Guests:

- Aparna Dhinakaran, Co-Founder and Chief Product Officer, Arize AI
- Prag Sharma, Global Head of AI Centre of Excellence, Citi

Transcript:

Jorian Murray ([00:01](#)):

Hello, I'm Jorian Murray and welcome to Good Things Happen, the show that invites change makers and enablers to share their inspiring stories of progress. Whilst change can be uncomfortable, unexpected, and at times disruptive, it's inevitable, and more often than not, change is for good. We'll be hearing from people from all walks of life who have been at the forefront of change, including their journeys to get there and their motivations. Because when people work together [00:00:30] for a common cause, good things happen.

Prag Sharma ([00:35](#)):

Artificial intelligence has become the hottest topic on the planet.

Aparna Dhinakaran ([00:39](#)):

I think for those of us who are on the ground, there's amazing excitement at what's being built.

Prag Sharma ([00:46](#)):

2023 was the year of the wow, okay? It was, wow, look at this tech. Look what it can do. This is amazing. But 2024 is year of the how.

Jorian Murray ([00:58](#)):

Today we have a very special episode [00:01:00] lined up for you as we dive into the realms of artificial intelligence. In this episode, we're joined by two industry experts who will shed light on what's happening with AI in 2024 and beyond. We'll be delving into the latest advancements, emerging trends, and the transformative impact AI is having across various industries. But what exactly is driving this rapid growth and what are the ethical considerations we need to keep in mind as AI continues to evolve?

([01:30](#)):

[00:01:30] Our guest today are Co-Founder and Chief Product Officer at Arize AI, Aparna Dhinakaran, and Prag Sharma, who is the Global Head of Artificial Intelligence Centre of Excellence. They bring a wealth of knowledge and experience from the front lines of AI research and development, and we're thrilled to have them join us for this conversation. Welcome, Aparna. So lovely to have you with us. Before we hear about you and what you do at Arize, [00:02:00] please tell us a little bit about your early life and development. How does one prepare for an industry that doesn't yet exist?

Aparna Dhinakaran ([02:08](#)):

Awesome. Hey, thank you so much for having me. My name's Aparna and super excited to be on this podcast today. Myself, I come from a technical background. So prior to starting Arize and to founding a Arize and leading up product here, I was an engineer for many years. Studied [00:02:30] electrical engineering and computer science at Berkeley, and I think a lot of just getting hands-on technical experience, building things and feeling the pain of deploying AI

models into production is what got me to go off and found Arize. So, being an engineer is a big part of my early background.

Jorian Murray (02:56):

Lovely. Same question to you, Prag. [00:03:00] Tell us about your formative years and what brought you to the role that you are now doing.

Prag Sharma (03:04):

Yeah, thanks a million for having me on the podcast. It's really me and not a deep faith. I think we should all sort of start saying that from now on as this technology gets better at really you have the opportunity to be at more places, more than one place at the same time. But delighted to be here. So, like Aparna, I have an engineering background. Why? Because it's a solid career choice and a course to do in college and it kind of sets you up for [00:03:30] many, many different things you could do down the line. I enjoyed my college life, so I wanted to stay four more years in college and one good way of doing that was to do a PhD. So I decided to do a PhD in computer vision. And computer vision is, as you know, a subset of artificial intelligence. And as we all know, artificial intelligence has become the hottest topic on the planet.

(03:54):

So I'm delighted that people are focusing on the stuff that I was accidentally interested [00:04:00] in back then, but the space has come a long way, delighted to be a part of this, enjoying my role here at Citi and really enjoying the world of technology and AI.

Jorian Murray (04:10):

Aparna, you said you studied engineering. Did you learn theories and processes and approaches back then that are universally adaptable or are you always educating yourself?

Aparna Dhinakaran (04:20):

I think it's definitely the latter, the always educating yourself. I mean the great part about engineering background and foundation is that it's not just [00:04:30] the things you learn once and then you use them over and over again. It's teaching you how to learn and a lot of engineering actually is getting hands-on experience and so building and myself. So my specialisation was in AI, and so learning back then what was the foundation of a lot of deep learning or transformer based approach, architecture type models. And then now [00:05:00] seeing all of that actually continuing to learn actually about all of the generative model architecture, what's happening as these new and new more foundational models come out, I think a lot of that foundation helps you build upon it and you're consistently learning. I think even for us in the AI space, folks who are deep in it this last year has felt insanely fast. The pace has been incredible.

(05:29):

And so some of [00:05:30] it is sometimes you just have to go out and read your research papers on your own. You have to go try this technology on your own. As you're trying it, you might actually be contributing back to the research we ourselves are publishing what we're seeing are users do and where the limitations of these foundational models are. So I definitely think it's never a one and done here. It's a continuous learning and it's part of the fun.

Jorian Murray (05:55):

Prag, what is Citi's Centre of Excellence? Explain how it [00:06:00] was designed and what is designed to do.

Prag Sharma (06:02):

The focus of the AI Centre of Excellence is to ensure that we use AI responsibly across the entirety of the organisation. And as you know, that word responsible has many dimensions to it. It's not just about ensuring that the models that you use, you can explain. It's not just about the data that goes into these models, it's also not just about the use cases. But it's about a combination of all of these things. Stepping back and understanding what [00:06:30] a typical use case of AI or more recently generative AI looks like, what impact will it have on our organisation but also on our clients, what impact it's likely to have more generally and more broadly, what governance and risk mechanism should be in place? What are our clients thinking about it? More importantly, what are the regulators thinking about it and what do our staff and people within the organisation think about it and the senior leadership who want to use it. So I sit [00:07:00] in the middle of this contract looking at how to use AI responsibly in Citi.

Jorian Murray (07:06):

Aparna, please tell our listeners what Arize does and why is it you are learning so much through what you do?

Aparna Dhinakaran (07:16):

Arize is an AI observability and evaluation startup. We've been around since 2020 and we work with a number of amazing top ML teams as well as Fortune five hundreds to help them make sure that the AI that's deployed actually works [00:07:30] in the real world. A recent example of this is you may have heard of an airline where they've actually deployed an LLM based chatbot to help customers ask questions about policies and really just ask questions about their product and services. And it actually made up a loyalty programme and the LLM model actually hallucinated. It made up answers that were not grounded in any truth. This happens [00:08:00] in LLM applications as you're deploying them into the real world. And what actually got ruled is that they had to honour that loyalty programme.

(08:08):

And so you have all sorts of these applications. And by the way, this isn't just in the LLM world happening now. This was happening even with traditional models. Prior to this, prior to founding Arize, I actually got into this because I was an ML engineer at Uber. We deployed many models into production. Everything from ranking to forecasting [00:08:30] and models unlike software are not deterministic, which means that as you deploy them into production, the results are statistical in nature. And we had a lot of experiences where the outputs weren't what we expected. We wouldn't find out about them until much later, and there just wasn't any tools to give us the visibility to catch those issues and troubleshoot it. That's what got me started on founding Arize. [00:09:00] And today we're the leader in this AI observability.

Jorian Murray (09:05):

Prag, just building on that from a global bank's perspective, AI must be equally exciting as well as it's daunting. It must be exciting about the opportunities, but daunting in terms of keeping to the responsibility that you talk about to. Bring to life some of those challenges for us, if you will.

Prag Sharma (09:25):

Yes, AI, like many technologies, is a double-edged sword. It has many, [00:09:30] many benefits, but it also comes with its associated risks, very much like Aparna just pointed out. And when you start to use these models in real-world applications, especially when you're interacting with the outside world or even when you're interacting with different businesses within your organisation, these issues or these internal capabilities or the associated risks become key drivers of how quickly you can adopt this technology. And that is not just true for regulated entities like banks, [00:10:00] but it is true for any use case. And that is why a lot of the new regulation that's coming in is looking at use case from a risk-based perspective. So you have low-risk use cases, high-risk use cases, and you even have prohibitive use cases in the European Union, for example. So for large organisations like ourselves, we certainly, certainly want to use this technology. Everybody's interested in seeing how these technologies can help them.

(10:28):

Having said that, the risks [00:10:30] have to be catered for as well. There are risks like hallucination that Aparna pointed out. There are cyber-related risks. And we just need to be comfortable in understanding what those risks are and in explaining away those risks where possible or accepting those risks where possible for a particular use case at hand. Because you can never explain away many of these things. And some of these things are inbuilt into today's transformer-based models, by the way. Bias is inbuilt because it is [00:11:00] inbuilt into the data. Hallucination is inbuilt today because the transformer predicts the next word in the sequence. So depending on what kind of probabilities you want for the next word, hallucination is there. So we have to live with that. So then the question is how do we live with that for the use case at hand, is it okay to accept that risk.

(11:19):

And balancing all of these things together? We look at deploying these models in production. And I can tell you today there are many, many use cases that we can use [00:11:30] these large language models for across the financial services and otherwise and manage the risk that is. But we certainly need to be mindful and cognizant of that.

Jorian Murray (11:40):

Aparna, every time I read an article, which isn't that often or hear a story about ai, the applications change and it broadens my mind about what the applications are, what the opportunity is. Give us a snapshot, from your perspective, and maybe take us [00:12:00] on a little journey because it feels like each few months it changes, it reforms.

Aparna Dhinakaran (12:06):

Oh, man, this is such an exciting time to be an AI. I mean you have the backdrop of over the last couple of years, language model is getting better and better. And what's happened basically since the last year is especially with GPT 3.5 and then GPT-4 releasing, is that you actually have a [00:12:30] large language model that is really, really good at human level reasoning. Not saying that there's no mistakes, not saying that there's not reasoning, but I mean you now have teams that are able to use an LLM to talk with their customers, to get them to make recommendations on behalf of the customer. They are so good at doing things [00:13:00] like being able to reason about why an error may have happened, what an error looks like even in code. And so you're now seeing every industry from the way you ask any engineer and engineers probably using GPT-4 to help them with their day-to-day coding life.

(13:18):

Now, I mean from engineering to you talk to folks not even in tech who are now using this technology to help them with their day-to-day tasks. It is [00:13:30] actually incredible. And what's even more exciting is now you have all these sorts of applications that can be built on top of that. The way that I think about it is, back when AWS launched. And AWS launched all sorts of different services from S3 and compute and all sorts of services, then were able to be built on top of that data platform. There is now a growing ecosystem of this AI platform that is being launched [00:14:00] as we're talking about. And there's going to be the companies of today and the organisations of today that are going to leverage these tools. But then there's also going to be new organisations that are going to be built off of now what's possible to do with this technology?

([14:16](#)):

And what Prag saying is right though is that there is an element of still not a hundred percent reliability with what these models are producing. And so I think that's the other [00:14:30] component that we're seeing shake out in 2024 is they're really, really good, but they're not perfect. They're not always accurate. And if you're going to deploy these types of systems into your business, you can't run 80%, 70%, okay, it's good enough. You then have to assume all the risks that can come with bad outputs. And so I think for those of us who are on the ground, there's amazing excitement at what's being built, but [00:15:00] we're also seeing these AI engineers have that caution when they're actually bringing that into production. And it's really easy to get a Twitter demo up today. You can get people to see what the application can do, but to get that to work reliably and consistent, I think that's kind of the journey that we're going to see unfold in 2024 as enterprises actually bring these into production.

Jorian Murray ([15:27](#)):

You both automatically talk [00:15:30] about large language models. I'm going to ask a very basic question Prag to the layman, why are large language models the heart of this, is the first question, and also how will reliability get better with the development of large language models?

Prag Sharma ([15:50](#)):

So we've all heard the term large language models, which is a sub-component of this term we've heard, which is generative AI, which is a sub-component of AI. And really [00:16:00] we are talking about large language models today because they have done something that was not possible only a couple of years ago. Which is analysing unstructured data, specifically language, in other words, text, which we consider unstructured, but it's not in tables, it's not numbers. So specifically analysing text to human-like levels. This was just not possible. So what is the large language model, a large language model is a model that helps us analyse unstructured [00:16:30] data, specifically a text to help us summarise, to help us extract key information from text, to help us translate text and to help us compare documents in any language. And this is why we are so excited about a large language models. More technically they're based on a newish architecture called a transformer model, which is based on what Aparna was earlier talking about, which is neural nets that we might [00:17:00] have studied in engineering from back in the day.

Jorian Murray ([17:03](#)):

It seems mercifully, Aparna, we are getting past this fear of AI and people are embracing more the creativity of it. Do you ever sit around and think what the social impact of this is going to be and how it's changing lives?

Aparna Dhinakaran (17:19):

Yeah, I mean I definitely think there's a camp that probably sits on the side of a lot more fear. Is this going to take away jobs? [00:17:30] Is this going to really reduce the number of people who can do what's being done today? And then there's probably a side that I sit closer to, which is does it actually empower those of to be more efficient? And then in doing that help enable create more opportunities for what could be out there. And I think what we're seeing is really more [00:18:00] of that ladder. We're seeing this technology is not at a point where it's actually replacing day-to-day individual, what folks are individually doing, but it's actually making things so much more efficient already. I was giving you the example of you ask any engineer today what they're doing and they're probably using a large language model to help [00:18:30] generate code, fix bugs in their code, but we're not actually seeing them take over any individual engineer's job.

(18:40):

And similarly, I think across other industries as well, it's going to help us become more efficient and then by doing so, help us create more opportunities. So I'm really excited and looking forward to that. There's another aspect we haven't talked about, which is up until now there's been a large focus on mostly language as interaction [00:19:00] modem. So you chat with it, you type, but with multimodal models coming out, if you all have seen the recent demo of Sora. I mean this is text to video, text to image. What's exciting about this is that now you're going to get more and more of during maybe more of that creative side that you were hinting at earlier, kind of start happening in the next... I'm excited about that in the next [00:19:30] half of this year is that all of these multimodal models and their applications I think start to become more mainstream.

Jorian Murray (19:38):

Same questions to you, Prag, look into the crystal ball for us, if you will. What excites you about the impact of AI on society going forward?

Prag Sharma (19:50):

Yeah, I think when it comes to AI, the crystal ball unfortunately can't see that far anymore. It can only see next month or two months from now. Looking into a year is [00:20:00] predicting across decades when it comes to AI because of the pace of change. But having said that, I expect a number of things to happen next year. As Aparna mentioned, multimodal models will become more common and more useful. Because inherently this means they can analyse more data and combine more data together from disparate sources to provide better outputs and results that are more useful to us. So that's one thing that's going to happen. We should also see that models may potentially [00:20:30] start to become smaller. We want smaller models so that we can fit them on our phones. We want smaller models so we can put them on IoT devices on the edges where they can analyse data.

(20:41):

So I'm hoping there'll be more development in that. I was at a presentation recently where a very famous scientist in the space was calling for more funding into AI safety. So I think this is going to continue to be a hot topic and grow as a topic in and of itself where more funding, looking [00:21:00] at AI safety at the kind of things Aparna and our team are doing in terms of observability and others that lend themselves to ensuring that these models are safer to use also in the long run will be another key focus area.

(21:14):

And then we are continue going to use 2024 as the year of the how. 2023 was the year of the wow. It was, wow, look at this tech, look what it can do. This is amazing. But 2024 is year of the how. We all need to get our hands dirty. We need to understand [00:21:30] how we can use them and what value this brings for us. Because it's not clear exactly how much value these models will bring for individual organisations because it depends on the number of things, it depends on the data the organisation has access to, it depends upon the senior management and the willingness of the people to adopt that technology, it also depends upon the hardware and infrastructure of the organisation to utilise these models. And finally, it depends upon the expertise that they can [00:22:00] hire, buy upscale, their own resources to really get the most out of this.

(22:05):

Looking further ahead is anyone's guess. These models will continue to get better. Robotics is going to get better. Combining these models, taking them out of our phone, taking them out of our laptop, putting them into devices that move around us potentially is going to be the next big wave. And then there are other areas where these models will continue to make their imprint [00:22:30] in the short term. I'm hoping these things continue to improve because I think we can certainly utilise them for many, many of the jobs that we do today.

Jorian Murray (22:39):

Aparna, please build on that year of the how. I guess you are uniquely placed to witness the challenges of this. I guess there's a gold rush going on, is there for all of the people that you're observing and analysing to come up with the next great application?

Aparna Dhinakaran (22:57):

Yeah, I mean, so what we're [00:23:00] starting to see is really maybe different groups of these applications. So there's first a group that started off last year. It's a little bit more customer-centric, call it customer-centric chatbots, et cetera. We're now starting to see another type of application, call it chat to purchase, where people are actually embedding these type of chatbots within their product and having them actually do recommendations to get them to purchase. If it's an e-commerce company, get them to purchase whatever they're selling, [00:23:30] whether it's a product or an experience. We're also seeing assistance. These can be both internal or external assistance. And these are actually really common because they're helping individuals within an organisation be more efficient, like I was mentioning. And then there's a number of other ones that are out there, but I think these are probably more agentic type of applications, but these are probably the more common ones that we see actually being deployed into production.

(23:59):

And [00:24:00] I think the biggest thing with these, there's two big ones, Jorian, which is, first, how do you connect to your own private data. That's important. So LLMs are trained on public data, but they might not have specific information about, let's say citi's. If you're building a chatbot about policies that you might expose to your own internal, let's say employees about how to respond to a customer if they ask about this or et cetera. Well, the LLM, if you're using a large foundational [00:24:30] model, is trained on some external large data set. So you first need a way to actually connect these LLMs to private data. That's really important. It's this process called RAG, retrieval augmented generation, where you're retrieving information and then using that in the actual generation of the responses.

(24:48):

This workflow has a lot of issues that can happen. You might not retrieve the right information. You might not have enough information about a specific topic to actually do the generation correctly. [00:25:00] But it's super important if you actually want to ground the LLM in some concrete facts so that it doesn't hallucinate. And then the second component that we are seeing as a massive, massive trend in the space is evaluations. It's essentially a way to evaluate how good is the LLM at actually producing a good response. And you need this because like I said before, it can hallucinate.

(25:28):

So evaluations [00:25:30] are, you're almost, it might sound a little meta when I say it right now, but you're asking an LLM to judge itself. So think about it like a human grading another human, you're having an LLM grade an LLM's response. And you're asking it to basically judge how good was my prior response based off of here's the information that I had, here's how I answered it, how good was this response? And we're actually seeing [00:26:00] this using an LLM as a judge, as a really popular way to give folks an evaluation of how good these outputs are.

(26:10):

And so these are two really common trends we're seeing in the space, especially as these applications are going into production that I think are only going to grow this year. There's just no way we're going to have successful applications if we are blind to what they're actually doing once they're launched. So it's kind of one of the [00:26:30] core foundational thesis that I think we strongly believe in.

Jorian Murray (26:34):

Sounds like excellent advice. I'd like to finish with a discussion about education, Bragg, both for adults but also children. What are your observations about how education should evolve in an AI enabled world?

Prag Sharma (26:53):

Yeah, it's a really interesting topic. Firstly, educating ourselves, keeping up to speed with what's happening in this [00:27:00] world is a different topic. Educating ourselves is interesting. But education as a whole system hasn't changed that much. If you think about it, there is still a teacher standing in the front of the classroom with a whole host of students. We still pay an exorbitant amount of money for some of the top universities in the world to go there and mix with people and listen to lectures. Why when we can get most of that information potentially for free, we live in an information rich world. There is no lack of information [00:27:30] on any topic. It can be highly technical, it can be highly general. If we dig enough, we can find that. Even so, I think so far it's been similar, but large language models might make a difference in this space where specifically they might make a difference when it comes to one-on-one education when it comes to that personal coaching, when it comes to that personal time where you want to really accelerate your own learning as separate from one collectively in the classroom.

(28:00):

[00:28:00] So that's where I think they'll make the biggest impact. Overall, I don't think the structures we have over the last thousand or so years of reputable universities of schools of learning through a teacher is going away, but they're certainly going to get partnered strongly with large language models supporting individuals either in the classroom or outside the classroom to bring them up to speed or to accelerate their learning on [00:28:30] a particular topic where it's applicable. And that would be my, I guess, I'm ready to take a bet on that with anybody, if they think differently, but that's where I think it's going.

Jorian Murray (28:39):

Are you prepared to take a bet, Aparna, or what are your thoughts on the evolution of education?

Aparna Dhinakaran (28:45):

I like Prag's answer of there's still this human component that that's so important in education today. Like you said, teachers standing in front of the classroom still hasn't gone away, but I think there's a lot [00:29:00] that students are now able to learn online, which wasn't the case probably when Prag and I went to school or all of us went to school. You see folks learning through videos, learning through probably majority through all sorts of online resources out there. And so I think that's only going to grow. It makes this ability to almost go to an amazing school or amazing college, less of a [00:29:30] barrier to entry for most people, which I'm really excited about. You can get high quality education from anywhere. But I still think the most important thing is there's going to be more of an emphasis on hands-on trying and building, which especially in this world is you can't just rely on knowing the theoretical aspects of all of this technology.

(29:59):

I think really important to [00:30:00] actually try it and build it yourselves and get your hands-on experience. Specifically in the LLM for less of education of let's say the general category of education. I think just as practitioners ourselves and people in the field staying educated. I mean, I think that there's a massive uptake in research papers, actually. You all have seen in the last year. I think a lot more of us are actually going directly and reading [00:30:30] the technical reports and the research papers. But then Twitter has become a great place to chat about all this.

Jorian Murray (30:37):

I think there's also an implication for education, having more of an emphasis on EQ rather than IQ, because we are AI-enabled with IQ. So teaching human skills, people skills, having those interactions with professors. So, yeah, I only see this as a good thing of humans becoming even more creative [00:31:00] than they ever have been. I've loved this conversation. Thank you so much. Thank you for explaining things so simply, and thank you for being a guest on Good Things Happen. Thank you.

(31:12):

On the next episode of Good Things Happen, we'll be joined by the World Food Programme's, Virginia Villar Arribas, and Citi's Head of Global Development Organisations, John Finnigan, to talk about global food security and how development [00:31:30] finance can catalyse change.

Speaker 4 (31:36):

Citi Group, Arize AI are not affiliated and are independent companies. The speaker's views of their owner and may not necessarily reflect the views of Citi or any of its affiliates. All opinions are subject to change without notice. Neither the information provided nor any opinion expressed constitutes a solicitation for the purchase or sale of any security. The expressions of opinion are not intended to be a forecast of future events or a guarantee of future result.

- ENDS -