accenture

# AI LEADERS PODCAST WHAT TO CONSIDER WHEN IMPLEMENTING GENERATIVE AI

## AUDIO TRANSCRIPT

**Bratin Saha** [00:00:00] It is real. It's here, and it's going to transform pretty much everything we do.

**Teresa Tung** [00:00:10] Welcome to the Accenture Leaders Podcast. My name is Teresa Tung and I'm Accenture's Cloud first chief technologist, where I look at innovations powered by the Cloud. I'll be your host for the session on getting started with generative AI, and I'm thrilled to be joined by Bratin Saha, Vice president and general manager of AI Machine Learning at IWC. Thank you, Bratin, for joining us.

**Bratin Saha** [00:00:32] Thank you for having me. Really nice to be here.

**Teresa Tung** [00:00:35] So generative AI, it seems almost magical. Many people have experienced ChatGPT and how smart it seems in answering common knowledge questions. I want to start with how does it do this? What's different about how Gen AI works that allows for the seemingly massive leap?

**Bratin Saha** [00:00:53] Yeah, I think there's really three factors that contribute to how Gen AI has moved forward and how AI moved forward because it's really just a continuation of, you know, the AI and machine learning progress that we have seen over the last few years.

The first of these is a massive increase in the size of the machine learning models that you can train. And to give you some statistics, back in 2017, you know, the state-of-the-art model then was resona you know, I remember that, you know, everyone used to look at that and, you know, it would be able to classify images and all that. And it seemed kind of wondrous at that time, but that it was to have about maybe 17, 20 million parameters and then board gaming 2019 and that was a giant leap that was actually kind of the you could call it the first kind of foundational model in the sense that the transformer architecture team there are you can call that the predecessor of these foundation models and that had about 330 million or so parameters. Since then, though now these models kind of hundreds of billions of parameters. So, you have seen like a super exponential increase. You've seen more than a 1500 X increase in the size of these models in just the past three years. And that has allowed us to pack more capabilities into these models. Now, when you have these larger models, you also need to train them with a large amount of data. And that is that has been the second aspect. And you can scrape, you can, you know, get a lot of data from the internet. And so, the second thing that has happened is just being able to get a lot of this data to train it. And then when you have large models and large amount of data and you need to train

them and these training runs can go on for weeks or months,, you need a huge amount of compute and you need that compute to be available at a reasonable price and if possible, in an on demand fashion, because any are going to create these models for a few weeks and then you're going to run some, you know, smaller experiments and so on. So, you just don't want to have to have the whole thing from, you know, you don't want to have to just factor in all of that expense. And so that is where the Cloud plays such an important role because it makes the data and the compute available for companies to innovate on. So, I think if you look at the core things, it's been the massive increase in the amount of models, the massive increase in the amount of data that you can process and the massive increase in the amount of compute that you can throw at this. In fact, there's a study, but, you know, the whole thing has really also been accelerated by the Cloud, because if you didn't have the Cloud, it would be very hard to make all of this available at scale to so many people because, you know, the key thing, if you look at the way machine learning has progressed over the last few years is that every company is kind of, you know, as strong as this. You know, at Amazon, we have done a lot, but other companies have also done a lot and if you look at the rate at which these models are also coming out in the open source and so on, so the Cloud and AI has played a huge role in kind of democratizing access to compute the data and the models that really sparked this kind of innovation. And to give you some really interesting data, if you think about Moore's Law, you know, Moore's Law drove data revolution and that doubled compute capacity every 18 months. Now, if you look at over the last five years, you know, maybe 2016 ish timeframe, when the Deep Learning era kind of started at that time, the total amount of system compute and amount of compute that we are throwing at the machine learning problem that has increased, that is doubled every three and half months. You know, so you can imagine that the progress has been so much faster because, you know, the thing has been moving so much faster.

**Teresa Tung** [00:05:00] Yeah, makes sense. So, without the compute, we could not process even the data at the scale create these really large models and then to even serve them back to us. So that that's really an enabler. As you mentioned, it's more of an evolution than a revolution. Your journey starts even decades back, right? It's not a surprise to you, but might be surprise to mainstream.

**Bratin Saha** [00:05:24] Yeah. And, you know, and it's that vast amount of compute that needs to be available and the ability to make it available in an on-demand manner. Like, you know, I need hundreds of accelerators, I need maybe hundreds and thousands of GPUs and accelerators, but making it available only for the duration, I need it that makes it a lot more economical than you know you have to just keep getting it because the technology is also moving forward every year. And so, you know, that actually helps you be on the leading edge of the compute as well. So, you know, this compute explosion, the data explosion, the explosion in the complex sophistication of the models, all of that is followed by the cloud, by, you know, AWS and others is really what has driven this.

**Teresa Tung** [00:06:15] So the good news is, because it is powered by Cloud and powered by data, those are two things that a lot of our customers have access to. You don't have to be a pure tech company to be able to tap into this, right? Our customers, many companies have unique data, right? Unique domain knowledge. And with the Cloud, like you said, it's on demand. So, one of the biggest changes is that, you know, Gen AI really unlock the imagination of what AI can do. It's so disruptive and will change how everyone works. It's so disruptive and will change how everyone works. Or we we've found research that among business leaders, 98% of respondents found that AI

Foundation models are going to play an important role in their organization's overall strategies, right? Not just tech strategies, but business strategies in the next 3 to 5 years. So, with that in mind, run, how do companies get started with Gen AI.

**Bratin Saha** [00:07:15] You know, I think the first thing is you still want your data and compute infrastructure. And, you know, usually what we have seen with customers is they get a lot of benefits from using a Cloud based infrastructure. So, I think that still is needed. Then you want to be able to use one of these foundation models. Now you can build them yourself if you want, but a lot of these foundation models are pretreatment already available, so you can save yourself a lot of effort by just using what is available. You know, most of the Cloud providers provide this at Amazon, we have Bedrock that provides this Amazon Bedrock that you can use pre-trained foundation models. Um, and from a variety of companies. So, you know, you want to be able to look at 40 all particular use cases we can works best. And then we also have a long chain of models that we support on save maker. Then once you have decided on, you know, which models you want to do, you then want to be able to figure out some use cases where you may want to deploy this. And you know, this could be like a search related to in a chapter in that thing, you know, supporting, you know, we have seen customers do like searching within documentation. So, I don't really have to browse documentation. I can just ask the question. It can be searching the enterprise content. But, you know, as you said, there's a plethora of use cases to do that. And so, then I think what you need to do is figure out what is the new skills that you're going to experiment with. And there will be an experimentation phase here because these models also have, you know, issues they had to submission, you know, those kind of issues. So, you then need to take and do some of these experiments. And at that point of time, you may have to actually do a little bit of, you know, a little bit of putting the system together. So, you know, that might be that the model just wants plenty of to share, which, you know, some sometimes consider short or it may be that, you know, you just need to give a little bit of contextual information in a few short manner or it could be that you also want to use other approaches like, you know, a retrieval of generation, like a and so on. So, you need to be, you know, so use you first need to figure out that, you know, you have the right machine learning infrastructure to do stuff that you want to pick. You know, most cases you will probably want to use a pre-trained model if for whatever reason, you know, you want to use your own model or build your own model or whatever, then you know you can do that as well. There's a lot of models available then getting to what is the use case. And you know, you probably want to start the use case that is not overly complex to begin with. So, you can get your hands dirty with what is how it actually deployed in production, how to actually guard against hallucinations and so on. And then once you've done that, then you need to get the POC going. And then once you have had a POC going, then you get to the point where you say, okay, this is robust enough and I don't want to be added to production. So, it's going to be a journey and you know, the basics of how you it's not different than the basics of how you basically want to deploy machine learning in general.

**Teresa Tung** [00:10:38] It's almost like when Cloud was being born, right when cloud was being born. And now you can buy it as a SAS, as a PAS, as the App IAS, right? Like part of that is based off of how much control you want to have and customizability you want to have. And it sounds like we're really seeing that play out in the AI space. So, one is I could buy something that is that I use, right something like Amazon Cold Whisper, right. I could use that for software development and just really augment and turbo boost my developers, you know, and they might not even know underneath it's Gen AI powering that and then you're saying sometimes we might need to augment and have more customizability using something like Bedrock gives us some of that capability and

you still myself so clients who might even need to go all the way to you know building from ground up and having that sort of choice kind of like we now have with Cloud, it's almost the same sort of choices that we have to make.

**Bratin Saha** [00:11:42] Yeah, and I was going to get to that as well, is that in some other ways, you know, if the easiest thing to integrate Gen AI into your enterprise workflows, maybe to start with the apps because they're, you know, none of this model building and you know, POC and all of that is significantly reduced because you know, the app builder in this case, let's say they're using code whisperer and software development is a really good place to use this. In this case, AWS has really done all of the heavy lifting of making sure that the modern quality is code, making sure that the code quality being generated is good and so on. So, you know, and we expect to have a lot more apps coming out. And in many cases, that might be the best way for you to be leveraging Gen AI in your enterprise workflows and then seeing what you do on top of that.

**Teresa Tung** [00:12:40] Well, you did touch on code quality just out. And so, we do hear a lot of concerns with Gen AI, right? So, we hear about Gen AI being used to answer nefarious questions or when it's revealed that some of the models are built by data that might be biased or have questionable IP rights contained. There's real enterprise concerns around IP and data leakage when users are submitting their data to a third-party model. So as a leader in this space, can you share how Amazon is tackling some of these security and responsibility considerations?

**Bratin Saha** [00:13:14] This is top of mind for us. You know, making sure that these models are being used in a responsible way is top of mind for us. And we have you know; we have been doing machine learning in the AI for more than two decades. And we have you know, we have that's part of our fabric, that's part of our legacy. And so, we have very robust processes to make sure this is done well. And I used the Code Whisperer as an example. So, you know, we have filters and there is a model being used in a particular app. We have filters that actually filter out inappropriate content. And in the case of Code Whisperer, but we also have a way of making sure that your code has the right attribution. So, if you are in an enterprise setting, you know, you don't want a situation where the code is being generated and the right copyright messages are not there. And Code Whisperer is unique in that it actually did the effort of making sure we provide the right attribution. So, you can use it much more easily. And then, you know, we do a lot of testing on the data that we use, making sure that, you know, it does not have harmful content, does not have toxic content. We came out with service cards last year for some of our services that go into much greater depth on what are the responsible parameters that we use in our services and it deals with bias and explainability and robustness and security and all of that. And you know, we are using the same processes as well. And we have we have what we call a quality or a set of quality checks that we do, just like before we release the models on our AI services, just like we would do for operational function of things.

**Teresa Tung** [00:15:03] So it sounds like even if you're using a model, you should definitely at the app level, you definitely need to understand the risks involved in making sure that that provider has done like what Amazon has done right and ensuring that the data that's being used and the protections are built into the app. Second, if you're going to take on some of that responsibility on your own building your own models or using something, maybe pre-trained from open source and adding your data regardless, you need to take that consideration to say, I'm going to take on the responsibility to make sure the data that I'm using is safe and is going to be propagated proper way. So.

**Bratin Saha** [00:15:44] Yes. Like when we when Amazon is, you know, producing the apps

that it's Code Whisperer and whatever else we come up with down the line. In that case, you know, we do all that hard work and heavy lifting. And so that is one way we make it easy for customers to use it. But you're right, if someone else wants to build the whole thing from scratch, then they have to take care of all this.

**Teresa Tung** [00:16:05] Well, Gen AI, you mentioned it's an it's an evolution. And so, this evolution is impacting most companies' data and organizations and architecture. So, introducing new Gen AI models, starting with the data foundation. And that's more important than ever, being able to get your data with the quality and being able to process it at scale. There's going to be more domain experts in addition to data users, right? Being able to tune the models and to use the models. So, what are some of the changes that generative AI brings to your existing data and organizations and architectures?

**Bratin Saha** [00:16:42] I think that are some of the things that carry over from just being machine learning. But then there are some other new things that are out there as well. I mean, ultimately you need an industry of scale, machine learning infrastructure, you know, because you would have to be depending on where you are in the stack for using apps, then no, you don't need it. If you are not using apps and you're building and using a pre-trained model, then the amount of infrastructure you need is less like you're building your own model, fine tuning it and so on. You need it more. But then there are also other things like, you know, prompt engineering in terms of being able to guide the model to the right place. You'll probably be using new kinds of infrastructure in many applications. Things like vector databases are useful, and you'd probably be looking at new techniques like, you know, retrieval, augmented generation. In many cases, you want to depending on how much more routinely you're doing, the human need to align the model to getting the right output. And that gives human feedback becomes important relative, you know, reinforcement learning and human feedback that becomes important.

So, it really depends on how you're using the stuff. If using the apps, then, you know, it's less software than as you go lower down. There's more of it. If you're going lower down, then, you know, these models are so big that that a new training techniques and new inferencing techniques you have to use to get the cost and performance and the profile.

**Teresa Tung [**00:18:15] I think infrastructure you started talking about Moore's Law as well, right? Infrastructure is going to make a difference. Maybe it's not seen to your end user, but certainly to be able to train these large models sustainably at scale, at speed.

**Bratin Saha** [00:18:30] Yes, very much so. And that is why at Amazon, we have we have come out with a custom hardware for training and inference of these large models. So, we available to stream the listen for it and share which can be used for training and inference of these models. You get, you know, more than 40% improvement in price performance or other comparable instances. And just as you said, that is that it's going to be really important because cost matters ultimately, if generative AI is going to become pervasive and we believe it's going to become pervasive. Then, you know, you want you know, you want to hit a particular cost profile, a particular performance provide to make it widely available and usable. And so, I think that is really an important part of that equation.

**Teresa Tung** [00:19:19] So any final recommendations? What should companies know about generative AI?

Bratin Saha [00:19:25] So the first thing is, you know, it is it is real, it's here and it's going to transform pretty much everything we do. You know, you mentioned 40% of the of the working hours will be used by that. You know, one analogy that I sometimes used is, you know, if you go back to the Internet revolution now,

if you look at like maybe 80%, 90% or so of odd working hours is actually influenced by the Internet. But it's, you know, this communication, we have email, slack meetings, you know, the documents we are writing that are stored in the Cloud and so on. So, this is going to be similarly helpful and will enhance our productivity. So, the first thing is, you know, it's really important to get started and then figure out a way to get started, could be adapted or you could kind of go deeper and build your own stuff. Get started on the POCs, you know, there's a lot of innovation happening out there, a lot of innovation happening in open source. And so just, you know, making sure you have a Gen AI based strategy, like you're thinking, how can I enhance my customers experience with and how can I enhance my employee's productivity with that? So having that kind of an activity and then just executing on some of these projects.

**Teresa Tung** [00:20:42] Completely agree. The new normal is being invented right now in every industry will be disrupted. So, I really appreciate that. I think really the biggest promise is actually for a lot of companies who own the data and play a critical role in reinventing their industries. So, this is not just a technology play, but with that in mind, Amazon Technologies and you guys are going to handle the hard stuff for many of us. And together we're going to look forward to what's to come.

**Bratin Saha** [00:21:14] Think Amazon Bedrock would be a great place to get started. If you want to use these models that are others as well. So, you know, people should do their own evaluation. What kind of enterprise readiness would be an added thing? Code Whisperer as you mentioned, we have done studies where people have been able to be to finish tasks almost 50% faster and they have been almost 30% more productive. Code Whisperer would be a great place to get started in the software world. But, you know, there will be other stuff coming as well. So, I think there's just a lot going to be happening in this space from us and from others.

**Teresa Tung** [00:21:52] Okay. Bottom line is to just get started. Thank you, Bratin.  Thank you for joining us.

**Bratin Saha** [00:21:56] Thank you. So.