

Tony (00:04):

Welcome to Code Together, a podcast for developers by developers, where we discuss technology and trends in industry. I'm your host, Tony Mongkolsmai.

Tony (00:17):

Generative AI has become extremely interesting and powerful in the last few years. Text to Image models like Dall-E and Stable Diffusion have become very popular and NLP models like GPT-3 have given rise to applications like the newly released OpenAI chat bot, ChatGPT.

To understand the power of these tools, I asked ChatGPT, "Why is Empire strikes back the best Star Wars movie?" The reply was this:

It is a matter of personal opinion whether Empire Strikes Back is the best Star Wars movie. Some fans and critics may consider it to be the best film in the franchise for a variety of reasons. For example, the film continues the story of the original trilogy in a compelling and satisfying way, introducing new characters, and revealing key plot points. The action sequences are also considered to be some of the best in the series, including the battle on Hoth and the lightsaber duel between Luke and Darth Vader. Additionally, the movie features some of the most iconic moments in the Star Wars franchise, such as the reveal of Darth Vader as Luke's father.

I don't think anyone would argue with that assessment. Feel free to disagree with me later on Twitter.

Generative AI is changing how we think about computer human interaction, content generation and even learning. Today we're going to talk about what this means for us as developers, some potential societal impacts generative AI may have, and some of the ethical and legal issues.

Tony (01:42)

Today I'm joined by two AI experts from within Intel. Ezequiel Lanza is an AI Open Source Evangelist, works on several AI open source projects, helping people discover the world of artificial intelligence. Ezequiel is a frequent AI conference presenter and the creator of several use cases, tutorials, and guides that help developers adopt open source AI tools. He holds a Master's in Data Science. Welcome to the podcast, Ezekial.

Ezequiel (02:00)

Hi Tony, nice to meet you. Thank you for inviting me.

Tony:

Ben Consolvo is an AI Solutions Manager at Intel and lives near Houston, Texas. He has a Master's degree in geophysics from the University of Western Ontario and enjoys digging into AI to solve some real world problems. Nice to have you Ben.

Ben:

Nice to be here Tony.

Tony (02:22):

Alright, so let's just go ahead and dive in. Ezekiel, who is our generative AI expert here is going to tell us a little bit about what is generative AI, how does it work and why should we be interested in it.

Ezequiel: Well it's pretty interesting. Basically, how we can start explaining what is generative AI: we can think of it as a program that can for instance use the existing content it can find like text, audio files, images or whatever, to create new plausible content. Basically, what it means is it enables computers or other devices, to learn the underlying patterns related to an input, and based on that they can create new general content or similar content.

How it works underneath is, over the last 5 or 6 years, there was a huge advance in a new paper called "Attention is All You Need." All these algorithms are based on the transformers architecture.

Basically, and without going in detail, what the transformers architecture uses is it has a layer, a particular layer, compared with the other deep learning architectures that it's called attention. What it basically does is, once you are training your model, this particular layer enables the algorithm or the model to understand which parts are most important in this input. It's similar to when you are talking with a person and you have a lot of noises around you.

For instance, you have street noise and you have a lot of different noises, but you are paying attention to the person. You are paying attention to some particular thing that this person is saying.

So you are using your attention layer in your mind to pay attention to that particular part. And these algorithms can do something similar like that. So when you, for instance, when you train those huge models because they are very huge and you have basis for instance like Wikipedia, Reddit and all the text that you can find on the on the internet sites, you can feed this algorithm, you can train this algorithm, and this algorithm will be able to get those insights of those patterns in the phrases.

So, once the algorithm knows how a phrase is written, how people write, they can try to emulate for instance something similar. So, if we like to teach a transformer with a lot of books of Shakespeare for instance, the model can understand how is Shakespeare's way to write. So, if you say, "Hey can you complete the phrase", they can complete as it was Shakespeare.

So, it's pretty interesting; the reality is that as you may think as you as you may imagine is that you need a lot of data, you need a lot of compute to train those kind of models. But this is what some companies are doing and once they have this model trained, what is very useful for the developers is that they can use this kind of models to adapt to their solutions.

Basically, what the generative AI scope: it's going really big in the next decade, in the next years or even in the next days or weeks. It's pretty amazing, but everything started five years ago or four years ago with this transformer architecture and now we are starting to see some real Benefits, which is pretty awesome.

Tony [06:27]

Yeah, and I think both of you guys are really interfacing with the community at large. Ben actually does a lot of work with trying to interface with developers. Ben, how do you think that this kind of affects developers, this type of generative AI model?

Ben:

Yeah, there's so many ways that it can help with things that developers are working on. I mean obviously, in particular, you think about code generation and you know looking at things like GitHub Copilot where you can have some code generated off of a docstring, or off of an initial line of code.

And then same with debugging, maybe entering some code and then having the bot be able to debug some of your code or find an error with it, that might take you a lot of hours.

How I see the AI helping is freeing up more time for creative thinking, because some people would say "well this AI is just going to replace my job" and that could be partially true but I think the creative thinking aspect of kind of stitching things together won't go away, but that's my opinion.

Tony (07:52):

Yeah, I think that sounds probably right. I think one of the interesting things that we see a lot of online, as I look around, is people saying, "how do I prompt the generative AI the right way to give me the answer that I want."

I've seen a lot of people talking about the different parameters that you need and it's almost the same thing where somebody says, "Hey, can you help google this for me because I'm trying to find the right answer online. I need the right google search words."

It's kind of the same thing here, where how do I know what to tell either ChatGPT or Copilot or even Stable Diffusion or DALL-E what I want, so I actually get the output that I need.

I don't know when you guys were playing this technology, if you guys had any experience with something funny where you type something in and you kind of got a goofy output from what you would have expected.

Ezequiel: (08:43)

Yes, I mean it's always funny when you try to have a conversation with this kind of algorithm or models, because they can behave as a normal person, but for instance when you try to

to ask for some code, or when you try to find some bugs, or even if you say “Hey, which is faster to run in one place compared with other places so the answers are always really, really fun.

But I believe that it's something that, as Ben said, I mean we need to think the technology as something that can help us to co-create or help us to guide. Of course, it could be dangerous—a lot of people are scared about, okay this thing will be replacing people, will be changing everything. I think that it's a really good tool to try to help. And, it's also really needed to define the boundaries, because we need to know where we should use that.

And I think that will be happening in the next years, or the next months. Maybe some companies will start to monetize that, will start to find those use cases where this can apply. It could be really fun if you like to use stable diffusion, you can create a video, you can create things that are really, really fun. You can create filters or different images from your face for instance.

It's really, really funny but when you try to see, okay, how does it work in the real case scenario? Is it useful for a use case? Is it a business case behind that? I think that this will be another thing that will be in the future to allow this thing to do to be even bigger.

Tony: (10:43)

Ben why didn't you tell us about, we were talking about prompts that are weird, and you mentioned one that you saw online on Twitter.

Ben:

Yeah, I'd seen it on LinkedIn. It was somebody basically typing in something about proving that a rational number was irrational. And the interesting thing is that it comes up with a proof by contradiction to prove that  $\sqrt{4}$  is irrational into ChatGPT, and it goes through the proof and concludes that  $\sqrt{4}$  is irrational. And so yeah, there's obviously some, you know, you can't trust this completely. And it's I would say really dependent on the, you know, the input data largely, you know, is what informs these models. And, so it's just interesting, you know, that sometimes it can come up with something that might look correct, you know, if you didn't know any better. And it comes up with kind of an interesting statement, but it might be entirely untrue.

**Tony:** [00:11:56] Yeah. And I think that that's one of the things that we want to look at a little bit, right? What are the limitations of these models and the applications of them? So, for instance, some people online were saying that potentially something like ChatGPT could replace Google because it gives you a very well-formed answer to a question that you may ask. But I think the challenge is as you kind show there, Ben, is that if you ask it to tell you something that is incorrect, it will tell you in a very nice way that sounds very true that what you've asked is, you know, that  $\sqrt{4}$  is irrational. But yet that isn't the case, whereas at least with Google, whether you like Google or not or any other search engine, you can see the source of

that information and kind of determine whether or not you feel like that's a valid source of information. [00:12:41][45.1]

**Ben:** [00:12:43] Yeah. And I think that one, you know, one potential help or, you know, to help answer that kind of problem would be, because these generative AIs can generate multiple answers and they usually just select one of the best ones, you know, based on some accuracy score or some metric. So, you could actually, you know, have ChatGPT give you like six or seven different generated responses and maybe be able to critically think about some of those options just like you would in a Google search. And you're not going to go necessarily just with the first you know, the first result. But maybe you'll look through, you know, 10 or 11 sources and see, okay, well, maybe from all these different perspectives, just this is what makes sense. So I'm going to, you know, formulate my thinking on it based on looking at multiple outputs. [00:13:35][51.7]

**Ezequiel:** [00:13:37] Yeah. And also, I think that what can be pretty interesting that maybe it's something that OpenAI I believe that it's working, is that all the information that we have in the ChatGPT or these models is information that it's already I mean, it's trained and the model has this information. I mean, it can be true. It cannot be true. The problem is when this model can start to reason or start to do some reasoning. But I think that the next level could be okay, now that we have the model that we have that we can understand. What if we allow these models to try to do a web search instead of doing your answer?

I mean, they can be trained, but they can double check with web search, which is another challenge, right? Because when you do a web search and you need to okay, the algorithm should read the web page and maybe it's not text, maybe you have JavaScript, maybe have pictures, you have images. You need to read these images and it has a lot of technologies that can that can go over there also, because you need to convert the image to text and you need to reason after that.

So, I think that this could be another thing that could be very interesting in the next years. Okay, we have the models; they are very capable to understand the patterns, the language and so on. But let's try to use, or let's try to fix the boundaries, give them more tools to double check or to cross-check, instead of having a person right in this particular case. I think that this could be pretty awesome also. And I know that OpenAI is working on something like a web API or something like that, but it's trying to provide web access. And if you try with chatGPT and you ask some questions and they say, "I cannot browse; I cannot go to the Internet to look for your answer." Right. [00:15:42][124.9]

**Tony:** [00:15:43] Yeah. And one of the I guess that we can move on kind of a little bit from the technology to kind of the ramifications of the technology. One of the many questions that I see online and where people are talking about, especially the image generation, but it also applies to the text and code generation type generative AI, is, what are the copyright issues? Who actually owns the output of these models? So, kind of in general, I guess, just what are your guys' thoughts on that? I know that there's lawsuits and things like that already around Copilot,

but what do you guys think in terms of either, not the legality, obviously, we're not lawyers, but what do you guys think about like how does this affect the applicability of this type of generative AI? [00:16:31][47.6]

**Ben:** [00:16:33] Yeah. I think just thinking about the kind of the training data in the first place, at least I looked into GPT-3 training data and it essentially is a conglomeration of internet sources, you know, Wikipedia, a common crawl, some books and you know, all of that training data is like, you know, from all kinds of sources on the internet.

It's like almost like a general human, you know, knowledge base or at least what people have chosen to present on the internet in these different places. And so in that sense, it feels like it should be open because it's, you know, the data source itself is coming from all these different people all over the globe.

But then, you know, I think about too the generation of the model and the work that goes into that and the company that's responsible for that, too. And so there's that side of things where you're looking at, well, this company is the one who put in the work to make this happen and make it a reality for people. So those are just a couple of thoughts that I have. [00:17:47][74.3]

**Ezequiel:** [00:17:49] Yes. Completely agree: since GPT and all these other models they are trained with the data that you have in internet and that that is written by people that has internet access that we can think that is most of the people, but it's not the 100% of the world. So when you are trained as models, you can get some bias because you are not representing the 100% of the population out there or 100% of the way to think or to write.

Even it happens if you go to the Huggingface model and you try to use it to complete phrases or to complete a phrase, and you say and it's really, really frustrating when you say, okay, the white person works as an \_\_\_\_, as something, as a lawyer. And if you put the same thing with the black person works as \_\_\_\_, and the result is completely biased. So, this is a really problem that we can find when we are working with this kind of models that are trained with popular data, with the data that is available.

And this is a problem that we should be aware when we are working on that. And this is why responsible AI and ethics AI is so important to try to avoid those biases, to try to make it more compatible, to make it more reasonable, instead of just working for the majority, right? And I completely agree with the copyright stuff, because now if you see an article written by a person and compare with an article written by a ChatGPT or GPT, and I mean, you can try to get confused, right? I mean, are you able to understand, or are you able to identify which is the person, which is the computer or the algorithm?

And this could be a really, real problem, because what if in the future we have a text or we have something that is created by these algorithms and who is responsible, right? It's not the algorithm. Because when you the, when you download the algorithm, you have an open

license, when you are responsible from, they are not responsible from what the algorithm can give you. They put disclaimers, they put a lot of things. So, who is responsible then?

So, if you are building an application using this algorithm, you should be responsible. So, you will be against the law or not. But who knows? And this could be a really challenging, and it happens with all the other technologies; for instance, that the laws or the things that people that the lawyers start to see or to pay attention maybe it can happen in the next years. Right? So, first we have the technology. [00:20:49][179.7]

**Ben:** [00:20:50] Yeah, no, I was I was reading about this bias issue in the GPT-3 paper and I thought that they summarized it really well, so I'll quote just like a couple of sentences from their paper on the bias aspect of the training data. So, they say "Broadly, our analysis indicates that internet-trained models have internet-scale biases; models tend to reflect stereotypes present in their training data. Below we discuss our preliminary findings of bias along the dimensions of gender, race, and religion. We probe for bias in the 175 billion parameter model and also in ... smaller models, to see if and how they are different in this dimension."

So, they go through, you know, religion and race and gender, and they kind of highlight some of the biases that are present in the model that they've trained. The interesting thing to me is that the biases are reflective of everything that's present in the internet data. So, it's at least a representation of whatever's present online. You know, that's not like a good thing necessarily because there can be a lot more, you know, potentially biased information on these topics online. But it at the very least reflects, okay, well, this is what the training data reflects.

But most people, when they just look at an output there, they're not going to think about that. They're just going to look at, well, that's not right, that's not fair. So, it's an important thing to address. But as a data scientist, it reminds me okay, well, this is where, you know, the data came from, and that's how we can explain why this is the case. [00:22:37][106.7]

**Tony:** [00:22:39] Yeah, that makes sense. And as a developer, I was obviously really interested in GitHub Copilot when it came out, and also ChatGPT because you can ask ChatGPT to generate some code. So, for Copilot, there was, I think Professor Tim Davis at Texas A&M posted on Twitter that he had a library and it seemed like that when he asked for Copilot to generate code, it was pretty much generating his code verbatim.

And that's a very specific type of example. It's a very what's the right word? It's not a general algorithmic problem. It's a very specific algorithmic problem. So, it would make sense that Copilot wouldn't have a lot of examples to draw from. But similarly, I tried to do something with ChatGPT yesterday and try to have it generate essentially a oneAPI SYCL Mandelbrot code, which I would also think is relatively fairly specific, being kind of in the space where Intel and oneAPI live.

And it generated some code that looked very reasonable and would run correctly, but it actually didn't match up to any of the source code that I could find online through Google,

through all of the common sources where I would expect it to have drawn from. So, in that case, I don't even know whether or not that would be considered something that should be copyrighted or not?

I thought what was really funny, Ben, is when you were saying, I'm going to quote their paper, that's something you don't get from chatGPT or Copilot. I have no idea what the source is there when I get it from them, whereas at least from you, I know where's the providence of this information and how trustworthy is this this source. [00:24:14][94.9]

**Ben:**

[00:24:15] Yeah, definitely. Yeah, looking at, you know, when you're speaking to somebody, you know, that it's coming from their own thoughts, from their brain. And then, you know, as I'm citing a specific paper, you know, you can see this is where it's coming from. But yeah, when it generates something, you don't necessarily know the source of where it's generating it from. You know that it's broadly from this huge database of I think they said in the paper, like, there's a trillion words or and basically more data than they can even train on, like to get through the whole training set. But yeah, you don't know specifically when you ask a prompt, you know, where that's originating. [00:24:56][40.8]

**Tony:** [00:24:58]

Yeah. And this technology is definitely moving kind of how we think about various technologies in society forward. So, I also saw on Twitter about this the other day because I was looking for Twitter to see what interesting takes people had. The Box CEO said, and I'll quote him, "There's a certain feeling that happens when a new technology adjusts your thinking about computing. Google did it. Firefox did it [I'm not sure Firefox it parenthetical], AWS did it, the iPhone did it and OpenAI is doing that with ChatGPT."

And that's an interesting thought too, because I think that a lot of times technological advances are not seen by the general public. But it seems like with something like ChatGPT, there's a lot of people who are not even in the technology space who are trying it. I've got friends who are artists, right, who are using Stable Diffusion and Midjourney and DALL-E.

And I've got, you know, all kinds of people who are not in technology playing around with ChatGPT just to see what kind of answers they get. So, with kind of this shift that's kind of been suggested by this generative AI, where do you guys think this technology might take us in the future? I know, Ezekial, you touched on a little bit, but is there some more directions or thoughts you might have on where does generative AI take us in the next five or ten years? [00:26:14][76.0]

**Ezequiel:** [00:26:15] Yes, I think that what the next years, we need to chase these boundaries, right, because, as you said, when the technologies, it's able to be used to people, which is not tech, it's not technical people, it's when it really explodes. Because it should be easy to use, easy to understand, of course, they won't to go in details of how generative AI works, but if

they can use it, if a developer can download it and can create a new solution, it will really, really help.

What I think in the future is, of course, the text part of the giving recommendations is now happening, which some companies that are offering, I know if you like to write your blog post or whatever, you can use these algorithms to give you advice. The future is about video, image and also music.

I'm not pretty confident with the music thing because I don't know if I would like to listen to a music artificially created. But I think that the next, probably the next step, will be related with the video, with the image to video, and how to get this information. Because what I think is AI in all the algorithms, you need data to train that and if you like to capture the data and if you have an algorithm that can capture this data; for instance, we can capture the YouTube data.

If you have Stable Diffusion model that can convert all the data that we have in the videos in text, I can use the data for something. I don't know for what, but it can extract this data. So, I see that this could be in the next years what could be used. But again, I mean, I believe that we will start in next year to listen to new songs created by these algorithms, which will be really fun to listen. But the realistic part, it's related with video and images. [00:28:34][139.6]

**Ben:** [00:28:35] Yeah. No, on the, on the music part, Ezekiel, I think it was maybe a couple of months ago, I listened to a video by TwoSet Violin. They're a music YouTube video channel, and they had a test where they listened to classical music that was written by an actual composer in the past, and then they had, you know, another piece of classical music that was written by AI, by generative AI, having been trained on, you know, some style or genre of classical music.

And in some cases, yeah, they weren't able to, even as trained classical musicians, weren't able to pick out the composer from the AI-generated music. In some cases the, the AI just was very repetitive and some of the themes didn't make a lot of sense. But in some cases, the music was very realistic sounding for that kind of composer and that timeframe. So, I think music is something that will be improved upon and generated and a really interesting part of this. [00:29:47][71.7]

**Ezequiel:** [00:29:48] And also it can even help because when you're creating, you know, I used to play the drums and when you are creating music, you need to try it. I mean you to have something in your mind and maybe it can help you to create a new melody. And from that a musician can create a new song I don't know. [00:30:08][19.3]

**Ben:** [00:30:08] But so then are we super... does it just make us super lazy or does it make this more creative? [00:30:14][5.4]

**Ezequiel:** [00:30:15] I don't know. I don't know. It's the same thing with the text, right? And if you are creating a blog and you have something that is helping you, it's helping you to make it

more creative. Or you can be lazy all the time. I mean, you can say, okay, let's let the AI to do everything and that's it. But I mean, this is not the point, right? I mean, we should use a technology to help us to make it better, more creative, and so on. Yes, it's not easy to find this point to be lazy or creative, right? [00:30:50][34.7]

**Ben:** [00:30:51]

Yeah, one of the things I thought about for text generation, like ChatGPT, I haven't tried this on the platform, actually the platform because there's so many people on it right now, it's kind of shutting down, like they have too much traffic, they don't have enough capacity. So you can't actually write a lot on it right now.

But I was thinking about, hey, can I just ask it to write an essay for me? Like if I were a student and I wanted to write an essay and I kind of give like a little bit of an abstract on the topic I wanted to write on, and then, you know, how easily would it generate multiple versions of that? And could I actually present that to my teacher or my professor as a grade school student and it seem feasible for that level of writing? That was one of the, I guess, potential abuses because, you know, it might make somebody lazy and not actually creatively think about how to write and express themselves. So that was one of the things that that I was thinking about. [00:31:46][55.1]

**Tony:** [00:31:47] Yeah, there was an editor, I want to say it was of TechRadar, who actually said, "write me a 350 word article about chatGPT in the voice of ...". And he used his name and because he is well enough known and has enough public writing, it actually wrote a 350-word article for him, which he literally put an intro saying, "This is the prompt I used." And then he just pasted the article and that was his story. And it was really interesting. I'll try to find the link and put that in there for the listeners. It was pretty cool. [00:32:19][32.0]

**Ezequiel:** [00:32:20]

Yes. And I believe that the artist part, I mean, the creative part, it's something that we always say that, okay, we as human beings, we have something completely different. Like is when you when you're talking about art, that is a particular touch that someone can have and so on that is different and you can't find yet.

And with AI, even with fashion, and when you're designing clothes and so on. And it recalls me and use case that it was, it's not the same case, but it's similar: 10 years ago, or 20 years ago, with the Gap CEO, he started to say, okay, let's try to fire all the designers that we have, all the fashion designers that we have and let's try to find or to design the clothes based on what the market says, right?

So, he started to capture data from their sides, from all the data, from the websites, from the stores to say, okay, people will like to buy this kind of thing from jeans or whatever. And it was, and the result at the end was pretty bad. Pretty bad because at least in fashion, it's a particular skill people uses what you see, it's not what you want, it's what you see. So, if you see a jean, if you see a LeBron James wearing, I don't know, his shoes, you probably would like to use it.

This is how our mind works, right? I mean, I'm not saying nothing new. But in the fashion style, for instance, it was a complete failure. So, this is why you need a fashion, you need a designer, you need people designing clothes and so on. And I see something similar with art, that we have this touch, I don't know, something completely different that AI can help, can provide us some insights, some data. But it's not like a real person. [00:34:24][123.6]

**Ben:** [00:34:27]

Yeah. One of the one of the interesting things with art is, or I guess just this technology in general, so it's using this corpus of training data, whether it's images or text, and it's limited in what it can generate to that corpus that, you know, that set of data. So, yes, it's generating something new, but it's really just kind of a combination of what's been before.

Whereas, you know, maybe with a fashion designer like you were talking about, somebody might come up with something, I guess, quote unquote, brand new, that hasn't, that's not totally a combination of something that's been done before, although I can argue with myself on that point, because a lot of what we come up with is based on some past experience. But I still wonder, is there an aspect of like the human creativity that comes into play there that, you know, a trained algorithm just can't come up with something that's totally new and unrelated to anything it's done before? [00:35:37][70.5]

**Tony:** [00:35:39]

Yeah, I'm hoping they bring back the nineties clothing, so I can use what's in my closet. I'm not sure the algorithms will come up with that or not. Since this is a podcast where we typically talk about programing, developers, and hardware solutions, you mentioned, Ezequiel, the amount of compute resources to actually train something like ChatGPT.

So, NLPs are well known for using a large amount of compute resources. One of the cool things that we have now with Stable Diffusion is it's small enough that I can run it here on my laptop. I can run on my workstation. I'll put a little plug in, Intel has solutions, OpenVINO. We can run it through PyTorch. You can get Stable Diffusion generating images on your computer. Is there a concern that you guys would have around the accessibility and the democratization of these types of solutions for kind of how other people can use it, how businesses can use it, how individuals can use it? [00:36:45][66.2]

**Ezequiel:** [00:36:47]

Yeah, I think that the open-source mindset, it's pretty important because all these algorithms, when people would like to use it, people would like to download the algorithm and they would like to implement into their solutions. So, what it means is that these algorithms should be freely available. Of course, it's up to you, your application, but it should be available to be used to any, any developer.

So, the problems that the developers now can have is, okay, I have a huge model that can understand or that can create something, like Stable Diffusion, but where can I run my

application? Okay, so now I should be worried about what is the compute? What is it compute that I will be using? We will be running that in the cloud, in a computer. We'll be optimizing. And this is why that there is a new trend that it's really interesting that it's about how can you optimize or how can you get the optimizations of those transformers? Because they are they are huge networks.

Of course, you don't need a huge power to train because somebody did it, but you need power to run this inference is when you are offering a solution or we are offering something. So, I believe that the power to make it available is really interesting to enable new use cases. But we also need this new trend to get optimizations, to get it running. I mean, the hardware is always needed. We always we need hardware to implement that, even if it's in the cloud or whatever. You need hardware and you need to have this hardware optimized to be working with this with these huge models. [00:38:35][108.2]

**Ben:** [00:38:36] I agree. You know, even with these large transformer networks, these neural networks, you need significant hardware for inference. But the initial thing that came to mind when thinking about this question is, originally, you know, when the, for example, the GPT-3 model was trained from scratch, you know, it required a ton of compute. I think it was over a thousand petaflops or something and they were only really able to train it once. They actually, in their paper noted that they had some overlap between their training and test datasets, but they weren't able to retrain it because it just takes so much money and so much time.

But I think one of the things that makes, you know, these models now so accessible is that, yes, we have the pretrained, we have the pretrained version. Now we can fine tune it with a lot, a little hardware compared to the original training and the fine tuning just does such a good job downstream of this original trained model that we can use it for a lot of different tasks and relative at least to the original training cost, it's almost nothing to do inference.

It's still, you know, a significant cost that you need if you're going to be, you know, 24 hours a day running some kind of GPU or accelerator. But yeah, I think what makes it largely accessible is the fact that, you know, you can fine tune this with relatively little cost, training cost and then the inference is still runnable on like a single GPU. [00:40:13][97.2]

**Tony:** [00:40:15] Well, I think that's about the end of our time today. I'd like to thank our listeners for joining us, and I'd like to thank Ezekiel and Ben for joining us and providing their opinions about generative AI. [00:40:24][9.5]

**Ben:** [00:40:25] Yeah, thank you for having me. [00:40:26][1.0]

**Ezequiel:** [00:40:27] Thank you, Tony. [00:40:27][0.0]  
[1738.3]