accenture

EP 62: TEST AND EVALUATION FOR AI READINESS

AUDIO TRANSCRIPT

00:00:00 Vijay
You need to ultimately have models that you can trust that give you accurate insights and that do that in a timely fashion.
00:00:12 Teresa
Hi, welcome to another AI leaders podcast. My name is Teresa Tung. I'm Accenture's chief technologist for cloud data and AI. I'll be your host for this session on enterprise AI readiness. And we're going to be really focusing.
00:00:27 Teresa
On testing and evaluation.
00:00:29 Teresa
I'm thrilled to be joined by Vijay and he's our field CTO for scale AI and so thank you, Vijay, for joining us and I would like you to start by introducing yourself and scale.
00:00:42 Vijay
Thank you. Thank you, Teresa, I'm really excited to Chad and and excited about all the work that you've been doing as well in terms of and it's AI readiness and testing and evaluation. So really excited to talk about our work together in the future of what this means for the industry. I'm Vijay ternary and the field CTO here at scale AI. And for those of you who don't know us, we are.
00:01:03 Vijay
About a 8 year old startup that now is one of the key players in the AI ecosystem.
00:01:09 Vijay
In part because we focus even in the early days on what it took to build trustworthy AI models. So our founder, Alex Wang started the company really think about really difficult problems with building computer vision and self driving tests, specifically what it took to make self driving vehicles Rd. worthy. And so one of the things that.
00:01:29 Vijay
It was a big challenge for AI researchers at the time was understanding. If you had a computer vision model, how do you know that model can live up to a safety milestone? For example, how do you know that your perception model on your vehicle is going to stop when there's a pedestrian crossing?
00:01:41 Vijay
Crosswalk and a really underappreciated part of building AI models, is focusing on the data that you need in order to achieve certain guarantees of performance. So if you're looking at a model that needs to stop for pedestrians having data around pedestrians that do walk across crosswalks, whether that's 2D computer vision data or LIDAR data, or mapping data combined together.
00:02:02 Vijay
With where crosswalks are known to be found in a given region, all of that can be incredibly important in order to train that model and then ultimately evaluate the model before you ever put a vehicle on the road. So Alex, when you built the company, started focusing on the data equation of what it takes to build better models and more trustworthy models.
00:02:19 Vijay
Now, over the years that ended up playing a really important role in other modalities as well, especially text and other forms of human creative output. So we really had an amazing opportunity over the last couple of years to partner with really industry reading research labs like open AI on building the future of text models.
00:02:39 Vijay
And one of the ways in which that culminated

was helping build the data.
00:02:42 Vijay
Line that was involved in testing RHF against GPD 3 and that ultimately ended up becoming ChatGPT when, when the researchers trained that model and incorporated RHF as a technique and a model that could actually chat with human users around the world and give them really creative, insightful outputs that I think really kind of.
00:03:02 Vijay
Really set the world on fire and it kind of accelerated everyone's timeline for what you could do with this tech.
00:03:07 Vijay
Analogy so early Jeff is a really important part of what we do. For those of you that don't know that, that acronym that stands for reinforcement learning with human feedback and at the end of the day it's a data centric approach to how you train models. It means you get human feedback, sometimes expert feedback and how models are performing and you incorporate that into reward models. They're used to train the next generation.
00:03:27 Vijay
Models that you might test as a user or that you might deploy in an enterprise situation and it's right now. A really key part of how this entire AI ecosystem is.
00:03:36 Vijay
Built from there, we've also gone into expanding what we do to testing and evaluating models as they're used in the real world. So as enterprises are adopting chat, BT or other models for all sorts of enterprise use cases, they've found that they need better testing, better evaluation of these models, sometimes evaluation against private data or against real world scenarios that are tough to duplicate out.
00:03:59 Vijay
Public and being able to do that is something that's still a big challenge. So we play a really important role in that across the industry and as well in the public sector. So we work really closely with the White House with the Department of Defense and we were chosen by the CDO, the chief data and Analytics Officer for Defense Department in order to evaluate how models can be deployed in general.
00:04:19 Vijay
That can be deployed across that entire department, as well as other agencies, and testing and evaluation is a really key part of that right now is it's understanding where these models perform really well, where they underperform and helping human beings use generative AI better by understanding what the real opportunities are and the limitations are when you use a given AI application.
00:04:36 Teresa
Well, thank you, Vijay. Certainly scale AI sounds like one of the only companies who's actually implemented AI and specifically Gen AI at scale and into protection scenarios. I think that importance of testing and valuation in all the cases that you mentioned is so key, right. We hear a lot about one of the barriers to adoption of especially.
00:04:57 Teresa
Generative AI is around Halo.
00:05:00 Teresa
Nations. So hallucinations when you have inaccurate results. But really, those foundation models weren't built for accuracy. They were built for generation and creativity and and other purposes. So what you mentioned around being able to add testing and evaluation, whether it's more classical.
00:05:21 Teresa
AI, where you do have mathematical measures, but also now, especially in the space of generative AI, sounds like it's much more important than ever for making it production ready so.
00:05:32 Teresa
How does that realistic testing and evaluation impacts the deployment and performance of AI applications?
00:05:41 Vijay
Yeah, that that's a great question and it's really the ultimate bottleneck that's kind of holding back to Geneva and Enterprise and where we see this tremendous potential to be unleashed over the next couple of years is understanding how do you evaluate AI applications so that you understand how they're going to be performing when you deploy them in a real world.

00:05:58 Vijay
Use case which looks very different than just giving a consumer a chat bot and letting them run with it and letting their imagination run wild when you're using general and the enterprise, oftentimes you need a model that can accurately give you insights into what's happening with your data. If you're using it for a data analysis function or as a copilot to assist someone in talking to a customer about their opportunities to do business with you or to.
00:06:20 Vijay
Expand on their relationship.
00:06:21 Vijay
You you need to ultimately have models that you can trust that give you accurate insights and that do that in a timely fashion. So testing and evaluation is really this principle that comes from software engineering that and really you know over the last 20-30 years this has been one of the most important insights about how we build computer programs. You start off by defining what the computer programs are expected to do and laying down in a way that you can test the programs.
00:06:43 Vijay
As you're writing them and then understand how those programs perform before you ever deploy them to a customer or in a real world scenario.
00:06:49 Vijay
You know, without doing that, you know, programs can do all sorts of things. They can surprise you. They can do something that's really harmful. And so testing and evaluation has become a key principle that we all learn as computer scientists when we're training. That hasn't changed in the general AI world, even though these models are incredibly creative and they can do really interesting things depending upon how you prompt them. And that's become a sophisticated field in its own way. What's called prompt engineering.
00:07:11 Vijay
Where you go about constructing a question or a query to a model to get really interesting outputs.
00:07:16 Vijay
Of that, in spite of that change and how we're doing engineering today, testing evaluation is even more important if because of all the range of ways in which a model could possibly respond and all the different ways in which the ways you pump the model or retrieve data for the model can cause bias in the model's outputs. So you mentioned hallucinations.
00:07:37 Vijay
That's a really important term. That's one of the things that we look at. Technically, it means it's making statements that are inaccurate.
00:07:43 Vijay
Or that don't reflect what the user asked of the model. The you know the classic example people give if you ask for a recipe for chocolate chip cookies and the model includes mushrooms and the recipe, that's probably a hallucination. Most people don't expect to see mushrooms in their cooking, and so you go back and you try to diagnose why did that happen? Why did that prop up and there a range of different ways when you think about deploying general applications that that could come in?
00:08:03 Vijay
One is how the models trained itself and what's called pre training on all of these large animals today are trained on trillions of tokens of data that were retrieved from the Internet.
00:08:12 Vijay
And so any sort of associations that they might have seen in that pre training data can later crop up in how that model thinks and reasons about a problem that you get the model. So if the model has been trained on social media or blog posts and those blog posts say something inaccurate about the world around us, you might expect to see that in the output of the model. But one of the things that mitigates against that problem is post training and really, you know, one way to think about post training is if the model.
00:08:34 Vijay
Comes out with the high school level education on certain topics out of pre training from having read all these posts. Post training is really give the model of the college education the critical thinking that it needs in order to understand what's an authoritative source of information. Been trained on how can I reference that line of thinking? How can I respond accurately to customer?

00:08:50 Vijay
Questions given a range of different ways in which we could go about kind of harnessing the human insight in order to to further train that model, so post training is really important part of what we do, but it really hinges on testing and evaluation, understanding where you need to do that post training in order to improve model performance. So testing could mean actually having a model answer, multiple choice questions in a given area, pretty static.
00:09:12 Vijay
Test sets that you can.
00:09:13 Vijay
Correct, this is really useful to understand. Does a model perform really well on biology or chemistry questions? You can ask it. 1000 Multiple choice questions in biology and see if the model performs really well against those. And there are huge ranges of differences, especially last year across different open source models the most, thus performing closed source models are out there. You could see a range of different performance areas using just that.
00:09:33 Vijay
But there are real limitations. Just relying upon that for testing and evaluation alone. One is that the models are being asked in real world to do a lot of creative tests and construct really nuanced outputs that you might ask the model to write A blog post for you, or write an e-mail for you. Or you might ask them all to send you an alert if it sees something suspicious and a range of different document.
00:09:50 Vijay
Out there and those sort of just simple multiple choice questions are not sufficient for that. So then we get into more sophisticated evaluation and that's really where model assisted valuation becomes important. This is where you incorporate human insight. You have real world examples of human beings sharing what a great e-mail looks like or what a great blog post looks like at a given subject. And you can use models to scale up that insight to test a variety of different scenarios for.
00:10:12 Vijay
And assistive testing means you could try out a variety of different prep engineering techniques. You could try to retrieve different sources of data as part of your query and you could see how the model performs in a range of different scenarios rather than just hinging your answer on whether the model did really well on just a simple multiple choice question or not. So we're really, you know, pushing this forward. We do what we're called hybrid model evaluation.
00:10:33 Vijay
We use both human insight. You know, human feedback, sometimes direct feedback on how the.
00:10:36 Vijay
Well puts look and ML assisted feedback, so scaling that up using ML assisted tooling in order to do 1000 types of inquiries that a human being alone wouldn't have the time or the effort.
00:10:46 Vijay
To be able to.
00:10:47 Teresa
Well, thanks for breaking it down. I mean, I think like you were mentioning in software development, software engineering, it's taken us quite a while to come up with a a framework right in the standards. So everything from unit tests, all the ways to automated and continuous.
00:11:03 Teresa
We're we're just getting started, especially in the age of Gen AI. I would say even before that, AI tended to be even more of a cottage industry, right? If we think about even your classical more diagnostic or predictive AI, it tended to be within the same team, I think.
00:11:22 Teresa
Very few things became production at scale and so now that everything is going to scale and that's what we want to get to as quickly as possible, being able to rethink.
00:11:32 Teresa
I think all these areas, so where should organizations begin, right when they're testing and evaluating AI models?
00:11:43 Vijay
Yeah, it's a great question because oftentimes organizations right now are thinking through potential opportunities, search under VI and a lot of them think through which opportunities have the highest impact on me as an organization and which ones are really going to move the needle.

For me as a business or which ones kind of achieve mission critical objectives, especially if I'm a government agency and we see a lot of organizations starting with a list of different ways where they could be using generative AI.
00:12:06 Vijay
And then trying to think through next steps or how they evaluate each of those different scenarios.
00:12:10 Vijay
So one of the ways in which we start is thinking through which of those scenarios are most realistic to accomplish in the next six months in the next nine months, given the the operating overhead of deploying a model in those scenarios as well as the risks associated with Jenna in different scenarios. And you can look at a different sorts of implementations.
00:12:30 Vijay
Having a spectrum of risks that you're taking on when you rely upon generative AI and to talk to a.
00:12:36 Vijay
Customer or to to improve an operational scenario. So you know talking to a customer might be the example of a really high risk category. If you're talking to the customer about insurance policies or refund policies, you obviously want to be incredibly accurate when you're talking to customers about something that impacts how they do business with you, how they perceive you as a business, whether they consider you a reliable source of information about.
00:12:57 Vijay
Your work and your products and what you could do for that particular.
00:13:01 Vijay
User and so you need to do a lot of upfront thinking about how you could build an accurate, reliable system that could talk to a customer. So when you start with that, you start thinking about, well, what data sources do I have to test whether this system is reliable? If I'm talking to customer about insurance policies that I sell, if I'm an insurance business, do I have the right set of information about how those policies differ from one another on certain guard rails around things I should never say to a customer?
00:13:24 Vijay
On insurance policies that should never claim to a customer that they absolutely are going to renew next year if we know that there's a certain test that the customers would take before they renew those sort of policies are things that you can build in into your system orders and how you going to.
00:13:36 Vijay
And really, once you have that, you are already off to the races with thinking about, testing and evaluation. You were thinking about what are the risks that are associated with this model and how do I go about testing that and what sort of data do I have in order to construct those tests? And that's often where we come in at scale. We take a look at the range of data sources that you're going to be using for that application and the range of different ways you might be talking to a customer or you might be assisting someone internally within your workforce.
00:13:59 Vijay
And how you go about scaling up that insight into a set of tasks that can help evaluate these models.
00:14:04 Vijay
Are performing. From there you can then get started with thinking about different sorts of models you could be using different sorts of architectures, what's called retrieval, augmented generation, or even text to SQL in order to get the data that you need to answer that and you have the test and evaluation frameworks set up to evaluate how those perform. Oftentimes you have many different choices about how you build that architecture as well as today you have many different choices of models that you could be using.
00:14:26 Vijay
And lots of different parameters you may be concerning there. The security of those models, whether you retain the IP as you fine tune those models.
00:14:33 Vijay
But once you have the test evaluation harness, you're on a better playing field to understand how everything's going to perform once you go to production, and then you can see where the most realistic path forward within the next three

months or six months, rather than getting bottlenecked at the end of that journey and seeing, oh, I didn't actually live up to what I need to do in order to deploy that copilot or deploy that chat bot, which is the worst case scenario to be. It's terrible to invest a lot of time and effort into Jenny.
00:14:54 Vijay
And never have the testing set up to understand what does it actually take to get.
00:14:58 Teresa
This into production, so it sounds like organizations should start with what they know best, right? It is what needs to be true, so both the data around the.
00:15:07 Teresa
The decisions to be made, how they make that decision, what a correct answer looks like and why right so it is their core business and with that in mind then you could use that for a lot of purposes model evaluation and making it instrumented as part of retrieval, augment and generation and the rag, right, like you could use it for all these cases.
00:15:28 Teresa
That you still start with that same data about this is like ground truth, right?
00:15:36 Teresa
With that in mind, UM how does testing and evaluation differ? Or does it differ when adopting pre packaged AI solutions where it's a full stack app, right versus developing custom applications?
00:15:50 Vijay
Yeah, that that's a great question. So you know examples of prepackaged AI solutions might be just leveraging an Open Access model, one that's closed source and doing some prompt engineering and using that in order to answer questions on behalf of an internal support team or a customer. But basically not having to do a lot of complicated architectural decisions in order to deploy.
00:16:11 Vijay
But that's often a great place to start, and it's a place where you can prototype out an idea as an enterprise. But it comes with certain limitations on how you can deploy that that kind of architecture 1 limitation is that a lot of the important data that you need in order to answer customer questions about your business or to assist someone working in your workforce, someone that maybe is new and doesn't know all of your policies as a business.
00:16:34 Vijay
A lot of that data is actually internal to you as an enterprise and you have certain access controls that you've already set up on those that data or that information about you.
00:16:42 Vijay
Business, for example, if you have databases that store customer transaction data, you have all sorts of access controls you set up to make sure customer data doesn't leak because that's particularly sensitive to reveal. You know how your customers are using your products are where they're using your products or how your business has trended over time. All sorts of information of those business you just can't have leaked out into the world and it becomes really challenging to think about and off the shelf.
00:17:05 Vijay
System that you could just trust to send that sensitive data over without understanding all the different constraints that you need to put in place. You know, do you have the right sort of auditing about how that data is being used? Do you help the model understand where it should refuse to answer a question? If you're asking something that's too sensitive as it is?
00:17:21 Vijay
Off guard rails in place to prevent the AI application from revealing to something someone that something that's incorrect or something that you just didn't have the right kind of infrastructure in place to help answer that question before that ever causes harm to your organization or to your customers out there. And so that's often where we see customers thinking about what they can do internally, what they can do to build an application of guardrails in place.
00:17:42 Vijay
Around these.
00:17:43 Vijay
Levels and what they could do to improve the the state of the situation. One aspect of this is understanding how the data is stored and how you can prepare that data to be better suited for

generative AI. When you have data with really tight access controls on it, you want to make sure those access controls are preserved as you're preparing that data for being used by the generated application. And so there there's a lot of work to be done there.
00:18:05 Vijay
We actually built some systems with our platform, what we call the scale General AI platform that take a lot of guesswork out of that. They help you understand what sources of data are structured in SQL databases, what sources of data might be in PDF documents that you store internally on, how you can set the right access controls around access to that data for your application and how you can go about maybe deploying models where they need to be deployed.
00:18:26 Vijay
To where the gravity is of that data. So some data may be more easily accessible and available on the open Internet. Some data may be closed and kept within your own on premise data center.
00:18:37 Vijay
Sometimes you need to deploy models directly within those data centers in order to keep them closer to the data and the security of that data for that given application. So our platform helps you get a big picture view of how all that data is sitting, where that is flowing and how you can go about setting up an overall architecture around that. And the second piece of puzzle then is evaluating performance. So the performance of a general.
00:18:57 Vijay
Location can mean a couple of different things, but a lot of times it means do you have accurate answers? Are your answers precise? Are you recalling the right information or to answer the customer question and you need to set up that testing and evaluation, or understand where the different components of the system may be underperforming and where you need to focus.
00:19:12 Vijay
On that or where you may be getting adequate performance just by using off the shelf software. So we often see that's the conversation is you know, let's start off with this basic prototype, but then let's think through what are the more sophisticated tooling and testing that we need in order to get this into production level performance. And that's often where customers turn to our platform in order to help make that journey possible.
00:19:33 Teresa
And that goes.
00:19:33 Teresa
Back to the term, the reinforcement learning human feedback, the RLH that you mentioned right in the beginning when you describe scale, right, being able to connect the the loop that starts with the experimentation, you're using the model, but then ultimately that accuracy and.
00:19:50 Teresa
Performance. Sometimes it's not mathematically clear right within UM generative AI. Sometimes it is right we we know that mushrooms are not likely in chocolate chip cookies, but whether this is a even you mentioned you know even like code generation like what is good code look like? Whether I'm looking at interpretability or whether I'm looking at.
00:20:13 Teresa
You know performance, they all might be good in some measure more subjectively, like what is a good marketing description for product also could be very different. So I think being able to connect that is really key.
00:20:30 Teresa
I wanted to talk a little bit more about maybe fine tuning next, right. So you mentioned that starting with some of these pre trained models is a good way to begin. But many companies as they're moving into production and they start building upon this corpus of data that they have about their troops, right.
00:20:49 Teresa
How can organizations fine tune their AI models to further meet their performance goals?
00:20:56 Vijay
Yeah, that's a great question. And and 1st, I apologize, I know you and me are gonna get 1000 letters with recipes for mushroom chocolate chip cookies after this, and I appreciate that. I'm sure there is a lovely mushroom cookie that's out there. So I didn't mean to offend anyone. One of the real insights

that we had over the course of the last, you know, couple of years of working in this space.
00:21:17 Vijay
Is really just how important it is to consider the data equation when you're thinking about fine tuning or you think about evaluating the performance of these models because it really starts with the data and the data that you have.
00:21:27 Vijay
And one of the big opportunities now in enterprise usage in AI is that compared to what's publicly available with either open source models or the most powerful closed source models which are trained on a huge range of data on almost all of that data, that those models have been trained on is available on the public Internet or is licensable in some way. So they might have been trained in a range of news.
00:21:47 Vijay
Articles or a range of blog posts, but they don't have access to a lot of data that for whatever reason needs to be kept private and sensitive, which actually enterprises have a lot.
00:21:57 Vijay
They have a lot of information on patterns that they've seen from their customers on ways in which they've talked to their customers and ask customer support scenarios. The tone of voice that they found to be useful in how you write an e-mail or how you go about writing any sort of messaging that's customer facing, all of that knowledge that's kept within private data and private data stores and it's being incredibly valuable.
00:22:17 Vijay
For building AI applications and you ivory, one of our enterprise customers spends time thinking about how they can harness the capabilities of that private.
00:22:24 Vijay
Data. So there are couple of different approaches here and and you mentioned you know rag and retrieval augmented generation, that's been one of the most important patterns that we've seen. Customers need a platform in order to do at scale. It's really leveraging all the different sources of data you could potentially have whether those are SQL databases, unstructured data on finding the right ways to store that in a vector database or another type of embedding.
00:22:44 Vijay
Store so they can retrieve the right documents that are relevant to a given query at the right moment. That's really changed the equation for search and how we go about doing search within the enterprise, and it's been a really transformative approach to building these apps.
00:22:56 Vijay
Almost every enterprise scenario we deploy and involves rag at some level or another, and so having platforms to help you scale up the opportunity to use RAG ends up being really important and really empowering for these organizations as they think through five or six different general applications they might want to build in the future. But once you start with rag, that's often where you have rag and you have testing and evaluation.
00:23:15 Vijay
And you realize there are certain performance goals you're trying to hit that you're not able to achieve just off the bat just by using simple rack. Some of these might revolve around hallucinations, so you might start to see the model retrieve something that it doesn't quite understand, and then it hallucinates in the output because it didn't really understand the context in which.
00:23:32 Vijay
Was said you might see the models get confused about the timeliness of information. So if you have research reports or you're using for wealth advisors as a bank, some of those research reports might have written five years ago. Others might have been written last week and the model may not understand off the bat from a given chunk of text. Whether this was relevant right now or is relevant a long time ago. I mean, you may start to see hallucinations as a result of that.
00:23:52 Vijay
You may just end up with models confused about the range of data that you're presenting to them and refusing to give an answer because they just don't understand how they identify the patterns within that data.

00:24:00 Vijay
So testing and evaluation is an important part of figuring that out. Once you start to ask real world questions to these models and you scale that up, you can see where those models underperform in different sorts of scenarios, and you can also tweak parameters there within rag, and that's how you build more advanced rag systems. An advanced rag system is one where you're considering the embedding model you're considering. The chunking strategy you're considering different ways of re ranking.
00:24:20 Vijay
The data before it's returned to the model and you're doing all that in order to achieve some sort of performance goal at the other end of that. So we spent a lot of time focusing on that.
00:24:28 Vijay
First, lastly, fine tuning comes into play when you have a performance goal, you need to hit. You started looking at in a more advanced rack pipeline and you realize that the model is underperforming because it doesn't have a lot of domain awareness of what you're asking it to do, and an example of this might be a model that you're asking to do. Code generation and example you gave. If the code you're writing is data analysis code and it needs to query.
00:24:48 Vijay
Specific databases or data stores maybe about financial information or healthcare information or insurance information. You will very quickly realize that there's performance barriers you hit just because the model doesn't understand how to translate an insurance term into something that needs to be an efficient SQL.
00:25:02 Vijay
Query and the model needs to see examples of other SQL queries people have written and needs to see ways in which data analysis, edit or refine their queries in order to get the best possible window of performance that you can achieve out of that model. So we go about doing that. We generate a set of data that can be used to fine tune the model and that's really a a pipeline in and of itself of translating examples of queries.
00:25:22 Vijay
Examples of data that we have available creating a structured set of data that could be used to fine tune the model.
00:25:27 Vijay
And then go about going through fine tuning and observing out the performance gain so you get out of that and that ends up being a very incredible, you know, the often the last three to six weeks of deploying this application, we see the biggest wins from just fine tuning the model and getting a model that's domain aware that's capable understanding your specific data stores and capable understanding where authoritative sources of data within your organization.
00:25:47 Vijay
And it gives much more accurate answers at the other end.
00:25:49 Teresa
So again, going back, it goes all back to the data. So for a company, a lot of what you're doing is you're building upon this corpus of of of data that you then apply with increasingly.
00:26:03 Teresa
More accuracy for your use of Gen. AI, but it could just start with as you evaluate some of these out-of-the-box models. The use of rag and then into fine tuning. So as we look at how organizations should validate that improvement right, we mentioned that there's this evolution and we're going to get more performance.
00:26:23 Teresa
From fine tuning, we validate those improvements, right?
00:26:27 Teresa
The through testing.
00:26:29 Vijay
Yeah, yeah. So testing obviously is a really important piece here and it's really testing against real world scenarios and and translating those real world insight into how you can go about systematically testing the models. So it's great to have humans in the loop as part of this, you as an enterprise, you have all sorts of experts against your specific ways of talking to customers.
00:26:50 Vijay
Your ways of thinking through documents. If you're tasking a general application with finding areas of legal documents that maybe pose a compliance.

00:26:57 Vijay
This risk you have all the experts right there with the organization that have done this work in the past that have identified scenarios that maybe reveal problematic passages or changes over time that you should be worried about. And so leveraging that expert insight that you have an organization into how you test the model since being really important, we've actually built within our scale.
00:27:17 Vijay
Interv.
00:27:18 Vijay
Form this amazing painted glass to construct testing evaluation using human experts and that's really where you start. You give examples of real world tests that the models are being asked to do. You have human experts coming in and observing where those tests are underperforming and sometimes even just giving written feedback to the models to help understand, hey, this is how we could improve the models understanding of the task or the world around it.
00:27:39 Vijay
Or even help them understand what sort of domain that is.
00:27:41 Vijay
Being asked to to work.
00:27:42 Vijay
And we translate a lot of that expert insight into automated testing by using analysis and tooling, and that can involve just rewriting the prompt multiple times. It can involve recharging the scenario about how you go about the giving specificity of what you're asking model to do a range of like different ways of evaluating model responses that go way beyond just multiple choice questions.
00:28:02 Vijay
Or simple arithmetic answers and really get into evaluating what the models giving precise and formal language when you needed to, or more casual and sympathetic language when you're talking to a customer.
00:28:14 Vijay
And scenario all sorts of ways you can evaluate how the model performs, so translating a lot of that internal private knowledge that you have within your organization requires expert insight right now, but we hope in the future, as you're scaling up testing and evaluation, you have a range of different ways you have of testing models and you've kind of created this, this bank or or library of tests that you've constructed.
00:28:34 Vijay
That can be used across a range of different applications and help you understand how those applications.
00:28:38 Vijay
Perform the second piece that we see customers using testing for and it's interesting. When we did our AI readiness survey this year in 2024 of thousands of different enterprises that are all adopting gender of AI across a range of different industries, we often hear that security is one of the most important pieces that they're testing for. So as you're deploying models and they're able to access data.
00:28:59 Vijay
You want to make sure that doesn't opening you up to security vulnerabilities to to customers or partners being to inquire about things about your business that you're particularly sensitive.
00:29:08 Vijay
And so creating a really robust test suite that looks at things are important to you as an organization and ways in which you want to be secure ends up being incredibly critical, and you can use that across a range of different generic scenarios. Once you have a good set of cyber security policies in place, testing for those good set of harms that could possibly come about as part of these models, a range of different things that you've observed as being.
00:29:28 Vijay
Problematic outputs that the models could create. All of that can be internal to organization and can be stored in this Bank of really good security testing that you can do.
00:29:36 Teresa
So one of the good things that you're doing is you're scaling as well that expert knowledge, right. You mentioned that I'm able to take that time that I have with the expert and then scale.
00:29:44 Teresa
To create all these different prompts that test for much more than what the expert would really

need to know, he or she doesn't need to know all those different variations you have to test with. They need to help you with the expertise of what what good looks like and what it should be true. So with that in mind, Can you imagine scaling across industries?
00:30:05 Teresa
Or cross functions, right? How do industry or function specific standards and benchmarks contribute to responsible AI deployment?
00:30:15 Vijay
Yeah, that's a great question. And it really gets to how regulation is evolving in this space. So one of the things we've seen here in the United States, which is really kind of benchmark approaches to regulation for the rest of the world as well, is that every different agency within the US government that's tasked with regulating a given industry has been encouraged by the Biden Harris administration as part of the I Act.
00:30:35 Vijay
I think through how testing is going to work for their industry.
00:30:38 Vijay
On how they're going to go about evaluating different harms that are industry specific, how they're going to go about evaluating reliability, these models are for different scenarios that might be mission critical scenarios or might be more operational efficiency, different scenarios. So great example might be the SEC involved in setting regulation for, for how you know you do reporting and you do testing.
00:30:59 Vijay
All around the financial services industry and thinking through ranges of ways in which before you ever deploy a chat bot and from a customer or copilot to assist a financial services employee, you understand how those models perform and ways in which those models could be underperforming. To help deploy that responsibly.
00:31:16 Vijay
Ultimately, we're getting to better ways where AI can be used responsibly across every industry, and that comes with a lot of industry specific insights that a lot of these agencies have. So what that means at the enterprise level is that in order to get ahead of these regulatory changes, you need to understand as an enterprise, what do you know about your industry, what do you know about ranges of potential security risks or harms that are specific to you?
00:31:37 Vijay
How do you go about testing those? How do you go about demonstrating that these models are free from bias, where that's an important consideration for her?
00:31:44 Vijay
In the models, a great example of financial services might be if you're deploying a general model to look through mortgage lending applications. There are obviously you know really well known situations where you need to look at bias and how those models perform. You don't want a model that's biased by the particular city that a customer's in or particular part of that city that customers in. If that's not an appropriate way of considering a mortgage application, you don't want to be biased.
00:32:04 Vijay
Maybe by all sorts of ways in which a customer's background may not be that relevant, but maybe could influence the decision that AI application makes. If it's kind of deployed now.
00:32:13 Vijay
Really. So testing for bias is an important piece of how you go about thinking about testing and evaluation. Setting that up during a range of real world scenarios, looking through real world data and insight from your experts and ways in which they might have seen bias in the past and really looking at more robust sources of data to help mitigate that bias. So you can help encourage the model through prompt engineering or can help look at RAG and finding.
00:32:34 Vijay
More robust sources of data. All that can mitigate bias and important ways.
00:32:38 Vijay
For example, rag can look at a range of different geographic regions where mortgage lending applications are coming in, not just cities but rural regions and suburbs and other parts of the country, including all those as part of your retrieval. And rag can really help remove some sources of bias that may come about from from a particular city being, you know, problematic for

for how these applications consider.
00:32:59 Vijay
Mortgage lending.
00:33:00 Vijay
Iterations. So when you think about that, you can think through hike and optimize on the right pipeline and and prompt engineering and ultimately that influences fine tuning as well. You know you set up a robust set of evaluations to look at bias. So you can now fine tune a model to to to consider that and hopefully have less biased answers. As a result, it really just starts with understanding.
00:33:21 Vijay
In a real world context, what are ways in which that sort of data can be influencing the decision the AI application makes?
00:33:26 Vijay
And ways we can find to the model to.
00:33:27 Vijay
Avoid that in the future.
00:33:29 Teresa
Yeah. No, I was thinking like industry specific benchmarks or function specific benchmarks like if we were all in oil and gas or you mentioned insurance just coming together and working together on some of these common corpuses that you could start with as the benchmark for that industry function. And then you can then add your company.
00:33:48 Teresa
Specific benchmark right afterwards but.
00:33:51 Teresa
Possibly something like that might be needed to bootstrap and get us to production and scale as an industry faster than we might be able to do by ourselves.
00:34:01 Vijay
Yeah, that's a great point. It's something that's a big need for is industry specific benchmarks for financial services, for healthcare, for insurance. We actually we published within scale a set of research reports this year from really starting with public sector use cases of AI.
00:34:14 Vijay
But that we have a group within scale called the the SEAL lab. The the safety and Evaluations Analysis Lab led by summer. You, who's the former one of the former early Jeff leads for at Google Deep mind for when the the Bard models were constructed. And one of the objectives of this lab is to publish more public benchmarks that are robust to a range of different industry scenarios and reveal better ways of doing testing and evaluation of these models.
00:34:38 Vijay
Ways that maybe aren't biased by whether the model is pre trained on a set of data or not, but actually kind of live up to the real world uses of these models. So we've published some initial work. You know one is focusing on really common math questions that are asked of these models. There's a data set called GSM which is great school level math and it turns out models actually performed pretty poorly and a lot of math.
00:34:58 Vijay
Questions out-of-the-box. So it's a really good test of how the model was trained and its ability to be a robust to.
00:35:03 Vijay
Scenarios, but it's a very general benchmark that's good for value models. I think in the future you'll see those general benchmarks translate into industry specific benchmarks that we're able to share, leveraging a lot of licensed data sources, ways of thinking about industry specific scenarios where we can glean that from public data. That's can be incredibly important because I think if you look at some limitations of benchmarks today.
00:35:23 Vijay
In tasks, for example text to SQL or translating natural language to SQL, they are often constructing academic setting using sometimes pretty naive situations like you might have a database with a a pretty simply defined schema that's really easy to understand.
00:35:37 Vijay
And it turns out in enterprise, they sometimes have 20 different databases with lots of different tables, all with the same name or maybe 30 columns, of which only five are actually used, and those sort of tricky noisy scenarios that you see in the enterprise or something. We need more public benchmarks against as well, just to sort of how we're doing it across a a real world

task. So hopefully we'll be able to push more then and we're also encouraging others industry to collaborate together too and publish more of those.
00:36:02
Touch marks 2.
00:36:03 Teresa
Well, thank you, Vijay, for spending some time with us and the amazing work that scale AI is doing to implement AI at scale. My lessons learned here is that to do Gen. AI at production, you need to get your data right, that your data is your competitive advantage without which you won't have.
00:36:22 Teresa
AI or Gen. AI in particular, that is relevant for your business and UM so making sure we start with that with testing and evaluation whether we're using a pre packaged.
00:36:36 Teresa
Application building our own RAG fine tuning. Creating a custom model. It all starts with that expertise and companies should really like that because you are the expert. So thank you so much.
00:36:47 Vijay
Thank you, Teresa really enjoyed talking and and hopefully we have a great future ahead with testing and evaluation and more robust scenarios that we're able to evaluate this miles again.