Speaker 1: This is ACM ByteCast, a podcast series from the Association for Computing Machinery, the world's largest education and scientific computing society. We talk to researchers, practitioners, and innovators who are at the intersection of computing research and practice. They share their experiences, the lessons they've learned, and the own visions for the future of computing. I'm your host today, Scott Hanselman.

Hi, I am Scott Hanselman. This is another episode of Hanselminutes in association with the ACM ByteCast. Today I'm chatting with Dr. Peter Lee, President of Microsoft Research. You've got a resume as long as my arm, and it is an absolute joy to chat with you today, sir.

Speaker 2: Oh, it's a joy for me to be here. I really appreciate it.

Speaker 1: I want to go in a little bit of a controversial direction because I want to start out with the fact that I had a birthday last week and in this year, if I understand it correctly, you're going to turn 65. Is that correct, sir?

Speaker 2: Yes, which I can't believe.

Speaker 1: Yeah, yeah. I turned 51 last week and I have been reflecting on the space that I've been in the last... I've got 32 years of software experience and you have more. And what does that feel like? Because, if you can see behind me, I've got a PDP-11 that I built from a Raspberry Pi, and I've been learning about that stuff. I've got my Commodore 64 that I had as a child. I'm really reflecting on how far this has come. And with the experience and the space that you have occupied in computer science over the last forty-plus years, what does that feel like as we sit here on this AI moment?

Speaker 2: Yeah, first off, since you mentioned PDP-11s and Commodore Amigas, there's a lot of fondness in my heart for those. Actually, my first real paying job was a system administrator for a PDP-11, and so I remember learning how to wire-wrap the core memories. There, and then for Commodore Amiga, Amiga not 64, I took time away from my graduate studies at Michigan to be part of a startup. And in that startup, one thing we were trying to do was put a productivity software like word processing and spreadsheets and paint programs onto Apple IIe and Commodore Amiga. So if you ever used AmigaWrite?

Speaker 1: Yeah.

Speaker 2: Has my code.

Speaker 1: Really?

Speaker 2: And that company didn't really succeed, but we had to do things like make our own mouse because of course those computers didn't have mice at that time, but you needed a mouse to use those sorts of modern productivity programs. And so really a lot of fondness for those. I reflected about this, and Scott, I assume you have a mobile phone in your pocket.

Speaker 1: Of course.

Speaker 2: And the reason I think you have that is you look pretty comfortable and relaxed because I think nowadays if you forgot it at home or left it in a hotel room, you feel bad. You feel naked and vulnerable, like you can't really function properly. And from a research perspective, I count six major triumphs of computer science research that are in that mobile phone. There's VLSI design that emerged in the 1970s. There's Linux or Unix inspired operating system kernel. There is a software-defined mobile wireless radio and so on.

And these things were hardcore academic research and computer science departments, great ones around the world and great laboratories like Bell Labs and others. They got published and then they made it into a suite of technologies that you literally feel like you can't live without all day every day. And it's kind of amazing when you reflect on things like that. And so when I look at, for example, what's happening today with LLMs and generative AI, the question in my mind isn't one about AGI or not, it's about when and if say large language model technology will become yet another example of a technology that you can't bear to be without at any point of your waking hours. And I think it might happen, but just as I get closer to retirement at some point and just reflect on everything that's happened, it's really incredible how much things have evolved to the point where we really literally depend on them all the time for our mental health and sanity.

Speaker 1: That idea of not being able to live without it is such an interesting way of phrasing it. I have a very personal relationship with my phone because it also runs my artificial pancreas. And I was in Johannesburg last week for two weeks and I bought a backup phone because even though I can drive stick shift and just take needles and take blood sugar with my finger, I could do that. And I did that for 20 years. Why is it now that I feel like I cannot survive without... I have to have the phone within Bluetooth range near me 24 hours a day and if I lost the phone or if it was dropped or stolen, I would immediately need to hook another phone up because I want my continuous glucose meter. And my wife, who's not diabetic, but is a nurse, said, "Well, you could just take needles and stick your finger like you did back in the '90s." And I was like, "That's unthinkable. How dare you?"

Speaker 2: That's actually something called the flip. And you see the flip most vividly in the technology adoption history in healthcare and medicine. And my favorite example of that is ultrasound. So ultrasound had been invented quite a long time ago, but it wasn't until the late 1950s technologists proposed that ultrasound might be a good diagnostic tool in the practice of medicine. So if you're a pregnant woman and you go to see your gynecologist or obstetrician, you'll see ultrasound used to examine your unborn fetus and the reproductive system of the mother. But in the late 1950s when this was first proposed, it scared people. Is this going to damage my baby? Is it going to damage my ability to reproduce? Are doctors trained to read these grainy images and all sorts of problems? And so it took about a decade before ultrasound became the standard of care.

And so now, what do I mean by the flip? Well, today, if you are a pregnant woman and you go to your obstetrician and your obstetrician says, "Ah, I don't believe in that fancy ultrasound technology, I'm going to put my hands on your belly and use manual palpation." And not only would you be horrified at that, but you might even report that as malpractice. And so expectations not only in the standard of care, but in patient demands and expectations consistently flip in the history of technology, adoption medicine. And your example is another example of how things flip. You flip from being able to drive

manually and monitor your own blood sugar levels to a feeling that of why would you do that, it's so primitive, it's so inexact, so dangerous, unconscionable, and in fact irresponsible to do it that way.

Speaker 1: I'm curious though that I find myself feeling like old man who shakes fist at cloud because I feel like the young people are saying it's going to change the world. And if you are you know, okay, boomer, if you're against it, then if you're not with us, you're against us. But I remember when the TI-83 and the TI-81 calculators came out and the math teachers were like, "Ah, no one's going to be able to do math in their head anymore. You won't always have a calculator young man." And here we are with a pocket supercomputer. I am finding myself questioning my own opinions about tech based on my age and my generation and how I... I'm really trying to be introspective. I get that it's going to change the world, but I also think that cognitively we don't know how these are going to change our brains, and I don't know if having a pocket supercomputer has changed our brains for the positive.

Speaker 2: Yeah, I'm always a little bit more circumspect. It's easy as two people who are tech executives to look at all the historical examples of... So the adoption of advanced technologies that are really enabling. There is a consistent pattern. And so you could always argue as we in tech do that the right side of history is that these things advance and the world gets better and better. And I think that's generally true, but I think it's also true and it's worth some reflection that we lose some things along the way. And that isn't to say that the world isn't getting better and life isn't getting better because I think it is by any measure. But I think it's wrong to deny that we're losing some things and some skills and some abilities that maybe were important. And one thing I would say about, especially in this new AI era is I do see that the world's leading thinkers are actually earnestly trying to be thoughtful about this, that there is a big debate.

In fact, it's a debate of leading thinkers that I haven't seen since the human genome was mapped. When we finished mapping the human genome as a scientific community, it sparked a huge debate about what would this mean and would genetic engineering be a good or a bad thing? And the fact that there are so many leading thinkers and the whole academic and research area of bioethics really just mushroomed into hugely important kinds of thought leadership and research. I think that's helped us a lot to gain as much of the goodness out of our understanding of genetics while mitigating the potential downstream harms and risks. And I see the same intensity of debate going on with AI as people are trying to grapple with what this means. And so I think that that's a good thing and it actually makes it easier for me to be optimistic about what's happening right now.

Speaker 1: Yeah. As someone who is and was a professor of computer science for many, many years, did you plug ethics in and did you have people thinking about the tech? I mean, sometimes it's fun to just do techie stuff for techie reasons, but I've always come back to are we making someone's life better by doing this or are we making someone's life demonstrably worse? But a lot of people, when I do these informal polls, I say, "All right, who here has ever taken a computer science ethics class?" The hands are not going up. The majority are plurality. It feels like we're not teaching computer science ethics and it feels like it's even more important in a post-GPT-4 world.

Speaker 2: Yeah, I have to admit early in my career, it really wasn't a serious thought. And in fact, if anything, I tried hard to adopt the Silicon Valley stereotype ethos of tech is good and

more tech is better. And just the belief that technology could be the answer to all of our problems and that we can solve all problems through technology. I do think we've gotten to such a more enlightened state about this. I think a lot of that has come through lessons learned the hard way. We've seen so clearly that tech isn't different than any other technology and being dual edged, and in fact, it's more interesting and has greater positive and negative potential because information technology can be democratized.

It can be given to literally everyone. Unlike genetic engineering where you have to have multi-million dollar wet lab to do anything, here literally every person on the planet could gain access and harness this for both good and evil. And so I think ethics has become a much bigger deal, whereas I think earlier in my career, personally, I think it was an age of innocence where we only saw the goodness in things.

Speaker 1: Yeah, I really appreciate the democratization of it, like open source itself. The pancreas that I use is a toolkit that one could build from scratch and I did. And that anyone could go and do that, and now we're seeing open weights and open models. What's your thinking about frontier models that are closed versus frontier models that are more open about where they came from, what they were trained on in the corpus that built them up?

Speaker 2: Yeah, I think that this is also evolving really quickly. The thing that's so interesting to me is the cost of doing something first versus the cost of doing something 10th. And so if you are OpenAI or one of OpenAI's frontier competitors, you're trying to get to certain levels of intelligent capability first. And there you can see a pretty methodical investment strategy. To get to GPT-2 maybe requires on the order of five to 10 million of compute cost to train a model at that level. And GPT-2 allows you to see perfect loss curves, so that gives you confidence that further scaling is going to get you somewhere. And then to get to GPT-3 level and to do that first pretty consistently in whatever domain you're trying to train your AI models consistently requires about 10x the compute cost. So now you're talking 50 to a hundred million dollars.

And then to get to GPT-4 level, whether it's in language or molecular dynamics or weather modeling, whatever seems to take yet another 10x leap if you want to get there first. So now you're starting to get close to a billion dollars and so on. The thing that's so interesting is like in all things that have to do with technological invention, once smart people see it as something is possible, then it becomes so much easier for the second and third and fourth and 10th people to do it. And what we're seeing in the industry is that not being first means that you can also... Of course, you don't benefit from being first, but you benefit from being able to be smarter and a hell of a lot cheaper to get to the same state.

Not that getting to a GPT-4 class base model is cheap. It's still quite expensive, but it's a lot cheaper than it took for the very first innovators to get there. This is the pattern of tech that you and I have lived through our entire careers, so we shouldn't be surprised by this, but just to see it playing out with just this stunning speed is just incredible. And business-wise, the question is there's always been a big premium in our business in being first. Now the question is, given the rapid pace of change and evolution, is there still the same amount of volume being first as there has always been.

Speaker 1: The idea that we shouldn't be surprised, but we consistently are surprised. I remember the first time that I saw a chatbot for lack of better words, and now I'm looking at my phone.

I've got an app called MLC Chat, and I've got 535 Mini, it's a four-bit quantized F16, but I'm having a conversation in airplane mode on a 15 tops iPhone. That seems insane to me, and I get it and I understand it, but we're at a point now where the full stack, you always hear about the full stack engineer, is so deep that I find myself going back to building things from original parts. Just so I can remind myself I've got an Apple 1 that I'm building from 7,400 series chips just to get my hands back in the silicon because the stack is so high. I'm trying to get my hands dirty so that I can emotionally accept that there's an SLM on my phone that works in airplane mode.

Speaker 2: By the way, I'll take one of those Apple 1s if you make two of them.

Speaker 1: Yeah, those are great. Yeah, it's a company called SmartyKit. I'll send that to you.

Speaker 2: There's something that you're saying there though that is another reflection just looking back at my career and I think you and I, when we started in all of this, we were able to wrap our heads around and know the intricate details of the entire stack end-to-end. From the silicon or in my case, the wire wraps all the way to all the code, every line of code in the applications. And things are just so much more complicated now. And when we think about what AI is going to enable, I think AI is very quickly going to enable us to construct systems that are so complex that they will really completely defy any ability for humans to comprehend them fully. That in research where I think this might happen first, right now there's a vibrant set of researchers that are using generative AI to write mathematical proofs.

And when you ask an AI system to write a mathematical proof, you, generally speaking, ask it to write it in a proof language. A popular one's called Lean. There are others as well. The interesting thing about a proof language is that they're set up so that you can use a simple type checker, just like you would have in any programming language. And these proof languages are set up so that if they type check you know for certain that the proof is valid. And so I foresee in much less than five years that we'll have an AI system generate a proof of some mathematical theorem. We'll be able to type check that proof to know that that proof is absolutely valid and correct, but that proof itself might defy any ability for any human being, even the world's smartest human beings to understand it.

And that's sort of like a point example of AI getting us to a place where we're able to construct things, build things that work. And we know we can see them work, we can validate that they work, but we won't know how or why. And I think it's going to be... When people ask me about a GI super intelligence, that will be the first mark of it for me. And your roots and my roots, we still want to understand everything and we do things to keep that fresh for us, but I think it's going to get hard.

Speaker 1: ACM ByteCast is available on Apple Podcasts, Google Podcasts, Podbean, Spotify, Stitcher, and TuneIn. If you're enjoying this episode, please do subscribe and leave us a review on your favorite platform.

That's a really interesting point. I think you're right that I and perhaps people of our generation plus or minus a few years are still unwilling to let go of the fact that we took a rock, we flattened it, we infused it with lightning, and now it talks to us. And I want to understand the lightning and the squishy rock part, but my children, my 19-year-old is perfectly willing to just accept that the magic black box is doing a thing. And when I

have to have conversations with him about maybe let's not anthropomorphize the AI, let's talk about what's really happening. He's like, "Nah, it helped me with my homework and I'm cool with that." And I was a little bit taken by your book, The AI Revolution in Medicine that was kind pre-GPT-4 general availability. You were kind of poking at the model in ways where you're like you were anthropomorphizing it, but you were also trying to understand it. What would you change about how you interacted with the model now that you know more since this very good book has been published?

Speaker 2: Yeah, that book, we wrote it over the Christmas holidays in 2022 while GPT-4 was still a secret project. And we organized ourselves so that the book would get published at the same month that GPT-4 was released to the whole world. And so that was a time when we were just so amazed and baffled. I talk about this as the nine stages of grief. In fact, I think at some point in the book I talk about the nine stages of AI grief. When GPT-4 was first exposed to me by the folks at OpenAI, I was intensely skeptical because they were just claiming that this thing could do things that I just thought were not possible. And then you pass from that stage of skepticism to a stage of frustration because I felt like I was seeing my colleagues in Microsoft Research falling, getting duped by this thing.

And then you start to feel worried because I detected that, wow, Microsoft might actually make a big bet on this thing. But then you get hands on and you start to encounter things that are just amazing. And I remember feeling the joy that, wow, this thing is... I never thought I would live long enough to see such a technology, let alone have my hands on it. And then you get into a period of intensity. So there are these stages that you go through, but in those early stages of joy and euphoria and you lose sleep, it is those aspects that feel empathetic, that draw you into anthropomorphism that end up being so interesting. And in the field of medicine, this has been observed over and over again. In fact, one month after we published our book, UC, San Diego and Stanford jointly published a paper in a medical journal where they used GPT-4 to respond to emails from patients.

And they compared those to the emails that doctors, human doctors would write. And then they had a blind test and they had patients and doctors grade the quality correctness of these responses to patient queries. Not only was the AI equally accurate, but by a factor of nine to one, the AI-generated notes were judged by patients to be more empathetic. And of course, it seems crazy to say a machine can be empathetic. What it really means, I think, is that a frazzled doctor can't take the time to write more than two or three sentences and just get to the point and then get onto the next email. Whereas the AI can write a couple of paragraphs and might remember that during the encounter they were talking about going to a Seahawks game and other stuff like that and put in those nice personal touches. And so there's something there that is both interesting and disturbing, but also seems to really touch people in a very meaningful and very practical way. And I think we're still just as a society trying to come to grips with this.

Speaker 1: Yeah, I'm struck that one could theoretically say that it can be infinitely empathetic and infinitely patient given appropriate prompting and an appropriate good attitude on the part of the controlling human. I keep coming back to empathy. We need more empathy in the world right now. We need more empathy in tech. When I use on GitHub Copilot, I don't ask it how to do my homework. I basically use it as an enthusiastic pair programming partner, and I find it to be infinitely patient. It never judges me. It's never mean, it's never unkind. So then it has theoretically infinite empathy if I talk to it right. I'm struck by the epilogue in your book. You say, "It just can't be that next word

prediction could be intelligence or can it?" Am I just a statistical model of what's the most likely next word from Hansel want to say? Is that the animus of me?

Speaker 2: So I think at the time we wrote that book, I was really baffled by this, but I think my understanding and acceptance of what's going on has evolved a lot more. It's true that the fundamental pre-training of these large language models is to predict the next word in the conversation. And so to that extent, you could say that these large language models aren't trained to do anything useful except that. But here's an example that I like to give. Let's take the sentence and the murderer is blank. Okay, so now if you want to pick with the highest quality, the word that would fill in the blank, well, that sentence and the murder is blank is in the context of let's say a whole Agatha Christie murder mystery novel. And if you were to just approach this purely a statistics question, well, there are lots of thousands of murder mystery novels and short stories, and you could try to make a statistical pattern on what are the most likely names of murders, and you wouldn't get any good answer there.

Instead, in order to really optimize the quality of the fill in the blank capability, you somehow have to be able to do some deductive reasoning. You have to have an understanding of the psychology of humans in different situations, what motivates them, how they react under certain kinds of questioning and all of those things. And so the way to understand what's going on when we try to optimize fill in the blank or next word prediction, is a very, very large astronomically large stochastic process that has a chance of accidentally discovering neural circuitry that implements some aspects of those reasoning functions. And the fact that that can happen at all, even by accident is amazing. But we're operating a scale where indeed it is actually happening. And so it's not that it's next word prediction that is causing us to appear to be thinking, but the process of highly optimizing an ability to do very, very good next word prediction is giving us a chance to really discover and solidify these bits of neural circuitry to do things.

Speaker 1: Yeah. One analogy that I've used to explain to some young people, and I don't know if it's a good one, is that we are as humans limited by the size of our stack for our recency. And then the older you get, you get this larger and larger, larger heap that you can pull from, and then you're constantly pulling things out of the heap and into the stack. And then there's certain presenters, there's certain thinkers, Jamelle Bouie, is one I think of who's an opinion person for the New York Times, who seems to have this huge corpus of information that he has pulling upon all the books that he's ever read. He's had an amazing vocabulary, and I admire him and his ability to page in and out these pieces of wisdom while I struggle to root around in the totality of my existence. And I think that AI will feel like AGI when its stack, when its context window is beyond those of even the smartest person. And that's going to be the thing that it has just such a large context window. It's bigger than a human lifetime.

Speaker 2: Yeah, yeah. Up till now, for the most part, large language models and transformers specifically, were pretty imperfect in memorization of things. There's a massive compression of the training corpus that goes on when you train it into a transformer. And so I was always trying to explain that to doctors as they're trying to come to grips with generative AI because unlike a normal computer, in fact, one popular application, maybe the most popular application that doctors use online is something called UpToDate. It's essentially a search engine for a very highly curated medical knowledge. And so in

UpToDate, you ask a question and you get a medically precise answer. And we've trained ourselves to use that just like we use web search.

You make a query and you expect to get a set of answers that are pretty precise, but the transformer doesn't have that capability. It has very, very imperfect memory. And of course that's now evolving because there's no fundamental reason why computer-based system couldn't actually have perfect memory recall. And so I agree with you, I think we're going to get to a point where these AI systems really are going to be benefiting from the fundamental capabilities of perfect memory recall that we've always assumed machines would have, while also making all of these associations and engaging in this kind of reasoning.

Speaker 1: And forgive my ignorance if this is a question I should know the answer to, but when did your deep interest in the medical aspect of technology in everything that you do... Recently of late, you always want to make technology to make people's lives better. And it seems to come back to healthcare. Do you have a background in healthcare I'm not aware of?

Speaker 2: It's really been accidental, and it almost didn't happen. When I joined Microsoft in 2010, it was to join this great organization called Microsoft Research, and I was very proud to be part of Microsoft Research, and I rose the ranks to the point where I was the leader of Microsoft research worldwide. And then in 2016, Satya Nadella and the CTO at that time, Harry Shum reassigned me, took me out of research. And asked me to take on this 100 person team, a Skunk Works team to rethink Microsoft's approach to healthcare and healthcare technology. I was devastated by that reassignment. I thought in fact, I was being punished for some reason and actually contemplated quitting. Not only did I not have any background in healthcare and medicine, but... Microsoft, if you go to any healthcare organization, any clinic around the world, you will see Microsoft products there.

We sell to literally every single healthcare organization on the planet. I think one of our smallest accounts is a one-nurse clinic in Nairobi, Kenya, and then all the way to giants like United Health Group or Kaiser Permanente and everything in between. And so that meant there had to be maybe a dozen powerful corporate vice presidents all throughout Microsoft, all doing their own things in healthcare. And so I also had to think, who's going to listen to Peter Lee on anything? And so that's how it started. And you had to kind of think, what would we do? One question, Satya was worried that we weren't thinking enough about the cloud and AI in healthcare. So our first thing was to try to think, well, can the cloud be used to store healthcare records? And we learned early on that, no, that there were compliance faults. There are certain data standards that we weren't supporting.

And not only could our cloud not do it, but our competitors clouds that Google and Amazon couldn't do it either. And so that at least got us started with something to do, something to fix. I mean, we made a lot of progress in that. We also started a second project in collaboration with a company called Nuance and a doctor at University of Pittsburgh Medical Center, Shiv Rao and that project was called Empower MD. Because we learned that doctors were really suffering with having to write clinical notes to enter into an electronic health record system after every conversation with a patient. And so we thought, well, we could use AI to listen to the conversation and then at least draft a

clinical note automatically. And so that was a project called Empower MD, and that got serious enough that we actually went ahead and decided to productize that and acquire Nuance in the process.

And for Shiv Rao, Dr. Shiv Rao at UPMC, the venture arm of UPMC, UPMC Enterprises, agreed to provide seed funding for Shiv to spin off a company to do the same thing. And that's a company called Abridge. And today the top two products in that space are Microsoft's DAX Copilot and a product called Abridge. So that's how we kind of got started. And so in that process of about five years of working on that, I got up to speed a lot. I also became a founding board member for a new medical school, the Kaiser Permanente School of Medicine and a new school of medicine has to come up with a curriculum. And so I was able to study the curriculum and at least in pre-clinical studies, learn quite a bit, all the way to the point that I actually got elected to the National Academy of Medicine.

And then finally, Kevin Scott, our CTO in 2020 hired me back into research. And so I thought, okay, I can separate from healthcare, get back to what my true love, which is fundamental research in computer science. And then the pandemic hit. So then Microsoft decided, "Well, Peter, you're our healthcare technology guy, and all of our customers and stakeholders need help from Microsoft to cope with the pandemic, so you need to coordinate that." So it kept me in the healthcare. I thought that that project will last a summer of 2020, but of course the pandemic ended up being much more serious thing than that. And then when GPT-3.5 and GPT-4 came out, there was again the question, wait a minute, is this stuff good to use? Is it safe to use in healthcare and medicine? And I was logical person to help lead that. So I guess you could say I've been trying hard not to be in healthcare, but I keep getting pulled back in and I'm not unhappy about that. But it's odd 'cause none of it has been planned.

Speaker 1: Some of the best careers are not planned, and I think that's a testament to that.

Speaker 2: I'm glad I didn't leave Microsoft. That would've been a lot-

Speaker 1: I think we are all glad you didn't leave. That would've been... If you felt you were being punished and quit immediately, we would've definitely lost out. As we get ready to close-

Speaker 2: I know this here because Satya Nadella is such a remarkable leader, but I think what people on the outside sometimes don't know is he sometimes does ask people to do very, very hard things and it really moves things forward.

Speaker 1: That is true. When you're called to serve though, you either step up or you don't. And it's good that you stepped up. I was going to say that as we get ready to close, here we are kind of in the middle of March. What are the things that are coming out of Microsoft Research that we should learn about as we close this podcast? What are some things that have either been announced or are being announced here at the middle of March?

Speaker 2: Right. We've been doing so much particularly in two areas. One is what we would call AI for science. So in the same way that we've discovered that generative AI architectures like transformers and diffusion based models seem to learn so effectively from our words and thoughts and actions. So you can take a big corpus of human text output or word

output, and it's amazing what is learned there. And similarly, from pictures of what we do out in the world and videos, those same architectures, what the world is learning and is a subject of huge intensity in Microsoft Research also work for observations of natural phenomena. Like atmospheric wind patterns or the dynamics of proteins and small molecules or the movement of electrons in the electrolytic material structures. And that is really incredible.

That means that if we follow the same path of AI scale in those areas, we might be able to do things like predict severe weather events weeks in advance. Or design drug molecule for known drug targets or identify new drug targets in pathogens. Or be able to design new materials for everything from say, solid state batteries to enzymes to make your vegan food taste better. And so that kind of AI for science thing, I think is something that was a big showcase effort with a whole bunch of new models that are now in the Azure Foundry coming out of Microsoft Research while being simultaneously published in the top scientific journals in material science, in chemical engineering, in physics, in climate science and so on. The one big difference, the one thing that holds us back in all of that is access to training data.

There's no internet of molecular dynamics simulations. So the question is where do you get the training data? We have the compute infrastructure, but we need the training data. And on that, the second thing that's emerging is quantum computing. And I think in 2025, we are going to see the first practical scalable quantum machines. And the very first application that at least I and some of our colleagues in Microsoft want to see is to run classical precise simulations of these natural phenomena to generate large amounts of perfectly labeled training data. And if we can do that, then we can literally have the GPT-4 or GPT-5 of proteins, of materials, of weather patterns, and I think that'll be pretty stunning.

Speaker 1: Yeah. And then that research to practicality bridge it doesn't just become theoretical at that point. I love that you're putting out papers and simultaneously releasing a model on the Foundry. Because I go back, call back to the whole beginning of this conversation, I've been told that my diabetes will be cured in five years every year for the last 30 years. Show me the money. Show me the practical thing that's going to prevent someone from losing their home in a tornado or prevent someone from dying of glioblastoma. Those kinds of practical things. You're saying good things are coming from AI and from these models.

Speaker 2: The way I've tried to explain it is that across many, many scientific domains, we have achieved GPT-2 level of capability. And the only thing that's really preventing us to get to GPT-3 is access to adequate training corpus. And so the minute that we're able to solve those issues, we'll get to GPT-3 class capability and beyond. And then GPT-3 for us at Microsoft, for you and me, Scott, has been important because GPT-3 in large language models was the first stage where we could try to make a product out of this. And that of course was the first GitHub Copilot.

Speaker 1: Yeah, that's the beginning of the hockey stick. When it starts to curve, then things start happening. Well, thank you so much Dr. Peter Lee for chatting with us today. We really appreciate it.

Speaker 2: Well, Scott, thanks for having me here. It was really fun to chat with you.

Speaker 1: We have been chatting with Dr. Peter Lee, the President of Microsoft Research, and this has been another episode of Hanselminutes in association with the ACM ByteCast, and we'll see you again next week. ACM ByteCast is a production of the Association for Computing Machinery's Practitioner Board. To learn more about ACM and its activities, visit acm.org. For more information about this and other episodes, please do visit our website at learning.acm.org/bytecast. That's B-Y-T-E-C-A-S-T. Learning.ac.org/bytecast.