domain, we could reduce the cost of using AI in health care. Is the answer to that not the same answer to why haven't they pulled their data and put it in the public domain though? So, I think there's a distinction because previously, there was no way to create the incentive to put the data out there. In this case, by sharing your data, you get value back in the form of being able to use that model to solve some problem that you had.

[00:33:41] That's true. So, you can create a shared resource that you wouldn't be able to do on your own. I guess it's harder to imagine that being as valuable as like creating a shared chest x-ray model or something like that, that probably has less value than you can imagine coming from like a big LLM that understands all of medicine. Yeah.

[00:33:59] [00:34:00] Raj, do you, do you want to follow up or should we hop to the lightning round? Let's move to the lightning round.

[00:34:15] Alright. The rules here of the lightning round are your answers have to be brief and they don't have to just be one sentence and they are a mix of silly and serious and it's on you to decide which category each question comes from. I actually think you've alluded to some of these already in your answers, but I'm curious to see if my language model has correctly predicted your response in this case.

[00:34:37] So what single person has had the single biggest impact on your life? Dilip Desai, the mentor who got me the U.S. My language model is well calibrated. If you weren't in medicine, what job would you be doing? I was going to be an astronaut when I was a kid. Very cool. Why did you give up on that dream?

[00:34:58] Turned out that when I came of age, [00:35:00] India didn't have a space program. A valid reason. A valid reason. Bootstrapping your own space program outside of people named Elon Musk seems like too much to do. Next question. Overrated or underrated? The AI scene in the Bay Area. The AI scene? Yeah, like, the AI ecosystem in the Bay Area.

[00:35:23] Overrated or underrated? Way overrated. Way overrated? Why? I think logistic regression has become AI now. Fair. Yeah. And, uh, our coefficients on the east coast are just as good as your coefficients on the west coast, so. Oh, absolutely not. Ours are better. They're at least, uh, probably more fit and better tanned, so.

[00:35:47] They have better weather. Better weather. Will AI and medicine be driven more by computer scientists or by clinicians? I would say neither. Unless our [00:36:00] clinicians, uh, we clinicians step up, it's going to be driven by the finance officers. Yeah. It's, it's the people who buy the product. I mean, just like EHR, no one ever got fired for buying Epic and no one ever probably will get fired for buying GPT-4.

[00:36:19] So, okay, next question. If you could have dinner with one person alive or dead, who would it be? That's a tough one. I would probably go for alive, and I'd probably go for Barack Obama. I think that that, I think we had, he's a popular choice for that question, uh, so far. I think that Euan, Euan Ashley also said Barack that too, yeah.

[00:36:41] Our first, first episode. Um, also a Stanford professor. He's also the only president I know that actually wrote, uh, a scholarly *JAMA* article. Oh yeah, he had, like, a *Science* paper and something else. Yeah, no, he had I think a set of them. I think he had, like, a double feature. Yeah. [00:37:00] In both of the journals.

[00:37:02] Um, alright, this is our last question, uh, and this one is a little bit out of left field, but I have to ask, thinking about what my dad has told me about this exam, should the U.S. adopt the Joint Entrance Exam for admission to colleges? And so just to contextualize this for our listeners while you think about that, the Joint Entrance Exam or the JEE is this notoriously hard exam that a very large fraction of the Indian citizenship or any Indian citizens who are entering college takes before they go to school.

[00:37:38] So sort of universal test across, across much of India. Should we have an equivalent of the JEE in the U.S. for college? With open to everybody in the world or just for the U.S.? Let's say open to everybody in the world. From an access standpoint, I would say yes. It might put a lot of people on edge and [00:38:00] defensive because when you have, a billion other people competing with 300 million people, just by sheer probability and priors.

[00:38:09] So, I would be actually say we should be creating a national entrance exam, at least at the country level. This whole madness of applying to 40 colleges one-on-one, it's just like, one, it's inefficient and it just creates these weird pockets of misaligned incentives where more people apply and everybody's admissions rate look like 1% or less.

[00:38:34] Alright. You fix the numerator, and we keep inflating the denominator, and everybody gives high fives, like what are we doing? It's a crime against math. The other thing about the JEE is that it's, uh, I don't know if this is still the case, but when my dad explained this to me when he was going through the exam and then into college in India is that it is much greater weight.

[00:38:57] It may be actually like the sole [00:39:00] criterion for admission, at least at the time into IIT and into many of the other schools. So that's quite different than the states here as well, where, uh, many schools are moving away from the SAT, but when, you know, let's say 10, 15 years ago when SAT was near universal or the ACT, one of the two was near universal in the U.S.

[00:39:19] Even then it was only a small, or one of the main, but one of the, one of the features to determine admission. So that's a, it's a very different model, right? To sort of use JEE as a way to rank, uh, everyone and then to decide on what schools you're able to go to. Alright. So, I would sort of say that instead of having an exam, we should have a national admissions system.

[00:39:43] I think that's like both other criteria, but it's, it's standardized at the sort of national level, right? It's centralized at the national level, but it's not solely a sort of written or technical exam. I would say my niece is applying to college now, and there is like a common application that you [00:40:00] can apply once and get applied to a bunch of different places, but it's not universal.

[00:40:03] Exactly, exactly. Like match for residency is near universal. The admissions processes are not universal, Andy, right? So even if there's a common app, the admissions is split over all the schools. Well, Nigam, I think you passed the lightning round. Congratulations. Thank you. Thank you. Alright. So, we'd like to zoom out a little bit and talk about some big picture stuff, um, with the time that we have left.

[00:40:24] You've touched on this a little bit already, but I think I'll rephrase how I'll ask the question. How nervous should we be about the AI ecosystem, given that it's essentially dominated by two large tech companies, both as academics, but also as patients. What should our nervous level be about that? I think it should be high.

[00:40:45] And I would say instead of nervousness, we should approach it with a matter of concern in the sense that are we willing to abdicate such a crucial part or seemingly future crucial part of our national [00:41:00] infrastructure to third parties over which there is no national control? Like we would not let just two countries run the entire electric grid of the country.

[00:41:10] Internet is not run the same way. And if health care is equally important, why would we cede control in that manner? So, it's not fear, but I think in, uh, it is more about national interest and maintaining equitable access and fair play. That we should all be concerned about it. I guess if we're going up to like the federal level, how do we, so, I mean, it's hard to get things that most people agree on done currently in the current political environment.

[00:41:40] This to me feels like almost like a big particle physics project where you need a CERN for AI in the U.S. if there's going to be an alternative to the ones owned by big tech. So, like, I agree with the vision. I just stop when I try to think about how we would operationalize that. So, like, what is the path forward there to [00:42:00] realize the vision that you just laid out?

[00:42:02] I think we need an external shock to the system. Like when Sputnik flew over the Americas in the sixties, that created the space program that put a man on the moon and bring him back alive. So, we need some external shock. Like U.S. and a lot of other, our institutions, like the places you're at and the place I'm at, we hate being second.

[00:42:22] So we got to tap into that and, you know, maybe we engineer an external event where like nationally there's this imperative like, oh, we're second. Come on, we got to do something. Yeah, I think spite is always a good motivator and maybe we can use institutional competitiveness to catalyze some of that. I think one of the things that strikes me as ironic in this moment is the only reason we have legitimate open source alternatives it's kind of because of Meta/Facebook, uh, they've published Llama and Llama 2.

[00:42:55] A lot of people have been built on top of those models. There are startups that make [00:43:00] open-source models, like Mistral, and MPT, and, you know, even the UAE has an open-source large language model. And these models impress me because they continue to punch above their weight. You know, if you're looking at, you know, accuracy per dollar spent on the model.

[00:43:15] They punch way above their weight. So how do you see the open-source ecosystem evolving and how do you, how do you think that folds into health care? Raj and I are both editors at *NEJM AI*, and we ourselves have written a lot of GPT-4 papers. You don't see the same level of health care investigations from the open-source models.

[00:43:35] Is that because the closed-source models are just better, easier to use, all of the above? How can we catalyze the open-source models in health care so that there's a legitimate alternative to the closed-sourced ones? Yeah, that's a great question. Something close to my heart. I think people are not experimenting as much with open-source because there's no good platforms

[00:43:55] on which you can easily experiment. People experiment with [00:44:00] GPT because there's ChatGPT, a browser-based interface where you can paste stuff in and do a project. We don't have that for the most part for all of these public models. The engineering lift to get it up and running is significant. So that's one barrier.

[00:44:17] But I think that barrier can be solved, like going back to one of our prior conversations. Imagine even 10 health systems or imagine insurance coming together, three large insurance companies, saying we will fund an open-source AI foundation model, large language model, for all of your billing, answering the patients, and so on.

[00:44:38] Because ultimately, all the payments come from insurance. And if we want to contain the total cost of ownership of AI, why would we want to pay the CapEx of having built these large giant models that somebody else built for whatever reason. And kudos to Meta to have released [00:45:00] Llama, which gives everybody a foundation on which to build a solution that doesn't cost as much.

[00:45:06] I don't need my patient message response model to also draw me a picture of a unicorn. It cannot have that capability and it's okay, but if it's 10x cheaper, like all the better. I think that's a really good point. Time dilation is a real thing here because ChatGPT is only a year old. Now it feels like a decade old.

[00:45:26] And the underlying model was still 3.5. They did some RLHF that you mentioned earlier to make it slightly more pleasant to chat with, but the barrier to entry was zero. That wasn't the big breakthrough and why it went viral is that there was a website that you could go to and chat with it and it just worked.

[00:45:42] I think that that, now that you say that, seems to me to be the missing piece here for the open-source side of things is we need a, uh, and Hugging

Face hosts some of these things too, but, um, it’s just not at the scale that OpenAI has done for the GPT models. And maybe, maybe that would be a good collective open-source effort to make [00:46:00] something like that that would be easy to experiment with.

[00:46:03] Absolutely. So, we are strong proponents of that. In fact, we just dropped two models on Hugging Face. At NeurIPS and MLHC, the embedding kinds of models, not the generated ones, not the ones you talk to. Because we need shared experimentation, which we don’t have. I mean, you have folks at Harvard, Matthew McDermott, pushing really elegantly on creating the right infra that we can all compare and experiment and share results.

[00:46:27] Like, we gotta do that. Yep, totally agree. But Nigam, it occurs to me that it’s, I think it’s something you just said, you know, moment before that, which is that why did ChatGPT get so popular, right? And what was it? It was, it wasn’t the existence of the model weights, right? It wasn’t the sort of API. It was this chatbot.

[00:46:49] It was this interface that I remember, being in NeurIPS in New Orleans last year. And one of my students just comes up to me and says, have you heard of this [00:47:00] ChatGPT thing? I was like, no, what’s that? Takes out his laptop. And then pretty soon everyone in the hallway there is just, interacting with it, asking it, it kind of breaks for a little bit, right?

[00:47:10] Time’s out. And then it just takes over the conference, right? Like in sort of the, you know, the cocktail party conversations. And it was this universal ability to just interact with it in so many different ways that I think immediately captured so much of our attention. And so, I like the idea of building similarly useful tools for opening or open models.

[00:47:29] Not OpenAI, but open models, right? Open source. Lowercase O there. Lowercase O. Lowercase O. Lowercase O, exactly. But also, like who would do that, right? So, who would, because it’s, it’s an engineering challenge, right? To build and support something like that. It’s one thing to upload it to Hugging Face which I think you deserve.

[00:47:46] Everyone deserves a lot of credit for doing things like that. At the journal, we’re very supportive of that. We have a whole article type that’s devoted to sort of shared benchmarks and datasets. But it’s another thing altogether to kind of support a service [00:48:00] that is available to everyone and that becomes kind of a shared experience for tens or, hundreds, uh, whatever the number is, millions of people to use and talk about together.

[00:48:11] But we have precedent. We have precedent. What is the best? There's PubMed. So, PubMed. So, this would be a sort of NCBI or kind of, equivalent entity to, to try to organize something, something like this. It can even start out academic. So, you know, GenBank, uh, there is GEO, Gene Expression Omnibus that started out as micro databases at a few academic institutions, which then sort of got lumped in and got sucked into NIH's national infrastructure.

[00:48:39] So there's precedent. But I think what needs to happen is that 20 years ago, there was this national conversation around Centers for Biomedical Computation. Eight or nine of them got funded. And, you know, there was one that Zak had at Harvard, and we had a couple here on Stanford campus. Where is that national conversation [00:49:00] about language models for research?

[00:49:03] Like, our previous NLM director, Patty Brennan, I was sort of joking with her that, uh, you know, we should, NLM should just rename itself to Large Library of Medicine as llm.nih.gov, right? I mean, yeah. But we need that conversation. Yeah, agreed. Alright, Nigam, our last question for you, and I think it goes pretty well with open-source models, is the following.

[00:49:30] How will the democratization of data empower patients? I think patients will start expecting customer service. One of the things I tell med students and our clinicians is that as a profession, we have to understand customer service. We got away with it for a very long time. Ha, ha, ha. Not having it.

[00:49:57] But as people are empowered, they will ask questions. [00:50:00] And questions that cannot be blown off and should not be blown off. And I'm one of the firm believers that information net net in the hands of the person affected is a good thing. Yes, it'll be confusing in the beginning and there'll be efforts needed to educate and so on.

[00:50:19] Knowledge and information is amazingly empowering. So, we will have empowered patients as soon as data liquidity happens. See, that's another Nigam phrase there, data liquidity. That's, uh, like such a good way to, to think about it. Do you think, so one of the things that also occurs to me is that, you know, bridging a couple of things we've been talking about is, LLMs as an interface to patient data.

[00:50:43] So do you think that LLMs will actually enable patients being able to sometimes talk to their data in ways that they haven't been able to before? Cause if someone dumps like a spreadsheet of data on your desk, like maybe that's not the most useful, but if you can say like, what was my blood pressure average over the last three years and then [00:51:00] actually get the answer back, is that a key enabling thing here?

[00:51:03] I believe so. There was actually a demo one of the Stanford students did. I forgot the exact name of that demo. I think it was HealthGPT or something like that. And the idea being that whatever you can get in your, in this case, just Apple's health app, you can talk to. Now, the model is there, there's an app, and then there's a data pipe.

[00:51:24] Now, 21st Century Cures Act will make the data pipe a lot bigger. The model will still be able to handle it. And so, you put the two together, we're not that far away from the first proof of concept that you can download 10 years of data, maybe including an image, and have a conversation with it. Does that scare hospitals?

[00:51:43] Because you can think about spotting medical errors and like litigation and things like that. There was this survey this morning about, you know, asking hospital leaders, CIOs, and IT folks, like how do they feel about their comfort level in complying or [00:52:00] meeting the demands of 21st Century Cures Act?

[00:52:02] And only 36% feel they're prepared. And rightfully so, like, the guts of health IT is a great way to time travel right back into, like, I don't know, late nineties. Raj, any follow up questions? I think maybe I have just one more, which is along the lines of, I think you just said, you're a proponent of information in the patient's hands, which I think Andy and I are probably both pretty, pretty well aligned with you there.

[00:52:32] Patients today, and physicians as well, but patients today, let's focus on them, are using large language models like ChatGPT to interact with their health data. This is already happening, right? Maybe you could just give us some parting thoughts on what you think about that? And maybe what potential good and not good uses of these models are for patients [00:53:00] interacting with data about their own health.

[00:53:03] I think an immediate use is explaining what happened. Like at a given visit, you know, after the visit is done. Like, I have a medical degree and I had a health scare in 2021 and after the visit happens and you spend 20 minutes, you come back and then you have like five questions. Now, given that I'm on the faculty, I could get those five questions answered by sending emails.

[00:53:25] But it would have been nice if that was available to everybody. And most of those questions did not really need the surgeon. They could have been answered by the surgery textbook. I mean, the guy was probably just indulging me as a colleague to respond. So, I think contextualizing what happened and then next level up is like the basic, basic, simple stuff, statin dosing, hypertension drugs.

[00:53:52] So when you say statin dosing, what do you mean? What specifically? This happens, what should I do? Should I go up? Should I go down? [00:54:00] The common pediatric ailments. I mean, you know, our children are now older, but if you have a young kid, I mean, how many times is it that you kind of panic and then it turns out to be nothing and happens to millions of people all the time, right?

[00:54:13] All the time. So, what happened after that comes assurance that you're able to say, it's okay. This is not an emergency. And so, the capabilities needed there are not the clinician replacing diagnostic capability. But only being able to detect when is it not a problem. But we don't build products like that.

[00:54:36] That's not how we think about deploying technology. Computer scientists and students end up horse racing a clinician as opposed to being the gatekeeper saying, I'll only let in things that are really worth your attention. It's a mindset issue about what we should be building. I totally agree with that.

[00:54:55] And I think that's a great place to end. So, Nigam, thanks so much for joining us today on *AI* [00:55:00] *Grand Rounds*. It's been great talking with you. Well, thanks for having me. This was a pleasure. Thanks, Nigam. Thanks. Alright.