**Ahmed Imtiaz:** 0:00

Synthetic data could be generated algorithmically, like you could have some rendering engine rendering the world and then you're creating synthetic data, like it could be like a 3D object that you design and your synthesizing images. We can have some ratio of real and synthetic data, but the most important thing is we need fresh real data compared to just having some ratio of real data. So some ratio of fresh real data is what can actually help us If we want to compare the amount of real data being generated and the amount of synthetic AI, synthesized data being generated. There's more real data being generated. There's no question about it right now.

**Craig Smith:** 0:35

Hi, I'm Craig Smith and this is Eye on AI. In this episode, I talked to AI researcher Ahmed Imtiaz about the phenomenon of model autophagy disorder, also known as MAD, generative AI Models. Ahmed explains how consuming their own generated data can cause models to lose diversity and become trapped in artifacts. He discusses experiments on image and text models showing that this effect emerges quickly, though it's not yet clear how prevalent generated AI data is on the wider internet. We talk about the potential strategies to mitigate MAD, including using fresh data and watermarking so generated data can be recognized in training datasets. The conversation provides an insightful look at this emerging challenge. I hope you enjoy the conversation as much as I did. Ai might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested, so buckle up. The problem is that AI needs a lot of speed and processing power, so how do you compete with costs spiraling out of control? It's time to upgrade to the next generation of the cloud Oracle Cloud Infrastructure or OCI. Oci is a single platform for your infrastructure, database, application development and AI needs. Oci has four to eight times the bandwidth of other clouds, and offers one consistent price instead of variable regional pricing. And, of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less, like Uber and Coher, take a free test drive of OCI at oracle com slash ion AI. That's E-Y-E-O-N-A-I I'll run together. That's oracle.com slash ion AI.

**Ahmed Imtiaz:** 2:56

Thanks so much for having me over here, craig. I'm a PhD student at RISE and I'm also a student researcher at Google right now, so I study the approximation theory for neural networks. So what I do is I think of neural networks as piecewise, affine functions and then I use this theory to see if we can explain how generative models work, if we can do neural network interpretability or even say something about why the neural network has different training phases, like why we see such training dynamics that we see. So I come from Bangladesh. I'm originally from Bangladesh. I did my undergrad there and after my undergrad I came here to do my PhD. I also founded right after my undergrad, so this is also something that's really close to my heart. I have a passion project called Bengali AI, which is a non-profit base in Bangladesh. What we do is we create datasets and we open source them for research to accelerate research in Bengali language technology. So LLM and also speech recognition are very primitive in Bangladesh.

**Craig Smith:** 4:14

So that's interesting because, if I'm not mistaken, I've read that Bengali is one of the languages that is underserved in the training data for the big and consequently chat, gpt, and some of these don't do well in Bengali. Yeah absolutely.

**Ahmed Imtiaz:** 4:35

It's not only chat GPT for Bengali. We don't even have good OCRs. It's in development, like OCRs and ASRs. These are also fundamental technologies that's required, especially when you think of Bangladesh. It's like in South Asia. It's so densely populated. So if you have this groundbreaking technology that can enable people to access technology in your own language, like if you have a Bengali chat GPT you get all the benefits of chat GPT, but in your own language, which makes accessibility easy to access. So that's the target that we have making these models more targeted towards the language specific crowd so that we can increase accessibility for the speakers over there.

**Craig Smith:** 5:24

Yeah, one thing I have to ask you is that your screen name is MTAZ. What's the naming convention, because I'm calling you Ahmed.

**Ahmed Imtiaz:** 5:35

Yeah, so I have a big name. If you noticed, it's like Ahmed MTAZ in my own right. The funny thing is, ahmed is a very common name in Bangladesh, so if you look for an. Ahmed in Bangladesh, it's going to be you who me so that's why MTAZ is what I go with. Mtaz is my preferred first name.

**Craig Smith:** 5:54

MTAZ. Okay, great, then I'll start calling you MTAZ. Yeah, so you're now at Rice, is that right? And how did this paper come about? And then we can start talking about it? I'm interested, and I don't know if I said it before, with the paper self-consuming generative models go mad and mad is an acronym. Maybe you can first tell us what the acronym mad stands for, and then we can get into how you came about doing the research and writing the paper.

**Ahmed Imtiaz:** 6:34

So mad is the acronym for model autophagy disorder. So autophagy is a term that refers to consuming. In someone like itself, consuming the self-consumption is like something that is a keyword that has come up in a lot of different literature. I recognize that it was also, like, mentioned in Greek literature and everything. So autophagy is the term that we chose to sort of denote this behavior that we see when generative models, when they consume their own generated data, they start behaving in a non-standard way, or by non-standard I mean like something that we do not want to happen. So the way, this research, so actually we started thinking about this problem like I think, more than a year ago now. We were so like we always traveled to conferences, like this idea of increased accessibility to generative models would obviously increase the prevalence of generated data, ai, synthesized data, online, right, and we that's what the company, that's what we want in general like we want people to be able to access LLMs, we want people to be able to access, with journey or like such technology so that we can, like they can, generate images that they want and that's consequently going to lead to more synthesized data online right. So we've been thinking about this problem and then at a conference, we've been discussing with people and we see that like there's this general consensus in a lot of people working with large LLMs and working with big data that we might be running out of data in the future Because, like, we want to go exponential and to be able to maintain that trend, we need more data, but we don't see that much data being generated or like it's very easy to like use up all the data that we have, right. So maybe, and there's also the implications of privacy. So there's this whole research direction called membership inference attacks where, basically, you could like this, there are these algorithms through which you can find out if some data was in the training set of another model or not. So if you think about that, if there's like, if all the data in the world, even sensitive data, is being used to train models, then through membership inference attacks, you could be like you could possibly find out, okay, so this sample was in the training data. So this is like true, this is like if I have, like maybe, some sensitive document ID, right, and then if I can infer that this ID was inside this model's training data, then we can say, okay, so this should be a valid ID, because it was in training data. That's ground truth, right? So there's this whole idea of like these, like attacks, or there is this whole idea of privacy when we're using this data set, right, these data sets. So there are people who want to make like train models in a private manner, so they might also be using synthetic data, like a synthesized data, instead of using the real data, so that when those models are being attacked like when they send it out in the open and someone attacks that model using membership inference attacks, like it's going to turn out that this is like these are synthetic data that's been trained on, so it's not going to matter much. So there's this: there are people who want to use just synthetic data to train, and there's also synthetic data going out in the open into the internet. That's like filling up the internet with synthetic content as well, and people are running out of data. So there are like these, like a couple of directions where we see this need for synthetic data, so that the natural question arises that what is the effect? So how differently is synthetic data going to play a role compared to real data because it's not real. Synthetic data has been like. So there's like. I want to make a distinction. So synthetic data could be generated algorithmically, like you could have some rendering engine rendering the world and then you're creating synthetic data, like it could be like a 3D object that you design and your synthesizing images. But that's like one direction. We're speaking of AI synthesized data because these AI models are getting like, like they're getting more prevalent. People are being able to use it and generate data. So we're thinking of synthesized data. What are the implications?

**Craig Smith:** 10:46

So that's okay, let me, let me just stop to unpack some of that. First of all, on the synthesize and AI generated Do both have this autofaggie effect? And I ask because I have a friend from the New York Times who said they started a startup AI reverie and Met about it. And he was, they were a synthetic data company, a visual data company. They were, you know, producing, you know, entire cities, rendering entire cities. So does that kind of data have the same effect on a large model as AI generated data, or would the effect be?

**Ahmed Imtiaz:** 11:34

different the effect would be. So, my, in my opinion, the effect would be different Because, like, when you're synthesizing data using, like, a renderer, then you wouldn't necessarily have this autofaggie loop, because you don't generate those images and then use those images to Retrain the renderer. So if there are elements in the renderer that do training, that do learn Based on the images that are being rendered, so if there's a generative model over there somewhere, then that could be affected by this feedback loop. But in general, like, like if we render data using these physics engines, it isn't like that is out of the scope of her discussion right and and the Before we, we go on with that.

**Craig Smith:** 12:20

Can you just describe for listeners what this is, this loop and what happens? And my understanding is that you, you, you lose the long tail and the data over time and you end up with just the mean. But is, yeah, just describe the, the, the effect, yeah, so.

**Ahmed Imtiaz:** 12:42

So the interesting part is that. So this effect Basically says that if I want to like shorten these, like the takeaways, the effects of like a self Feedback loop in any format, like how, like whether it's like just generated data that's being used in the data that being used to train, or whether it's in some ratio with training data, so the effect that we would, that we, from our experience, that we see is that we would have this loss of diversity, right, so we like, like you said, it's going to be converging towards the mean. So we would see more Generic features Compared to more diverse features. So you would have lost the tails, right. And another very interesting thing that we see is that we see this like this emergence of artifacts that are subject to the algorithm that are being used to generate the data. For example, I can see it talk about two different generative models that we studied, so we have. We can talk about like there's stable diffusion that has, like the diffusion model, a method that's different from Just generative adversarial network, that has a different way to train, right, and for both these methods We've seen this auto-faggy effect or these artifacts sort of appear. What we see is that it's subject to the algorithm itself, for example, for style ban or like the generative adversarial network that we've studied, we see that there are these like cross-hatching artifacts that are appearing, that are like that, that could be related to the, the way these models behave or the way these models learn, compared to diffusion models, which sort of have these blurring artifacts that are coming up. So it's very subject to the model, but we do see this auto-faggy loop increasing the artifacts in the images and it's yeah, it's not only just like something that happens like in the infinite limit, it happens like within maybe 10 or 11 steps, like in a short few steps of auto-faggy.

**Craig Smith:** 14:40

Yeah, and, and this applies not only to image generation, but to text generation as well yeah, right.

**Ahmed Imtiaz:** 14:48

So there were some concurrent papers who were studying this effect for LLMs, so they've also seen that for large language models. If you Retrain or fine-tune a large language model with its own data, you would see a similar effect of like converging towards the mean and the and the and the generated data Like slowly. It also loses quality as it moves, losing a semantic meaning as you continue this loop.

**Craig Smith:** 15:15

So it's a general phenomenon. Yeah, and One thing that I've had people have asked me and I don't know, so I'll ask you how much. I mean that the internet is, the volume of data on the internet is, you know, massive. Has anyone been able to measure or estimate how much of that data to date is generative for AI generated and are there any predictions for how that Percentage will grow? I can imagine today it's initial, but are there predictions about how that might grow as a generative AI Spreads through the global economy?

**Ahmed Imtiaz:** 16:04

Yeah. So Two questions right. The first one is what fraction of data is currently synthetic online? It's very hard to say what is synthetic and what is not even like that. We have these methods to do that. Deep fakes and like. A lot of technology is being built To be able to detect synthetic data, but it's not completely like infallible right. So there's I'm not aware of any studies where they are claiming that such a fraction of the data on the internet is generated, but we've seen that already in the data that's being used to train large models like stable diffusion for image generation. We see that in Leon, the large like open source data set that's being used to train these large models Already contains something that is data and it wasn't that hard to find. So if we say that, how? How did you find it? Oh, so there's a website called Know Your Data, so or. And there's also another call: have I been trained? So to these websites. You can explore these data sets that are being used to train to see if your image Was there or not. So you can search on these data sets using queries and and like. It was very easy to like, just like one or two Queries away, we were able to find some synthetic data on those data sets, like AI synthesized like AI synthesized, but what was the query that you would use? We use it like avocado on a chair. That's like the reason, yeah, yeah. Like an avocado chair a chair shape like an avocado stuff like that, like something that people would use to generate.

**Craig Smith:** 17:43

Yeah, and that actually I don't mean to get sidetracked, I do want to get back to the paper. But another question that people ask me and that I can't answer there. There are the model builders, like open AI and thropic and Google and whoever, and they're training on data. And People ask me well, where does the data come from? And I say, well, it's on the internet. But what are they? And then I? But? But there are there's a class of a middleman, data Preparers or data prep companies. They are packaging the data into data sets like lions or Others. I mean, lion is an academic data set, but are there companies that are doing this and how are they getting the data in? For example, I was talking to my wife this morning about this lawsuit against open AI by Sarah Silverman the comedian. That's a copyrighted book. How did open AI get a Digitized verse copy of that book? If it's copyrighted? Did they? Did they buy it? A Kindle version? I mean, how, what? How does it actually happen? Do you, do you understand?

**Ahmed Imtiaz:** 19:07

Yeah, in my opinion, I think it's just how humans work. I said Silverman's book is quite good and I think that's why some humans have decided to put it on the internet, like people do piracy all the time. Right there there are a lot of websites where there are PDFs and I think like those are maybe sources from which it might have come Into the training data sets. So it's like when we're like going out into the internet to just get all the data. It's. It's very hard to say like. Like. The way it generally works is like we get an initial set to then scan through and select right. So that's how data is a profession. That's how we do it at Bengal AI as well. We crowdsource that a, we crowdsource an initial set and then we annotate it or we clean it up, so my assumption is yeah, and just let me take that first step.

**Craig Smith:** 20:01

When you outsource, how is it? Are those just web crawlers scraping the internet? Or are people Downloading all of the Gutenberg Online library or all of the Google books online? Or how? The mechanics of getting that data into a data set? How is that?

**Ahmed Imtiaz:** 20:25

so I would assume it's being crawled, because it's actually Is generally like, given the scale at which we need to train these large language models right, it's actually impossible for us to have, like Mac, like anything manual in that. That's why, like even in the Leon data set, they collected like all the images and then they had like an automated method to score those images and say whether the captions are Matching up with the images or not. So these were all automated. They were like it's very hard to do anything manual there. So if it's like a smaller data set, then we can assume that there could be some manual intervention, but for these like super large data sets, it's mostly like crawling From the internet.

**Craig Smith:** 21:08

Yeah, so open AI, for example, they don't have, I mean Google, it's obvious. They have the data through the search engine. But open AI, would they have hired a data collection company to go out and scrape the web? And then I imagine there's some prep in putting it into a format that then the model can be trained on.

**Ahmed Imtiaz:** 21:39

Yeah, I'm not really sure, Like that's something, it's kind of specific to open AI. I'm not really sure whether I know, like, what they did, but we know from, like the public records that they for the, the RHF data sets. They did outsource it to some other Companies who did collect the like help to collect the RHF data sets. So I don't know if it's like outsourcing, or if it's like part of open AI, because these are big companies, right? Yeah, so, but, but I assume that there needs to be, like for RLHF, for example, there needs to be the human element, because we need the data set to align our models. But for when we just need, like, a large chunk of good quality text, I think the manual intervention would be more towards how we can curate the text in an automated way. So even that's hard, because there there can be so many different cases, right, so there can be the cases that you need to handle, maybe like put in failsafe so that you don't really like scrape the, the insensitive part of the, the Internet, where you have, like all the like everybody saying whatever they want to write like 4chan, maybe you know. So it could be like that there. Those are the parts where there could be manual intervention coming in and maybe, like, like you, would need a big team to be able to do that. It could be that they're using it like other teams as well.

**Craig Smith:** 23:00

Yeah, so there would be like a white list, not a white list, it would be too large a black list. You know that you cannot scrape from these domains. And then you set your web crawler free and it's gathering this data, and then is that data converted into vectors or is it dumped into a giant text file, or how does it? What's the next step?

**Ahmed Imtiaz:** 23:32

Yeah, generally the way it should work is so for when you're when you're collecting, like when you're scraping data from the internet at least, like at my company, what we do is, if we create image data sets, we generally take, take a key track of the URLs that are being scraped, and then I think even for Leon, when they release the data set, they basically release the URLs. So if someone from the internet, like someone from whoever likes a webpage, one image has been scraped, if they remove the image, then the URL is not going to work anymore, so that the data set sort of changes. So that way the data sets are dynamic, depending on, like, the actual owners of the data, and the only thing that the like Leon is releasing are the URLs. So that is like one method in which data could be shared.

**Craig Smith:** 24:23

And so. So, and I'm sorry, I'm saying Leon, you're saying Leon is. Am I mispronouncing? It's L Y O M.

**Ahmed Imtiaz:** 24:32

It's L A.

**Craig Smith:** 24:33

I O M Okay, l A I O M my mistake. Not a worry, yeah. So, Leon, when Leon is, it's a list, it's a set of URLs. How does a company or a lab convert that into training data?

**Ahmed Imtiaz:** 24:58

Well, they would like to download the data like downloading using the URLs that are valid, and then have their own methods to pre-select which sample they want to use. It could be based on keywords or any other filter that they have.

**Craig Smith:** 25:13

And is that data then downloaded in a text format, or or is it immediately vectorized and put into a massive vector database?

**Ahmed Imtiaz:** 25:24

Yeah, these are very case specific. So this would vary between like some labs, like for for some people it would be, it could be like more convenient to have like JPEG images, for example, and then so when we, when we think about such large data sets, there are more, more efficient ways to store this data as well as like like transfer this between like memory or like between like locations, right? So I'm guessing it would be very case specific.

**Craig Smith:** 25:54

Yeah, okay. So back to the paper. The internet has some fraction of its content. Is AI generated? Presumably a very small fraction at this point, but you're already seeing autofaggie or this mad effect. Or is that only in the, in the? The testing that you do with, you know, in a closed system.

**Ahmed Imtiaz:** 26:28

Great question. So we don't see. So it's so. So the autofaggie is very interesting, as in like the effects of autofaggie or the symptoms of autofaggie. They come up late, even after like autofaggie has been happening for a while. So we don't necessarily see symptoms of autofaggie or we don't really like, we haven't even like very rigorously explored as well, like if stable diffusion has any autofaggie effects. It's not, it's kind of non-trivial to see if there are autofaggie effects over there. But we do see that like in in our, in our controlled experiments, when we're where we're using the same algorithms used to train stable diffusion, the same algorithm behind like generative models, when we, when we do those controlled experiments, we do see that autofaggie is happening in like in a few steps. So there are like three different sort of settings which we explore. So one setting is where we're like completely training a model using its own generative data. So we take, say, suppose, some generative model, we create some data and then we take another jet like we, we start from scratch. So we retrain another similar generative model using that generative data and see, like what defects are. So this is like the most extreme case and it would be relevant for, like, like I was saying, like people who want to use just synthetic data for privacy purposes, right, so that is. That is where we would see, like DD, the strictest form of autofaggie effects coming up, like, so we would see very fast, like very fast, decay of the diversity and then you would see these artifacts come up very, like way faster. Another setting would be where, suppose we have like some training data set that we are always going to use, so we always have this fraction of real data, but then we're going to create synthetic data to augment the original training data that I had. That I have. And the reason a lot of people actually have been thinking about this is because, like, if you think of, suppose that, like, we have two different images, like one is my image and another is your image, right, and then you want to, like there are a couple of images between these two faces, like if we think of interpolating between my face and your face, then we would see, like a week there, there can be like a continuous change in the facial features that I have that maybe would slowly lead to your face, right, right, so, so if we have a generalized generative model, then a generalized generative model by definition should be able to do this interpolation. So learn the data manual so it knows what that is, like how to go from my face to your face. So a lot of people what they've been thinking of doing is like using generative models to create these intermediary faces compared to what you have in training data. Right, you have, like, when you generate synthetic data, you would get these intermediate faces compared to what's in the training data. So you, that might help in training a newer generation of models. But these people have been thinking of using synthetic data as well. So over there what we see is that the auto fact effect does exist, but it's a little. It's not as sharp as just training on synthetic data, but it doesn't stop auto fact. If I'm happy, we do see that if we have a fixed training set, if we keep generating more synthetic data it is going to slowly decay towards madness. One small thing that I that I wanted to mention is like the term mad comes from, like it's actually very much inspired from the mad cow disease, because yeah, oh yeah, like, like some person had the smart idea that, okay, let's feed, like cow, cow brains to other cows or something, and then like it. Maybe it was a good idea at first, but then it turned out to have been. The self consuming loop led to some very bad effects.

**Craig Smith:** 30:08

So, yeah, yeah, so that I mean it's also, it's, it's the same principle in, in, in, in, sust, or narrow gene or small gene pools, where the, yeah, the, the. So first, before we talk about how you might deal with us. So you're seeing it already. You were saying these, these hash marks in a style again, the fusion of the star, again generation. Yeah, yeah, yeah yeah, it is. Is there some metric about how much? And I understand it's model specific and algo specific, but you know that once you reach 2% or 5% or 10% or 20% of generative data, you're going to start seeing this effect, or is it impossible to say?

**Ahmed Imtiaz:** 31:22

The percentage of generative data versus, like, real data versus real data. Yeah, that is like a very important thing that we want to explore in the future and we've done some explorations as well. So the most important takeaway there is we can have some ratio of real and like synthetic data, but the most important thing is we need fresh real data compared to just having some ratio of real data. So some ratio of fresh real data is what can actually help us sort of like nullify the effects, like maybe asymptotically we are doing more studies here but asymptotically we could have like a situation where, like this percentage of fresh real data on every loop would basically keep delaying like the Matcow effect, to like asymptotically right, so we will never reach like autofaggie.

**Craig Smith:** 32:15

Yeah, and in terms of the internet, because of course, the popular mind, the layman's mind, goes to wow, you know, all of this generated, ai generated data is going online. Maybe in 100 years, 90% of the data online will be AI generated, and then, you know, this autofaggie will really be a problem. Is that something that you guys have contemplated? Is that, is the growth of generated data so high and, with regard to the generation of real data, that that's possible?

**Ahmed Imtiaz:** 33:09

If we want to compare the amount of real data being generated and the amount of AI, synthesized data being generated, there's obviously more people. Like there's more real data being generated there's no question about it right now but like, the thing is that the direction that we want to go towards is people adopting these technologies to help generate, like, to help write better, maybe, or help with the image that they're trying to create, right, so we want people to be adopting this, so we want all the people in the world to slowly start using these generated models, and that would be beneficial in so many domains. I think that is why, like it's it's we can like we can like it's not like a scenario. That's it. That's sort of scythe fun, right, it's almost here. Like, we have a lot of different websites that are completely generated. Like, there. There, I think there was an article I don't know if it was New York Times or Guardian where they were reporting that there are like these fake news websites which are like, which are completely generating data using LLMs, right, so we already have these sources that are creating these websites, create only creating synthetic data, and we have people who are using, like millions of people using chat GPD to write better, right. So so, slowly, like from this point on like, from chat, the, the, the chat GPD going public for use and then onwards we're going to see a shift in the language that's being uploaded online right In the way things are being written, and we we want it to be adopted more. So we'll see this, this effect, grow as time goes on, right? So that's why like it's, that's why we want to. We need to think about this right now as in like okay, so we, the, the new generation of GPT is maybe GPT, and when we try to train it, maybe after training, we'll see that it's very formal, like it sounds very formal because for the past 10 years, people have been using GPT like technology to write like more formally or in a very nicer way, right. So those effects would start coming up and for us to not like to meet that point where, like, our performance is getting saturated right. We want to start thinking right now like how we can deal with this probable phenomenon that we're going to see, yeah, so.

**Craig Smith:** 35:33

So what are some of the strategies that you might pursue that could deal with this?

**Ahmed Imtiaz:** 35:39

Yeah, like I was saying that there's the idea of having fresh new data, which can be one strategy of mitigating the, the mad phenomenon, right Another way could be like another easy way that we should like to try right now actually is just a watermark. The data that we are generating, like we already we already see, like when Dolly was we use, dolly had that like sort of the, the multicolored watermark block right at the bottom right, and there's also like this like recently Google released Google announced that they'll be using SynthID on all of their generated images. So SynthID is basically they put in some really perceptible watermark into the image that you can use to like see. Like that's something I think is also being used in YouTube and a lot of others like where people are uploading their like content to be able to like a monitor if it's being copied or something like. So SynthID is also going to be used in generated images. So that's like another thing. There's a lot of research going towards watermarks. So, watermarking could be one way to detect these images, but then if we want to use any of these watermarks, so there's this dilemma If we want to use these watermarked images to train generated models, we we have forcefully added this noise into the image and then we're using that to train the model. So that's going to increase the autofaggie effects, because autofaggie basically, like it, exaggerates these hidden artifacts. That's why we see this crosshatching coming up in StyleGun too. So there's this dilemma of okay, watermarking and then okay, so now we can detect synthetic data, but autofaggie might still happen. So then we need autofaggie aware watermarking, meaning that once we detect the watermark, we should be able to remove it as well before using it for training, so that we have, like, the actual information intact. So this can be like an interesting, like, like this would be an interesting dynamic field to sort of do research. I love adversely, like attack research. If you're aware, neural networks, like for some imperceptible change in the input, can behave completely differently and the way it became such a like, the way they feel was explored or the way it flourished was. There are a lot of people working on attack methods and a lot of people working on defense methods. So then there could be like. There can also be like in terms of these autofaggie effects. There could be a lot of people working on watermarks and a lot of people working on moving the watermarks to mitigate the effect of watermarking like autofaggie. So this can be interesting.

**Craig Smith:** 38:19

Although the point of watermarking would be in order to exclude those things from the training day Right.

**Ahmed Imtiaz:** 38:27

Like you might want to keep, like I was saying, like there could be some benefit of having some synthetic data in your, in your training, as long as you have fresh, real data, right. So you might want to keep some synthetic data, maintain that ratio compared to your fresh new data that you're getting and then try to retrain. And that's how we will keep getting better because we want to keep getting better. That's the target Right.

**Craig Smith:** 38:51

That's right and the fresh real. But in any case, watermarking and I I just interviewed the I can't remember who is the CEO or CTO of DigiMark, who's the biggest watermarking company in the world, and then Scott Aronson, who's a computer science out of Texas, who's working on safety for open AI. He's working specifically on watermarking text, which is obviously much more difficult by creating statistical patterns in the generated text that can be recognized by another system. But watermarking would be a way of excluding data, synthetic data, from training sets or at least knowing what ratio or what ratio you have. Without that, there's really no way to know, right?

**Ahmed Imtiaz:** 39:51

Without that like there are, like without watermarks, it's there. There are going to be ways to maybe find out what synthetic like. There are ways that are there right now, but as we keep improving, it's going to keep getting harder. Right, so so that's, that's, that's like one aspect of it, so there could be like other methods to like control this as well, because this is like a new, like this is a new phenomena that we will learn and there are a lot of like research groups who are starting to like jump in and study this so yeah, yeah, yeah, yeah, because Finding synthetic data to train a model on is fairly straightforward.

**Craig Smith:** 40:33

Once the internet is polluted with synthetic data that's not watermarked, It'll be very difficult to find fresh real data right and and it's obviously not practical and and would be too costly to Generate real data just to train a model.

**Ahmed Imtiaz:** 40:56

Yeah, yeah, this is like this is a very good question. I think there might be like something in the maybe companies in the future they decide, okay, we're gonna generate our own data and and like, maybe, like, like, when I, when I think about this situation, I like the first thing that comes to our mind is matrix, so it makes we have, yeah, humans generating something for the machines, right? So there it could be that we have like a situation where we want to generate the company might want to generate their own data, could be for privacy purposes, could be just to get the fresh, real data that we need to be to keep on accelerating, because we do see this exponential growth, but because of such phenomena, it could be that we saturate and then we would need to do interventions to keep growing, keep getting like better models.

**Craig Smith:** 41:43

Yeah, and when you say saturate, this is another thing that Somebody was asking me this morning. Actually. Is there a point at which, so the LMS are right now training primarily on public data on the internet? Is there a point at which all of that data would have been used to train a model or models, and Does that create some sort of a convergence between models, even though they're owned by different companies, if they're all being trained at some point on the same data? Yeah, that's a very good question.

**Ahmed Imtiaz:** 42:26

You have the convergence of models to one, one sort of similarity. Like it's very hard to like, sort of say, whether two different algorithms are going to converge to the same point or not. Like it is intriguing that you have to have a. It is intriguing that using the same data, that could be a case that happened. But we've seen that for a lot of these, like the open source model that we have, that's why I think open sourcing is so important. Because we have these open source large language models and because people like doing studies on these models, we've seen that there are a lot of different tricks you might also have to apply to get the utility out of your network. So just training on the data might not be the be all to get the performance that you desire. So there, I think there would be the aspect of a recipe as well. Apart from having the data. The data and the recipe together gives you like the perfect dish or like an amazing large language model. I think yeah yeah, yeah.

**Craig Smith:** 43:33

And then there's the other. I've started talking to people about other AI Architectures or strategies beyond that generative. You pre-trained transformer models. I mean, there is that. That is People. It's certainly in the popular mind today. That is AI. But AI is a very large field and just another question related to data, data collection, and Probably you won't be able to answer it, but you certainly could take a stab at it in a more educated way than I could. Open AI. You know GPT-4, for example, has that. What percentage of the internet has that consumed in its training? Do you have any?

**Ahmed Imtiaz:** 44:26

idea. I think, like I don't think anybody has any idea other than open AI, because, with the rising trend of Like companies, they don't want to share the information of how they're building their, their flagship products. So, like, apart from meta, who's like releasing everything? Like? I'm completely pro open source. So there's, I'm gonna add something about this a little while later, but to answer your question, like it's very hard to say, like how much they're using, but I would assume they're trying to use all the good quality data that they can get, so we.

**Craig Smith:** 45:00

And let's talk about Lama then, because, again, you're pronouncing it differently than me, is it llama, not llama?

**Ahmed Imtiaz:** 45:08

It could be llama. Yeah, that's that, yeah, that's something. Yeah, I'm not sure what the perfect pronunciation for llama is, but Is it possible?

**Craig Smith:** 45:20

At this point or very soon, it will have consumed all of the internet. The put, not the dark web, obviously, but but the public internet.

**Ahmed Imtiaz:** 45:32

Yeah, like there's gonna be this limitation, like it also requires some time to train these models, you know. So suppose, like you're in 2021 and you take all the data and you start training your model and then like it'll, it's some time to develop your model, so it'll take one more year, so you'll always be in a situation where you're maybe behind by a year, right, so that's what we see in chanjibu as well. We have data up to 20 20. So I think that that effect would be there. Like we would not always have like that, like everything consumed by an effort. But, yeah, I think, like any, any good amount of good text that exists, I think these models will be using it. The social media company. So social media data is very different from the data we have online in general, like in the block sites in Wikipedia and stuff. Right, social media data is very unstructured and it can be very Like. It's very hard to sort of quality control, like from my personal experience, like from our company, when we work with social media data, for we generally, what we do is we crowdsource data and we open source them to crowdsource the solutions as well. The way we crowdsource solutions is to compete. So we have like a kegel competition running right now for like $50,000 price money on Bengali speech recognition. So we basically crowdsource the data online to influencer campaigns and through targeted social media campaigns and then we just crowdsource the solutions through this competition, right? So in these settings, when we're crowdsourcing data, if we ever try to crowdsource data through social media, we always see that it's very unstructured in the sense that, like in social media, is very hard to control for like that that's natural, natural, like more natural text maybe, but more natural text is like it has like a lot of Like bad words, it has like Translutorations, it has like code switching and like spelling mistakes so many different things, right? So it's harder to work with social media data. So maybe, like it could be that Like the social media, the data that exists and out in the public, to be used, to be used for training. Maybe them, the companies will work towards like other, maybe they already have methods to clean that data up, like so it's gonna be case specific like social media data would have to be handled differently compared to Wikipedia data or something, but I think people would be doing that to get more quick quality data to train.

**Craig Smith:** 47:56

Yeah, and then you were gonna say something about LLMs. I have another question, but I don't want you to forget what you're.

**Ahmed Imtiaz:** 48:02

Oh, I was talking about, like the, the, the openness like that, like open sourcing, I feel like is so important. So there was a, so there's an archive paper that David Donald was like a legend in the field. He he posted a couple of days ago where he says that so he takes a stab at the, the existential crisis question, right, and, and what he says is like, and which resonates with me so much, is that there like the three things that have accelerated this field, and those three things are also gonna are also it's also gonna ensure that we don't have some like we don't have a situation when we are gonna like not exist in it, but and the three things is like codes, code sharing, sharing code and recipes between like people, like for free, coach, coach sharing the. The second thing is data sharing, which we like we have like in all the like, all in everything in AI has like happened because of data sharing. We have to invest in them right, like someone open source that faithfully, like has completely changed, like how they completely accelerated, how Compatibility in progress, and so, and the third is competitive, like competitions. So these three together Like sharing large language, like if we think of large language models, sharing the code and data for large language models as well as sharing the code and data for these, like for a generative model, like stable diffusion as well, right, so these are gonna. These, I feel like, are very important to be able to assess these networks so that we can actually tackle, like any existential crisis question that can come up in the future, right, so yeah.

**Craig Smith:** 49:48

That's interesting. That's interesting. The question I was going to ask is about open source, because I've been debating with some very smart people on this who take the view that open source will not dominate because there's too much value creation in proprietary models and the real. In order to create that value, you need tremendous financial resources, and the open source community will never have resources to match the proprietary models, and certainly meta is the exception. But the question is whether or not meta will continue to invest in open source models and there's a big question mark as to why they would do that, given the cost. And then there's you know that's interesting what you say about the existential threat. But these models are extremely powerful. Once they're open source, people can remove the guardrails and do whatever they want with them. And I mean I think Jeff Hinton, or at least Joshua Bengio quoted Jeff Hinton as saying you know, would you open source nuclear weapons? You know it's too dangerous to hand to any potential malicious actor out there and unfortunately there are a lot of them. So how do you feel about that debate? First of all, do you think open source can marshal the resources to dominate or at least compete with proprietary models? And what about the dangers of open source?

**Ahmed Imtiaz:** 51:42

Yeah, like working in the open source domain, like myself, I think I agree with this partially that it would be very hard for open source to match the performance of proprietary models Because, like, open source efforts are basically community efforts. So wherever you need capital for making something better, the closed source models are going to be better, right, they have more. So suppose, like we run out of real data on the internet, only the big companies have enough capital to hire teams to create new data. So the open source efforts might. So it could be that some big company does that and open sources the data. So that would be completely on them. But just for the open source models, like what the development that's happening with open source models right now? It might be obvious that it could be hard to compete with them, like proprietary models, right, but the benefit is not in performance when you open source, so the benefit is in understanding what these models are doing. There are so many research groups who are studying the open, like Lama two, for example. They're studying these large-length models. You see, like when they're failing, why they're behaving this way, how we've repressed them better, and when something is open source, like it's, it's theoretically you have like n number of heads who can work on on this problem, right, compared to when you're in a company you have, like you have, the number of heads is limited by the budget that you have to hire people, right. So, it's always going to be like open sourcing models is always going to allow more people to solve the problems with AI that we see that could harm people or that like them like having bias. So there are people studying the bias in networks. That's why they're being able to do that with the open source models. So I think open source has the benefits in that direction compared to like. Maybe they won't be able to match the performance of proprietary data proprietary models someday, but they'll be able to help the people working with proprietary models to make decisions on how to make them better as well, with benefits everywhere. Benefits humanity, like better models. Help everyone, yeah.

**Craig Smith:** 53:53

Yeah, it's interesting. On the dangers of open source, I don't remember how long a couple of years ago, I interviewed a guy named Connor Lee. Do you know him? He and he started a Luther AI, an incredible organization that is now incorporated, but at the time it was a totally decentralized group of hackers on Discord and I think it was when GPT two maybe was when GPT three came out. They said, hey, let's build an open source project. And they built GPTJ and now they have Neo I can't remember the name of the models that they have now but but but Connor became so freaked out by the dangers of these models being not only existing but being open source that he quit a Luther AI and he's has a startup now that works on the alignment problem. On AI safety, what do you think about the danger of open source, of having to turn these powerful models over to whoever wants to use them?

**Ahmed Imtiaz:** 55:10

Yeah, but we don't have one. There are so many risks involved, like like, when the open source, really strong models like like we see fake websites full of text that's being generated right, fake news these are going to be like super easy spamming people like you can like generate emails and send millions of emails out, Like we had the Nigerian Prince before. Maybe now we have like the AI overlord from Mars who sends emails to everybody. So these situations could happen, but like this is my belief completely, so what I'm going to say now. So I believe that when we think of something like, when we open source, we send it out to all, everybody, right, all the people. And when we have such a big number of people, the law of large numbers should come into play, right, so we would have some modal tendency of doing something and then we would have tails. And I believe that humans in general, we are the mode to be towards good, because humans, we want to survive in society. We have to like talking nicely with other people, right, and we have to behave nicely. We need to help others so we can get help Right. So I think, like, because of that reason, when we open source, the mode would be towards good, but there are always going to be people towards the tails who would want to do bad with it. And then comes that, then comes the question: how much resources do these people have? So if we have bad actors with a lot of resources, that's when bad things can happen, like if you're talking about open sourcing nuclear nuclear missiles, right. So if you have someone who has the K-pad, like the capital to build like nuclear missiles, like they have access to uranium and all the everything needed to build them, that's when it becomes so dangerous right. But then if you don't have access to that, then it's just yeah, I have another piece of paper that says, like how to make a bomb, a big bomb or something, right. So I feel like it's going to be subject to that person, to us humans. It's going to be on us humans who have, especially, who have capital right, who have the capability to use this in a bad way. We would. Then that's when we need to like, that's where we need to think about like okay, so as long as the people who want to do good are like, being able to like maybe triumph over the people who are trying to do bad. As long as the FBI is hunting all the people who like using this to do, maybe like the worst possible things on the dark web, like the good will try them that the mode will be towards the end.

**Craig Smith:** 57:40

That's my, that's my belief, that's what I think yeah, so.

**Ahmed Imtiaz:** 57:43

so, while we like this idea of self-consuming neural networks, like, we have these amazing collaborators on this paper, especially the like, I must mention the the first two authors, sina Ale Mohamed and and Josue Casco Rodriguez, who are like, who, like, literally spearheaded this research and brought these amazing insights up. So, like, like, everybody go check their papers as well. They're like amazing researchers.

**Craig Smith:** 58:10

AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested, so buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete with costs spiraling out of control? It's time to upgrade to the next generation of the cloud Oracle Cloud infrastructure or OCI. Oci is a single platform for your infrastructure, database application development and AI needs. Oci has four to eight times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing and, of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less, like Uber and Co hair, take a free test drive of OCI at oracle com slash I on AI. That's E Y E O N A I all running together. That's Oracle.com slash I on AI. That's it for this episode. I want to thank Ahmed for his time. If you want to read a transcript of the conversation today, you can find one, as always, on our website E Y E hyphen O N dot AI. And remember the singularity may not be near, but AI is changing our worlds, so pay attention.