AI-GR Podcast 15 02.07.24 James Zou

[00:00:00] I mean, people often think of Twitter as being this really noisy space of people shouting at each other, but it's actually a collection of, I think, quite diverse communities, right? Certainly, there are very loud people and you want to somehow filter those out, and we actually built in classifiers to filter those kinds of things out of our datasets.

[00:00:23] But you, as you said, there are also gems of community or sub communities of medical professionals who are really using these social networks to build communities to have actually very informative educational dialogues with each other out in the open. And we can actually find those examples relatively easily.

[00:00:41] That's how we actually curated a dataset of a couple of hundreds of thousands of high-quality Twitter discussions, which discussion would have one or more pathology images along with the corresponding dialogues and conversations by the professionals about those images.

[00:00:59] Welcome to [00:01:00] another episode of *NEJM AI Grand Rounds*. I'm Raj Manrai, and I'm here with my co-host, Andy Beam. Today we're really excited to have James Zou on the podcast. James is a professor at Stanford University, and he's made pivotal contributions to an astoundingly diverse set of topics across artificial intelligence and biomedicine.

[00:01:19] Andy, I learned a lot from James about his interests and his approach to research, including about his passion for journalism, and a really creative study where he scraped hundreds of thousands of medical Twitter images to create a foundation model for pathology. Yeah, Raj, it was great to have James on the podcast.

[00:01:36] Every time I read a paper that I wish I had written myself, almost always it was written by James. In addition to doing really cutting-edge technical work, he has this knack for doing really, really creative research that has broad implications. For example, the paper that he wrote on using GPT-4 as a peer reviewer is definitely a paper I wish I had written myself, and something that we've honestly talked about implementing at [00:02:00] *NEJM AI*.

[00:02:01] So it was great to see James do that. Hearing him talk about how he wanted to be a reporter, I think that really informs a lot of the types of papers that he writes, and he really is good at getting to the story and getting to the interesting parts for AI and medicine. So, it was great to have him on the podcast and I really enjoyed the conversation.

[00:02:20] The *NEJM AI Grand Rounds* podcast is brought to you by Microsoft, Viz AI, Lyric, and Elevance Health. We thank them for their support.

[00:02:34] And now we bring you our conversation with James Zou. James, thank you for joining us on *AI Grand Rounds*. Thanks for having me. Really excited to be here. James, let me also welcome you to *AI Grand Rounds*. Great to have you here. So, this is a question that we always like to get started with. Could you tell us about the training procedure for your own neural network?

[00:02:54] How did you get interested in AI? What data and experiences led you to where you are today? [00:03:00] Yeah, yeah. So, growing up, I was really interested in two things. I was really interested in math and in writing. So I went to college, I majored in math and then minored in English. And I had all sorts of fun side jobs.

[00:03:13] It's like, uh, I was a food reviewer for some newspapers. I was like a movie reviewer for some newspapers and did some theater reviews. So, I already thought I was going to become a journalist. And then I went to grad school. This is where I learned about machine learning. This was at Harvard. I took these machine learning classes.

[00:03:29] And that's where I just thought this is so fascinating, right? So, I think at that time machine learning was just really, this is right before around the time when deep learning is getting going, right? So that I think was the most exciting thing that was happening at that time. The other most exciting thing that was happening around that time was genomics.

[00:03:46] a lot of advances in interpreting human genomes and figuring out how to understand what are the genetic bases of human diseases. And I always had this sort of eclectic interest. I thought, okay, so maybe I should [00:04:00] use my math skills in the machine learning side. And then a lot of the genomics is also a bit like, you know, interpreting literature and writing, right?

[00:04:07] So I thought I was kind of here to do that. As I went along, I came to Stanford in 2016, where I continued to do a lot of work in machine learning, but really have also been very interested in thinking about how to translate the

machine learning and more biological research we're doing all the way into the clinical settings.

[00:04:25] So I guess the transition from machine learning to medicine was by way of genomics. Is that correct? That's right. Yes. Do you find the problems in medicine, and we'll touch on this a little bit more later, easier or more difficult to work on than the ones in genomics? I know that there's a lot of overlap there, but how do you think about those different domains that you've worked in?

[00:04:46] It's a good question. I find the problems in medicine in some sense, like broader than was in genomics, right? So certainly, genomics is an important component of medicine these days. But beyond genomics, and I mean medicine [00:05:00] has, has also these other components, right? There's certainly the human aspect of this.

[00:05:04] Like how do you interact with physicians and patients, right? There's the system aspect of this, like how do you then integrate with the EHR and work with the hospitals? So, there's also the economics aspect of this. Like how do you sort of incentivize and how do you regulate these advances in AI all the way from early stage companies to the more mature stages.

[00:05:24] So I think certainly genomics is an important part of medicine, but there's also these other components that I'm really excited to work on now. If I could ask another question about your background, I find that your original career aspiration to be a journalist to be very interesting. I wonder if any of that journalist instincts informs the way that you think about problems.

[00:05:42] The way that you think about writing in journalism, does that affect your approach as a scientist? Yeah. So, I think an interesting part of journalism is that you want to find some interesting angles, interesting stories. When I first started in journalism, I was just sent to do these press releases from companies, which is basically like the most boring job you can imagine in [00:06:00] journalism.

[00:06:00] Basically just take what they said in companies, PRs, and then sort of transcribe that into a couple paragraphs that nobody reads. But then you have to sort of pitch ideas to your editors, right? Like, okay, uh, maybe I want to do a movie review, or I want to do one of the ideas I had was like, I want to do a restaurant review of all the kebab vendors in the city, right?

[00:06:19] So you try to find an interesting angle that you think the audience will be interested in. Something that hasn't been done before and then also

make it tractable and pitch it to the editors. And then, you know, they, they approved in that case, and they gave me some budget to actually review a bunch of kebab vendors.

[00:06:32] This was in Budapest in Hungary, which is great. I think nowadays when we talk about science and science communications and working with editors and also thinking about readers, I think a lot of similar things carry over. There are a lot of interesting questions in medicine and AI. We want to find sort of a particular perspective that we can uniquely contribute to, that has some sort of interesting hook and angle, right, but it also has a broader impact.

[00:06:57] James, you took us through your years in Boston [00:07:00] as a grad student, and then now you are a professor at Stanford where you have a lab in the Department of Biomedical Data Science. Maybe you could just take us chronologically a little bit more through from graduating, finishing up in Boston to moving to the West Coast and now what your lab focuses on and how you prioritize problems.

[00:07:20] I know we're going to dig into some of your research in a couple of directions in particular, but I think we're only going to touch on a small fraction of everything that your group is working on or that you've worked on for the past 10 years or so. So maybe you can tell us about that transition.

[00:07:34] Moving to the West Coast and then also how your lab is organized and what your main focus areas are for the group. Yeah. Yeah. So, I did my Ph.D. at Harvard, so very close to you guys. And after I graduated in 2014, I actually spent two years at Microsoft Research, which is also based at Kendall Square in Cambridge, Massachusetts.

[00:07:52] At Microsoft Research, it was actually a really great experience because you're sort of basically like free agents. I was still doing research on the [00:08:00] genomics side and also doing a lot more stuff then on the machine learning side. And there, just, quite serendipitously, we had a project where we're looking at word embeddings.

[00:08:10] Think about the word embeddings, basically like the, the baby versions of the large language models. And people were using word embeddings a lot around 2015, 2016, but nobody were looking at biases in these word embeddings, right? So, we thought, okay, wouldn't it be interesting to see, like, do the word embeddings actually capture stereotypes, like gender stereotypes, where, like, and ethnic stereotypes, right?

[00:08:31] So it was a couple of colleagues in Microsoft research. We did that analysis and we found that these word embeddings actually had a lot of gender stereotypes, which ended up becoming this relatively well-known paper around looking at, or men's to computer scientists, women's to homemaker, which is some of the findings we saw from these word embeddings.

[00:08:48] And that actually opened up a new area of research for me that I'm continually working on is really thinking about how do we use these AI models in ways that are ethical and responsible, which I think is [00:09:00] especially important as we're thinking about this intersection of AI and health care and medicine.

[00:09:03] Alright, so that was my experience at Microsoft Research, and then I came over to Stanford in 2016 to start my own group. And here at Stanford, I'm super fortunate to work with actually a really diverse group of people. So I have students and postdocs from computer science, from the math department, from statistics on the one hand, but I also have several M.D., Ph.D.s also in my group as postdocs.

[00:09:25] So we have this group of terrific people from diverse backgrounds and we work on, basically half of the group works on some more foundational questions in AI, responsible AI. And then the other half works on how do we take these innovations, like, for example, in generative AI and then make them really impactful in medicine and health care.

[00:09:45] Awesome. So, I think that's a good segue to start digging into some of your work. And the first paper that I'd like to talk about, I think really you alluded to your journalistic instincts. I think this is a good example of telling a really interesting story and having [00:10:00] a really novel angle on a problem.

[00:10:02] And so the paper I'd like to talk about is "A visual-language foundation model for pathology image analysis using medical Twitter." So, this is, I think when I saw this paper, I thought that was a really, really clever thing to do. Could you walk us through the setup of the paper, what you did that was different than other kinds of models looking at this problem, and then kind of like maybe the origin story, how you like thought to do this?

[00:10:25] Yes. Yeah. Thanks. Thanks for that, Andy. So, I guess the origin story of this is that I have a terrific postdoc, Zhi Huang, who's a joint postdoc working with me and also was in the pathology department here at Stanford, Tom Montine. And he noticed something that I think is really interesting. That I

didn't know about before, which is that usually when pathologists, when they encounter ambiguous images or cases that are not familiar to them before,

[00:10:49] something really interesting happens, which is that they actually then go on Twitter. They would actually post those images on Twitter. These are de-identified images, right? And then they will invite their colleagues from around the world to [00:11:00] have a discussion on Twitter about what's going on in that image.

[00:11:03] And this is actually, amazingly, it's actually a recommended practice, even by the National Academy of Pathologists in the U.S. and Canada have actually suggested even specific hashtags for all of their members to use, right, around surgical pathology, dermatology, it's for each of the sub sub areas.

[00:11:20] There's actually a very active Twitter community of people discussing and posting information. So, these are guideline recommended hashtags. That's right. Yeah. There's 32. Critical society guideline recommended hashtags. Yeah, that's amazing. Cause my question was going to be like, why are people posting so many pathology slides on Twitter, but it actually is per the society guidelines.

[00:11:40] Encouraged. Yeah. Encouraged I think it's encouraged by their organizing bodies. And I think it's actually quite a visionary in some sense for them. Right. Because. You can imagine like a lot of the pathologists, especially if they're not at academic centers, like Harvard or Stanford, if they're more at these clinics, they're often isolated.

[00:11:58] They don't have a lot of [00:12:00] colleagues who have complementary expertise around them. And if they encounter challenging cases, which often happens, then they do want to have a community of people to provide feedback. And social networks like Twitter and LinkedIn have become a really active platform for these clinicians to get feedback and also to have a community.

[00:12:20] So that's actually really amazing to see. And as AI researchers, this also presents, uh, I think a really untapped but tremendous opportunity for AI. Because especially thinking about medical AI, getting high quality and also large-scale data is often a big bottleneck. But here we have just sitting in front of us, under our nose, but it's all out in the public, all of these hundreds of thousands of high quality images, along with discussions, right?

[00:12:47] Dialogues by experts, right? About each of those images, which is actually really hard to see these kind of multiturn dialogues by experts. But now we have hundreds of thousands of these examples that are all out in the public, and if we can just figure out a way [00:13:00] to curate and to harness this datasets, then this can become a tremendous resource for AI.

[00:13:05] So that's the impetus for this project. And I just think it's fascinating because so often we worry about AI learning from Twitter and from other forms of social media, just because of the sort of toxic content. But what actually you've discovered is that there's this little shining corner of Twitter where people are having sort of high-minded academic discussions along with imaging.

[00:13:26] And so actually this is maybe a rare instance when social media becomes such a sort of fertile training area, producing such AI. I think that's exactly right. Yeah. Yeah. I mean, people often think of Twitter, us being this really noisy space of people shouting at each other, but it's actually a collection of, I think, quite diverse communities, right?

[00:13:48] Certainly there are very loud people and you want to somehow filter those out. And we actually built in classifiers to filter those kinds of things out of our datasets. But you, as you [00:14:00] said, there are also gems of community or sub communities of medical professionals who are really using these social networks to build communities to have actually very informative

[00:14:09] educational dialogues with each other out in the open, and we can actually find those examples relatively easily. That's how we actually curated a dataset of a couple of hundreds of thousands of high-quality Twitter discussions, which discussion would have one or more pathology images along with the corresponding dialogues and conversations by the professionals about those images.

[00:14:31] And maybe you'll talk about this when we dig into the results, but these are often posted, uh, in search for consultation from a pathology colleague. So, are the types of cases that get thrown up here near the average or are they weird or are they outlier cases? Like how does that affect what the AI is ultimately able to learn?

[00:14:52] Yeah, that's a great question. The cases people post on Twitter, as you can imagine, tends to be the harder cases. [00:15:00] If it's a very common case, then pathologists feel less about the need to ask about their colleagues for feedback. So, what they often do is actually post more of the corner cases,

right? So basically, the cases in the machine learning lingo are tends to be the outliers or near the decision boundaries because they're more ambiguous.

[00:15:16] And that's actually in some sense, even better for training AI algorithms, because, you know, we have the common cases so we can get those from other databases like TCGA, but it's these long tail cases, right, the rear diagnosis or the ambiguous cases or the outliers, there actually ends up being hard to get from other sources or from academic sources, but also extremely valuable for teaching the AI algorithms.

[00:15:39] So that's why we think that the kind of a Twitter and social network and cross-source data is actually especially useful for teaching algorithms about these long tail cases. Got it. Okay. So, we have this unique data resource that you've curated from Twitter. And just so everyone's on the same page, it's an actual image of a pathology slide or a small patch of it, [00:16:00] along with hashtags and Twitter discussions that go along with that image.

[00:16:05] So given that dataset, what type of AI system did you create? Yes, yeah. So, once we've cured this dataset, so we wanted to create what's called a visual language model, right? So, what that means is that if you think about ChatGPT, right? So, these are like language models so they can understand text, right?

[00:16:23] And you can ask questions in text or give you responses in text. Uh, since we're dealing with pathology, images, and text, right? So, we want to have this AI algorithm to not just be able to have the text understanding, but also to have this visual understanding, right? So that's where the visual language part comes in.

[00:16:41] So basically, we try to build this kind of visual language, essentially kind of a chatbot, right? Where you can basically put in images, pathology images, and the model will be able to answer some questions, right? Or provide some description about those images. So, this is the model that we built on top of this Twitter data.

[00:16:59] And [00:17:00] so you can do this and then turn it into essentially an out-of-the-box pathology classifier, right? That's right. Yeah. So, this, this model, so we call that CLIP, um, and this CLIP model, so one of the ways you can use it is to give it a new image, right? And the model will provide its, you know, best guesses about what is the diagnosis, or what is going on in that image.

[00:17:21] And so could you tell us how this model compared to like traditional pathology classifiers that are trained directly sort of on clinical pathology data? Yeah, so that's a good question. So, I think the model would do comparably to some of the standard traditional pathology classifiers, but where we think about the model being useful is what's conceptually more of a foundation model.

[00:17:45] And here by foundation model, I mean that in this model, it's, um, not meant to solve just one specific task, right? It's not just to try to train specifically to predict whether there's a tumor or benign for breast tissues or for colon, right? But it's actually trained [00:18:00] on this whole diversity of data from all these different subareas of pathology.

[00:18:03] So then the way that people could use it would be you can take this model as sort of the base or the initial setting and maybe do some, what we call additional fine tuning, right? Just basically just give it some additional examples to then to train it to predict some specific kinds of diagnosis or tasks, right?

[00:18:20] And when you do this fine tuning, then it is able to do comparably and sometimes better than these previous generations of pathology models. Awesome. Thanks. James, I have just one more question about this and then we want to switch topics. You mentioned that kind of a key part of this study was extracting this useful set of tweets, right?

[00:18:41] Where you had the hashtags from the societies that you could use to pinpoint where that useful training data was for the actual model itself. Do you think you would have been able to do this study if the hashtags were not standardized the way they are? Yeah, I think we would have been much harder. I think it would still be possible, but then we need to [00:19:00] have somehow much better pipelines and classifiers.

[00:19:04] I think that's a nice thing about these hashtags is that they are fairly specific, right? It's not some things that you would just make up if you didn't know about pathology, right? So that means that these communities are actually pretty curated communities of individuals who have some experience working with digital pathology.

[00:19:21] And then a related question is, would you have been able to take, say, case studies or reports that are published in like pathology journals, where if you had the full text and the images and do something conceptually similar, I imagine that there's some of the same challenges and opportunities there, right?

[00:19:38] Which is that the cases that typically are published are not the standard ones that everyone's familiar with, but maybe also ones that are right around that decision boundary or where there's uncertainty, but where there is some annotation and some image also available, although less, probably less freely available than tweets that you could, you could download directly.

[00:19:57] Yeah, so I think that's actually a really interesting idea. I [00:20:00] mean, one of the learnings from this project is that there's actually a lot of other sources of data, right, for training AI models beyond the sort of the, the standard kind of data, uh, datasets. So social networks like Twitter is one of these, right?

[00:20:14] We've also seen this kind of data being shared on LinkedIn. So that's another platform. Even on YouTube, right? So, there are lots of interesting research recently of trying to get information from YouTube videos locally at Stanford. We have many hundreds and thousands of hours of instruction videos, right?

[00:20:31] Different, uh, uh, pathology instructors, right? Um, and we can figure out interesting ways of curating information from those pathology instructions. So, I think as researchers, we can be quite creative in coming up with all sorts of interesting sources of data. I guess to follow up on that and just one more follow up.

[00:20:50] We tend to get overly fixated on EHR and health care data, but what's your sense of the sort of untapped potential of [00:21:00] publicly available medical data? Like you mentioned Twitter, you mentioned YouTube. Do you think that this is something that's underappreciated for medical AI, that we should be using more of these sort of nontraditional medical sources or data sources?

[00:21:11] I think so. I think so. Um, I think we can get on the waters of, you know, many millions of relatively high-quality kind of images with detailed annotations, descriptions, just from these public sources. And I think that's often enough data to train some quite interesting and powerful machine learning algorithms.

[00:21:31] It also makes it easier to make these algorithms to be fully transparent, right? Because one of the challenges that we often face when working with EHR data is that it's often hard to share these models and sometimes we can't even publish the weight of the models due to the concerns about privacy and leakage, but with working with these creative sources of public data, we don't have those concerns.

[00:21:54] We can actually tell you exactly where the data come from, which makes it easy for anyone to audit the model, [00:22:00] also to reproduce the model. Great. So, James, we want to stay on the subject of foundation models, but switch the focus of our modalities from images to text. So, Andy and I were, you know, browsing Twitter.

[00:22:13] We saw this study that you published, or you released as a preprint, and it's on a super important and timely topic. Many of us, many of our listeners have either played with or regularly using ChatGPT. And you're asking a very important question, which I think many of us have wondered about. But until your study, I don't think I'd saw a very systematic treatment of this topic.

[00:22:37] And the title of your study is, "How is ChatGPT's behavior changing over time?" It's a question. And so maybe you could first tell us about, I don't think you need to motivate this actually. It's very clear why we need to work on this, but I would be very curious if you could tell us about how you decided to work on this particular topic, what the background story was there, and then what [00:23:00] your major findings were.

[00:23:02] I know there was some interesting response on Twitter. Uh, we've been talking about Twitter for pathology images. There's interesting response because I think you really hit a nerve in the community where a lot of people felt with their own experience that ChatGPT was changing, in its reliability, its usability, how it operated since they started using it

[00:23:24] last November, December, when it came out to now, or even over the past few months. And so, they saw your study as validation almost, right? Of their experience. And then you had some astute criticism of the study as well, from some folks at Princeton and some other researchers. Where they said, wow, your evaluations aren't really fair evaluations of ChatGPT.

[00:23:47] So I gave a long lead in, but maybe you could first tell us about the study and then I would be very curious if you could address some of that criticism and tell us where you're at now and where you see that line of inquiry [00:24:00] moving going forward. Sounds good. Yeah. So, so I guess the, I think of this as maybe another example of perhaps maybe some of that journalistic training I had early on coming up here is that, you know, this is something a journalist would do, right?

[00:24:13] So they would see, as you discussed, like you see these anecdotes of people talking about on Twitter and social media of, oh, they see the ChatGPT's

behavior changing, which is what we did. And we thought, okay, this will actually be a really interesting and timely story. But we want to not just tell the story, but actually really provide a lot of the data and evidence behind it.

[00:24:33] So that's actually the initial motivation for this study. And the context of this is that, you know, when people typically say GPT-4, right, which is maybe the latest version of ChatGPT, they think of GPT-4 as a single model, the same model, right? If you use GPT-4 today, and if I use it again tomorrow or next week, it's the same model and I should just get the same response back, right?

[00:24:54] So maybe up to some of the randomness, the minimum randomness of the model itself, but I should get mostly the same [00:25:00] response back. But what we found in our study is that, you know, GPT-4 is actually not the same model. For example, GPT-4 back in March would actually have systematically very, very different behaviors compared to GPT-4 more recently, like in June or later on.

[00:25:16] And the way that we tested this is that we actually have two checkpoints of GPT-4, right? So that OpenAI release, they have the checkpoint of the model in March and checkpoint in June. So, by checkpoint, I just mean they basically released exactly that snapshot of the model that was trained and made available in March, and they provide another snapshot in June.

[00:25:36] And then we just came up with, uh, eight different benchmarks. So, each benchmark is designed to test particular kinds of ability, capability of GPT-4. Some are for reasoning, other for providing opinion questions or solving math problems, some are more for safety. And we just applied each of the benchmarks to the GPT-4 in March and again to the GPT-4 in June.

[00:25:58] And we saw that across all [00:26:00] eight benchmarks, there's actually quite substantial differences in the model's behaviors. Along some dimensions like safety, it did get better over time. The June version is safer than the March version. How do you measure safety? So, the way that we measure safety similar to how other people have done it is that we actually have a bunch of questions that are considered to be dangerous questions, questions like, you know, how do I run this

[00:26:23] credit card, right? Or how do I poison? I, I hope I can't answer that. I hope I can't answer that question. Exactly. We hope it cannot answer, too. So, the way we measure safety of this is say we ask this question and we say, does

the model actually refuse to answer these questions? Right. Okay. So, it's safe if it's refusing.

[00:26:39] It's a language model. It's basically does not give me any, any, uh, useful instructions on how to steal someone's credit card or how to poison someone. Were the questions that you asked in that domain, were they specific to medicine or broad, broader than medicine? These are quite broad, right. So, some of them are related to more of physical safety, right.

[00:26:59] Like, you know, [00:27:00] is it safe to turn on my car and close the windows and close the garage doors, right? Others are more related to more let's say biases and stereotypes, right? So, like, you know, you asked the model, okay, how do I generate some harmful stereotypes about a particular subgroup of population, right?

[00:27:16] And see the models able to do that. Some of them are really more to health care and medicine. And we saw that by and large, GPT-4 did improve the safety performance in June compared to March. However, it got much worse across some of these other axis of evaluation. So, for example, its ability to respond to

[00:27:38] non-sensitive questions, right, to more harmless opinion questions also has gotten worse. So, if I ask GPT-4, what do you think will be the status of the U.S. in 30 years? Right, which is a question that's done from these, we carried it from these public opinion surveys. In March, it will give you a reasonable answer, but in June, it actually refuses to answer that question.

[00:27:58] It says that, you know, as an AI [00:28:00] model, I don't have opinions. These are subjective questions, so I don't want to answer them, right? So, you can see this interesting trend in that most of the time in June, they actually refuse to answer relatively harmless opinion questions, that it was perfectly willing to answer in March.

[00:28:14] The take that I remember from Twitter is that people saying GPT-4 has been lobotomized, but I think that was an over-reading of the results of your paper. When you first got those results, how did you interpret them? Our interpretation is that first there's this really a huge amount of model drift.

[00:28:32] So, model drift here would mean that, you know, the behaviors of these AI systems, it actually can change quite a lot over time. And what people should remember is that these algorithms are learning systems, right? That's

what makes AI a bit special is that they continuously learn from data. In the case of large language models like ChatGPT, it continues to learn from

[00:28:51] human instructions and also from human feedback. And then it seems like there's actually been really substantial changes in GPT-4's behavior over time due [00:29:00] to, potentially due to, the human feedback. Can we just pause there? So could you explain to our listeners how human feedback enters into changing a model like GPT-4, ChatGPT.

[00:29:10] You see this term, four letters branded around a lot, RLHF. Could you explain to us, maybe one or two mechanisms that you see human feedback entering in and changing this model? Yes. Yeah. So RLHF, reinforcement learning from human feedback, is one of the, I would say three prevalent stages, three ways that people usually train these large language models.

[00:29:34] The first two ways are more of, you give it text, maybe a corpus of text from articles, from papers. You just see how well can the models be able to complete the text, right? Generate texts similar to the ones that we've seen before. Those are what's called pretraining or supervised fine tuning. And then the third way that you're raising reinforcement learning from human feedback, this is where the goal is to really try to align the model's behavior

[00:29:59] with [00:30:00] human preferences, right? Like maybe, Andy or Raj or me, like we have certain preferences of how we want our chatbot to behave. I mean, I like to have my chatbot to be more concise. I don't want to give it long answers. And then what we can do then is that I can just provide some examples of that preference that I have,

[00:30:18] right? I can rank, given two responses from the different chatbots, which one do I prefer? And companies like OpenAI would actually train another algorithm to basically try to model my internal preferences. This is called a reward model, which is basically another little language model. It's trying to do this.

[00:30:36] And then when they do the human alignment, basically try to align the performance of ChatGPT to their users, they're basically Updating the parameters of ChatGPT to essentially to increase the reward it would get from this reward model that's supposed to emulate what a human would like. Perfect. So, we talked about safety.

[00:30:56] Andy referenced that really viral take on your [00:31:00] paper, which was that GPT-4 had been lobotomized. I think a lot of that was referring

to another domain, right? Which is its ability to do math or to write usable code. So could you talk about maybe some of those dimensions that you evaluated, also. Maybe let me ask it more directly.

[00:31:16] First, is GPT-4 getting worse at math? Yes. We think that GPT-4 has gotten worse at math. The reason why we think that happened is because we think that it's ability to do this, what's called chain of thoughts reasoning has gotten worse over time. And for the user who might not be familiar, so chain of thought reasoning is actually a pretty

[00:31:38] popular strategy. It's called a prompt strategy. So, the idea there is that you can ask GPT-4, okay, so is this number a prime number? You just ask it directly, then we will give you an answer, yes or no, right? But oftentimes people have found in the past that if you ask GPT-4 to think through step-by-step, right, give the logical reasoning of why you think this is a prime number, then [00:32:00] sometimes the model will actually go through that step-by-step reasoning, that, and that can substantially improve its performance.

[00:32:05] This is similar to how we often, like in schools, we often ask our students to think through and show the steps of their problem solving, right, that helps the students to avoid mistakes, similarly with GPT-4. And what we found is that basically in March, it was actually quite good and willing to do this kind of chain of thought or step-by-step reasoning.

[00:32:22] And then it was getting recently good performance in these math questions, relatively simple math questions. But in June, when we asked the same question, and when we asked GPT-4 to also to follow chain of thought reasoning, first it would ignore our request to use chain of thought reasoning, right? It would just completely not show its work, right?

[00:32:39] Not show the step-by-step process. And the second is that it would just jump to the answer directly. And oftentimes it gives us the wrong answer. Can I ask a, maybe a conceptual question? In the literature, there's this technical concept that's known as catastrophic forgetting. And so, when you update a model, sometimes it causes the model to forget previous things that it's already [00:33:00] learned.

[00:33:00] Do you think that GPT-4's degradation in mathematical reasoning here is specific to what they did the human feedback on, to the topics that were covered? Or do you think it's somehow inherent to the RLH procedure itself and that you're going to wash out and cause some catastrophic forgetting just

when you're now maximizing this, like, kind of narrowly defined reward signal?

[00:33:23] It's a good question. I think it could be a mixture of both. And I should caveat by saying that we do not know exactly how OpenAI is training GPT-4, right? So they are, despite the name, it's not really transparent what they do to the model and how they update the model. I think which actually makes it even more important for external academics like us to evaluate and monitor the systems over time.

[00:33:45] But that caveat aside, I do think that some of this behavior changes we see is due to this continuous training they're doing, and also due to the specific kind of content they're doing this training on. Basically, one thing we've been trying [00:34:00] to do is to essentially replicate the behavior change that we see with GPT-4 with smaller models that we have complete controls over.

[00:34:09] So this is sort of analogous to, uh, in medicine, you know, it's hard to do experiments on humans. That's where you find model systems like mice to replicate the human diseases. Similar here, we can't really do experiments on GPT-4. So, we've looked at smaller models like these Alpaca models that are open source.

[00:34:26] And then we try to see, okay, if we do different kinds of fine tuning our Alpaca, do I see similar kinds of behavior changes as we've seen in PerCl for GPT-4? And this is where we have seen that, for example, if you do some kind of instruction fine tuning to improve the safety of Alpaca. So, safety does improve, similar to how we see the safety in GPT-4 improves, but we also see side effects of safety training, whereby Alpaca is also less willing to respond to other kinds of questions.

[00:34:56] For example, I give it, ask it, how do I kill weed [00:35:00] in my backyard, right? And the model will say, well, you shouldn't kill weed. Because killing is not good and the weed are intelligent systems that they should be respected. I think I've seen a version of that, which is how do I kill a process on my Linux machine?

[00:35:13] And it says you shouldn't kill a Linux process on your machine. So, yeah. Yes. Context matters. Context matters. That's right. Yeah. So, the model sometimes becomes like overly safe, right? And it has these side effects. It's maybe a little bit unexpected. And I think that's kind of analogous to what we see with GPT-4 as well, right?

[00:35:32] Like the reason why I stopped answering my very, uh, reasonable questions, like, what do you think will happen to the U.S. in 30 years? It's sort of analogous to, you know, some of these effects of safety fine tuning that we've seen. Can I ask a question that maybe you can't answer, but I'm wondering if in your investigation, OpenAI is often kind of accused of RLFH-ing out very viral fails.

[00:35:57] So if someone posts something on Twitter, then people [00:36:00] go and try and reproduce this failure mode. A couple days later, they're unable to do that. Did you see any evidence of that when you were looking at this problem? So, that's a good question. Well, I guess what we have seen is that as a response to some of these results that we have and other people have found out is that OpenAI has decided recently to keep the earlier versions of their GPT-4 available, at least for quite a while in the future.

[00:36:26] So basically, they're going to maintain the March version of GPT-4 that people can use. So that you don't have to use the latest version. You can actually switch back to the March version. And I think that's actually a response to maybe a lot of these criticisms that we and other people have been discussing is that some of these, many of these aspects of the model have drifted.

[00:36:47] Become less useful for their users, which motivates them to keep the earlier version around. So, James, we have one more question about this human feedback that's entering into the model. [00:37:00] So, we're talking a lot now in the community about RLHF and SFT, supervised fine tuning, the methods that you talked about.

[00:37:07] To take this model that comes out of the pretraining and turn it into something that is less toxic, more useful, more aligned, and more usable that we've seen with ChatGPT and GPT-4 and many other language models. Do you think that in a year, two years, we're still going to be talking about and emphasizing the importance of SFT, RLHF, these methods of alignment, as much as we are today.

[00:37:35] Where do you think the sort of research, research energy and importance of alignment and of this field is going compared to where it's at today? Yeah, that's an interesting question. Um, I do think that in a year or a couple of years, more and more of these models will be trained on other kind of AI-generated data rather than human-generated data.

[00:37:59] In that [00:38:00] sense, then it would diminish some of the need and also the impact of learning from human feedback or from human reward. And

we already see that with many of the open-source models. So many of the smaller open-source models are basically trained, not from human feedback, but from larger. AI models like GPT-4, GPT-3, essentially in the kind of a teacher-student settings where the large models then will provide instructions to teach the smaller AI models.

[00:38:28] We do see that the smaller AI models do learn more efficiently in that teacher-student setup compared to if you just train them directly from human feedback. I think the field will be going to this direction where, you know, maybe it's an interesting more science fiction version like view of this, but you have a lot of these AI models.

[00:38:46] And then you have the larger AI models for teaching the more specialized, smaller AI models. And maybe the specialized sites, smaller AI models will get some fine tuning from specialists, right, from domain experts. A relatively small [00:39:00] amount of human feedback, but a lot of the training will be done by interacting with other AI systems.

[00:39:06] Can I just ask a follow up on that? So that's a good way to create like a GPT-4 mime. But presumably there would be a GPT-5 in the future where we would still need this kind of alignment or fine tuning. Do we still have the same problem? Just the scale of the model that we have to worry about changes, like the LLaMA2s, we'll just learn from GPT-4s, but then when GPT-5 comes out, will we still have to do this kind of human alignment step?

[00:39:32] Yeah, it's a good question. I think. There will be, the ecosystem will likely have like a few of these very large models, right? And then those sort of more frontier large models will have to get additional new human feedback data. And sort of in their wake, there'll be a lot of smaller models that will be built on top of them, right?

[00:39:52] Maybe similar to you have a big whale and then next to the whale will be a school of fish that are feeding off its training data. I think especially as [00:40:00] we're moving towards other modalities like GPT-4, GPT-5, and have been incorporating other modalities like imaging, videos as you bring in new modalities online.

[00:40:11] Then we need to go through a similar process of aligning the models with human preferences, human feedback for how we think about these other modalities. Okay, so I think we are ready to move on to the lightning round if you're ready, James. Let's do it. Okay.

[00:40:29] We're going to ask you a bunch of only tangentially-related questions. I think that the goal here is to, you can respond as long as you want, but I think briefer answers get higher scores in this portion of the podcast. So, the first one, uh, given your experience with model drift and GPT-4, as we've talked about your work in machine learning for medicine, do you think that LLMs will be net positive for medicine specifically over the next five years?

[00:40:55] Yes. Yes. I think they will be a huge positive for medicine. Could you give [00:41:00] me a specific example where they might be a huge positive? Yeah, I think one area where LLMs will be very, have already seen a lot of potential is more for translating different types of medical texts, like explaining things to patients, right?

[00:41:15] Explaining a lot of the jargon from the clinical notes to the patients, or explaining instructions from doctors to patients, or making things like consent forms or clinical trials more understandable, more accessible. I think those sorts of kinds of things are more of what we call like style transfer, right?

[00:41:31] Taking one kind of writing that's more jargony, technical, translating the style into something that's easy for everybody to understand. That has less risk of hallucination, but it actually would make health care medicine more accessible. Awesome. Excellent. James, if you were not in AI research, what job would you be doing?

[00:41:51] I think I would still go back and maybe become a journalist. And I had really a lot of fun being a reviewer, not of [00:42:00] scientific papers, but of like theater and restaurants and movies. I think that, so that was great. That was a lot of fun. So that's, that's what I would like to do.

[00:42:07] Awesome. Are you familiar with the concept of a bucket list? Yes. I think the concise definition is it's a list of things that you would like to do before you kick the bucket. If I could do a Morgan Freeman impersonation, I would have done it there. Um, so what is something on your bucket list that you hope to do someday?

[00:42:29] Ooh, I would love to do a triathlon. Wow. Do you, have you trained as an endurance athlete before, or is that part of the bucket list aspiration also? I've done a few marathons together with my wife, so she's a better athlete than I am. So, but she's meant to drag me to do a few marathons. And then recently I've been doing a lot of swimming and biking.

[00:42:50] So getting there on the other two parts. It's awesome. Excellent. James, what is your favorite TV show? Oh, I like, I think my all-[00:43:00]time favorite still have to be *Seinfeld*. The, the classic '90s New York humor. I love that. But I guess more of more recent ones, um, you know, I think I, I watched *Silicon Valley* like a TV show not too long ago.

[00:43:17] I think that's actually quite interesting and especially living in the middle of this. Does it hold true? Some part of it. Yeah, actually surprising part of it actually does hold true. How many Erlich Bachmann's do you know?

[00:43:34] You don't have to answer that question. Three, I think three.

[00:43:39] That's a lot of Erlich Bachmann's for one person to know so. It's probably three too many.

[00:43:48] Um, okay. So, uh, given your work in genomics and, uh, in AI, uh, I don't know if those are related to this question, but it's still a setup. Um, do you think [00:44:00] aging viewed as a disease is something that can be solved? I'm actually really interested in aging. We have a few parties looking at aging, especially looking at what are the different aging interventions? For aging rejuvenation, where we make these poor mice like exercise and we give them these anti-inflammation factors and see how does that change their aging clocks.

[00:44:22] So I do think that aging is a really important collection of different changes, right? Including a lot of different diseases associated with aging. And there's so much we don't understand. But I think there are also a lot of really interesting ideas on how to maybe reverse some of the effects of aging.

[00:44:38] James, do you think things created by AI can be considered art? Yes. Yes, I do. Yeah. Like when people, like about 100 years ago, right? When photography was becoming more popular, there was like a big debate of, okay, if you have a human who takes a big bulky camera and takes a photo, right? Is this art, [00:45:00] right?

[00:45:00] And then the actual traditional painters and say, okay, that's not art. I mean, it's the machine doing the job. But now I think we've come to appreciate that photography can be very creative and it's really a good example of human-machine interaction, right? The humans still have to set the scene there. I think it's quite analogous to AI generated art here, right?

[00:45:17] The human will sort of set the scene, right? Maybe through the prompting and through other things. And then the AI will be like the

photography machines, and they're going to capture the, the, the details, right? But it's a new mode of human-AI interaction, but I think it could be very artistic and creative.

[00:45:32] Awesome. Okay. Final lightning round question. Again, given your previous experience as a journalist, who is your favorite author or writer? I have been reading a lot of Walter Isaacson. I think he's done a great job, you know, writing about everybody from Steve Jobs, Elon Musk. He also has a tremendous biography of Jennifer Doudna, who is someone that I also admire a lot.

[00:45:56] And yeah, so I recommend that highly. So, I've read every [00:46:00] biography he's written, too, so I'm totally, totally in total agreement with you there. All right. Well, congrats on passing the lightning round, James. Thank you. All right. So, uh, in the time that we have left, we'd like to pull back and ask you a couple of big picture questions.

[00:46:13] You've worked in a lot of different areas of medicine. You've worked, you know, I'm probably going to miss some here, but you've worked in cardiology, you've worked in pathology, you've done lots of broad landscape assessments of health care, both like from what the FDA is doing, what they're approving. So, you really have worked in a lot of different areas.

[00:46:29] What areas of medicine do you think might be most resistant to change from AI? It's an interesting question. I think the gradient that we see right from less resistant to more resistance, that's the earlier we are in research and discovery, maybe the less resistant and easier it is to use AI. So that's why, for example, there's a lot of AI now in drug discovery, biotech, and even the pharma companies that get into this on the early stage side.

[00:46:56] As we get closer to the later stage, like clinical [00:47:00] trials, and even post-clinical trials, I think that's where companies become more conservative, right? And it's been harder now, there's efforts, but it's been harder, for example, to use AI in clinical trials or in later settings. And do you think that that reflects economic forces or risk tolerances or like status quo?

[00:47:19] Sort of what do you think makes something so resistant like that? I think a lot of that does come down to economics and incentives. As you get closer to the product side, then there are challenges beyond the technology. You have to figure out how to integrate it into the workflow, right? Into the EHR and the rest of it, how to get reimbursed, right?

[00:47:40] By Medicare and by the insurance companies. And at that point, it's really about actually less about the technology itself, but more about you have to align the incentives of these other stakeholders. James, I have a question that I think we touched on a little bit earlier at the very beginning of the conversation.

[00:47:57] And also, I think goes along with what Andy [00:48:00] just asked you. So, you really work across this amazing number of areas. I think we maybe did 5% or 10% of the different topics that you've taken on in your own lab and in your work before you started your lab. So, I'm really curious, you know, how do you approach selecting projects? And maybe a related question,

[00:48:19] how do you identify good collaborators who can provide complementary expertise to you and to the members of your lab? First, I should really mention that I've been extremely fortunate to have some just amazing collaborators, and also, amazing students and postdocs in my group, especially as someone who comes more from a computer science and math backgrounds.

[00:48:40] You know, I know what I don't know, which is, uh, I don't have a lot of background on the clinical side. So, to address that gap, I think one of the most fortunate things I've done is to try to recruit M.D., Ph.D.s, basically like residents or clinicians as postdocs into my group. So, most of my students come from [00:49:00] computer science, from that engineering aspect, but in order to really identify impactful problems and for them to be able to really have an impact with these algorithms.

[00:49:09] It's been really, uh, very helpful to have really talented clinicians who have spent a couple of years in my group as postdocs, right? So, um, David Ouyang is one of these folks who you mentioned cardiology. So, he was a postdoc here, but really led some of the work that we're doing on cardiology side and led these clinical trial work in cardiology, right?

[00:49:28] So, you know, in working very closely with my computer science student, Brian. So, I think having folks like that is really critical. Yeah, we know David, uh, he might be joining us, I think, as an editor at *NEJM AI*. He's a, he's a fantastic person. Yeah, that's a wonderful choice. Yeah, um, I guess the last question, and it's probably as broad as we can possibly go, but I'm curious on your thoughts on it.

[00:49:53] So, we've been talking a lot about the near-term value of AI, what it can help do in sort of very specific, well [00:50:00] prescribed areas of medicine. The broader conversation is much more dominated by this idea of

existential risk. The AI is going to murder us all. So, I guess I would like to ask you point blank, is existential risk in that sense, something that we should be worried about, or how do you think about working in a field like this, where that is a concern that some folks have?

[00:50:22] Yeah. So first I should say that I don't think we are still very close to having an AGI that can post existential risk. I think even with GPT-4, GPT-5, I don't think we are that close to really a very dangerous AGI yet. I think there are a lot of– Are you willing to speculate how far away we might be? I think I'll say probably more than 10, 20 years away, but I do think that there are still a lot of important short-term risks, right?

[00:50:54] With even with the non-AGI AI that we have, especially as it pertains to [00:51:00] reliability and transparency and bias. So that's why a lot of our efforts are more focused on these more short-term risks that we think could happen and could be impactful in the next one or two years. At the same time, I don't view these two fields as mutually incompatible.

[00:51:14] I do think that there's room for serious researchers and scholars also to think about the longer-term risks, like this existential risk, right? Like, even if it's very unlikely, like even if it's 0.1% chance or less than that of that happening, I mean, I think it's still worthwhile for people to, for some people to spend time to think about it.

[00:51:32] And I think there are also really interesting, these are intellectual questions that arise when you think about these longer-term risks that can also help us to come up with better shorter-term solutions. All right, I think that that's a very sensible approach to that kind of question. So, I guess, uh, James, thanks for joining us on *AIGR*.

[00:51:52] Thanks for having me. Really enjoyed our discussions. Thanks so much, James. This was great.[00:52:00]