SDS PODCAST EPISODE 761: GEMINI ULTRA: HOW TO RELEASE AN A.I. PRODUCT FOR BILLIONS OF USERS, WITH GOOGLE'S LISA COHEN

Jon Krohn: This is episode number 761 with Lisa Cohen, Google's Director of Data Science and Engineering for Gemini, Assistant, and Search Platforms. Today's episode is brought to you by Ready Tensor, where innovation meets reproducibility, and by Intel and HPE Ezmeral Software.

Welcome to the Super Data Science Podcast, the most listened-to podcast in the data science industry. Each week, we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn. Thanks for joining me today. And now, let's make the complex simple.

Welcome back to the Super Data Science Podcast. Today we have the incredible fortune of having Lisa Cohen on the show and at just the right time. Earlier this month, Google announced the public release of Gemini Ultra, their largest language model and the only LLM that is comparable to GPT-4. I've personally been experimenting with Ultra and have found myself using it on some tasks that I've previously relied solely on GPT-4. Ultra maintains attention across large context windows, competently generating natural language and code. Like GPT-4 V, Ultra is multimodal and so accepts both an image and text as input at the same time. Piggybacking on Google's excellence at search, I've found Gemini Ultra to be particularly effective at tasks that involve real-time search. The Google Bard project that focused on real-time information retrieval was renamed Gemini when Gemini Ultra was released.

Lisa Cohen, our guest today, is perhaps the best person on the planet to be speaking to about this momentous release. As Director of Data Science and Engineering for Google's Gemini, Assistant, and Search platforms, Lisa is responsible for Gemini's rollout to Google's billions of users worldwide. In addition, she was previously Senior Director of Data Science at Twitter and Principal Director of Data Science at Microsoft. She holds a master's in applied math from Harvard University. In this episode, Lisa details the three large language models in Google's Gemini family and how the largest one, Gemini Ultra, fits in. She talks about the many ways you can access Gemini models today, how absolutely enormous LLM projects are carried out and how they're rolled out safely and confidently to literally billions of people, and she talks about how LLMs like Gemini Ultra are transforming life and work for everyone from data scientists to

educators to children, and how this transformation will continue in the coming years. All right, you ready for this exciting episode? Let's go.

Lisa, welcome to the Super Data Science Podcast. I'm so excited to have you here. Where are you calling in from?

Lisa Cohen: Seattle.

Jon Krohn: Nice. So I came across you because you were a recent guest on Sadie St. Lawrence's Data Bytes podcast, which if people regularly listen to the show, they've probably encountered Sadie many times. She was most recently in our 2024 Data Science Trends episode that kicked off 2024, so that was episode number 745. But yeah, she's done that for three years in a row now, as well as having her own episode on her own topics. Amazing speaker, great host. Your episode with her on Data Bytes was fantastic. As soon as I saw the teaser clip, I instantly... I can't remember how I reached out to you, or maybe asked Sadie, I can't remember, but I remember seeing that clip and instantly reaching out to you because you are a brilliant speaker working on some of the most exciting stuff on the planet right now, so I am delighted to have you on the show.

Lisa Cohen: Well, thank you so much for having me. Yes, it was fun to chat with Sadie and really happy to be on Super Data Science.

Jon Krohn: Yes, yes. Well, let's get right into these exciting topics. So you are currently the Director of Data Science and Engineering for... The title that you have on LinkedIn. Maybe this has changed now with this release, but it's Director of Data Science and Engineering for Google Bard, Assistant, and Search Platforms. Now, recently, as of last week at the time of recording, you switched the branding of Google Bard to Gemini, so I don't know if that impacts your job title or if you want to talk about that.

Lisa Cohen: Yes, that's right. We're fresh in the midst of it, but I now work on Assistant, Gemini, and Search platforms. We have renamed Bard to Gemini to reflect our most powerful family of models that you can access now through gemini.google.com. We also released an app, so you can go through Android, Play Store, or on iOS to access as well.

Jon Krohn: And I have been playing with Gemini Ultra on all of those different platforms. So, let's break down even just in your title, you have these different products that are interrelated, I guess increasingly interrelated. So Gemini, Assistant, Search platforms. What do those different things mean and what are your key responsibilities as Director of Data Science and Engineering, overseeing that vast portfolio of products?

Lisa Cohen: Great. So my role and my team is leading data science for these products that you mentioned, and so what that means for something like Google Assistant for example, we have hundreds of millions of monthly active users using Google Assistant. I'm one of them. And we want to make sure that everyone has a really great quality experience. So when you say the hey G keyword, which I won't fully say because all my devices will wake up, and it responds when you want it to and it doesn't respond when you don't want it to. We want to make sure that when you try to execute a task, that it completes successfully and it does what you intend it to, playing the right song, calling the right person, et cetera.

So we have in the data science space a number of metrics that we use to track the quality of your experience. We analyze the user journeys, of course all anonymized and with personal information appropriately stored, but we're looking for just overall aggregate user trends to remove friction points and just continue building on where we have strong product market fit. For Bard, now Gemini, we're doing data science across the breadth of the product. So again, we're looking at user experiences, understanding the use cases, understanding how much users are using Gemini for coding or for writing or summarization. How does their journey evolve as they start using the product and understand what they can and can't do, and how does it progress as they become a more engaged, experienced user, how does their conversations evolve, et cetera, so that we can really evolve the product to again be most useful. It's such an evolving space with just changing human interaction patterns with gen AI, and so we want to continue moving with that trend and helping users understand how they can best leverage this tool to realize their potential.

So again, there's measurement and evaluation to see how helpful was the response? Was it correct? Did we prevent hallucinations? Ensuring the safety of responses. Ensuring great experiences on the

app, for example. And then things like experimentation, infrastructure, logging to ensure that we have the right data and measurement to innovate and measure and learn along the way. And then for search platforms, that's really the platform part of search, so covering areas such as latency, reliability, capacity. Pretty much every product I've ever worked on, latency has been so powerful. So when I worked on Visual Studio, I remember users would say, "If you just make the next version faster, I'll buy it." And so we did a lot of work to make things load in the background. There's cold start, warm start to load, everything. Similar with Twitter, we had time to first tweet.

So with search, latency is really impactful for user experience as well, and so there's a lot of analysis that we do, whether it's profiling-like analysis to understand where time is being spent, or looking at different experiences to see where users have more sensitivity to latencies. So for a weather query, you expect that to be really fast, or a more complex query you might tolerate a little bit longer time, and that helps us prioritize the roadmap. Reliability, again for that, every product I've worked on as well is seeing an impact with reliability such as Azure. So when you have an outage, then that really impacts users' usage for some time afterwards. We see the same for search, and so we help measure and identify any reliability issues which might just also be at the feature level as well and work with those teams to ensure a great experience, reliable experience. And then also working on areas like capacity, our internal developer experience and velocity and guiding different actions around our code maintenance for the search stack.

Jon Krohn: It's unreal. It's hard for me to wrap my head around how one person can be involved in so many enormously popular products. Google's search platform is just one of the three and yeah, you said over a hundred million, or was it hundreds of millions, of users of Google Assistant, and now Gemini is also going to be taking off. It already has hugely. It's amazing, and I guess it shows by having one person be responsible for these kinds of things like latency and reliability across Assistant, Gemini, Search platforms, it shows to me at a distance, and maybe this is something you can't really talk about or dig into, but it's interesting to me. It suggests to me the importance of the interconnectivity of those experiences for users that across search, across generative things that are happening with Gemini models, across the Assistant experiences, across all of your Google

Assistant devices that you might have, you're getting a high quality, low latency, reliable experience unified across all of them.

Lisa Cohen: Yeah, and so definitely it's with a broad team, and definitely grateful to I think be able to work on these experiences that impact so many users. I think that was important to me, even in my first job, to be able to work on something where users would immediately see the impact of the work that we're doing. But like you mentioned, there's a lot of connections across, whether it might be just gaining instincts around user expectations and tendencies and approaches that we can then leverage those learnings and apply to other areas. We also look at areas where we should be consistent intentionally or intentionally different based on the nature of the different experiences. And then from a data science perspective, I think the centralization across does give us opportunity for efficiencies of reusing libraries, techniques, approaches, and that peer group to both share ideas with brainstorm approaches and socialize across.

Jon Krohn: Wow. Yeah, fascinating. It is a tremendous role. And so digging into the particular part that we were talking about earlier, and I think we're going to focus a lot in this episode, that name change in your title from Bard to Gemini, that reflects the Bard product being renamed to Gemini. Why did that happen? What does this new naming mean?

Lisa Cohen: Yeah, so we updated the name to Gemini to reflect the most powerful family of Models, which is Gemini, which we released in December. And I think by using this cohesive name across just increases clarity for users about which model they're using, and we're going to be starting to use that more broadly at Google as well. So we already have more than a million people using features like "Help Me Write" through Duet AI. And going forward in the coming weeks, Duet AI will become Gemini for Workspace. Cloud customers will also see Duet AI becoming Gemini in the coming weeks, so that way we have a cohesive story.

Jon Krohn: It makes a lot of sense. I think particularly, for me, it's obvious that Gemini, particularly the Gemini Ultra model, is incredibly prestigious. It is pushing the capabilities of what any AI system could do on the planet, and so to be associating branding with that to me makes a lot of sense. And I think it simplifies things for users as well. You don't

have to be thinking about all these different products when ultimately you're using the same model family in the backend. It just makes sense to simplify that branding, so very cool. In addition to the public release of Gemini Ultra, last week at the time of recording, a couple of weeks ago at the time of publication, what else did Google announce last week alongside that?

Lisa Cohen: So we introduced the Gemini Advance subscription, so that's now available as part of the Google One AI premium plan. It's 19:99 a month. It starts with a two-month free trial, so definitely go check it out and give it a turn. Give us any feedback. So Gemini Advance is how you chat with Gemini Ultra 1.0, so that's the largest and most capable state-of-the-art AI model from Google. We also launched a new Gemini app, so really excited about the opportunity that that's going to provide for making Gemini more accessible and seeing the mobile-friendly use cases that people are starting to use now. And so that's available on Android and rolling out an iOS, on both platforms essentially as well.

Jon Krohn: Nice. Yeah, so anybody listing right now can try out that state-of-the-art Gemini Ultra model through that subscription. And included in that means that as part of that Google One subscription, you're not getting access just to a generative model, but you're actually getting access to a suite of Google products. I don't know if that's something that offhand that you can-

Lisa Cohen: That's right. Your Drive access and additional features that you'll get across the Google suite of products as well.

Jon Krohn: Yeah, so that seems like a pretty good deal. Especially if you're already using Google Office Suite and Google Drive for different things, it seems like a no-brainer to be adding in those generative capabilities which are in the flow of work for you. So you were talking about Duet earlier, which is integrated into the Google Workspace. So if you're in Google Docs or Slides, you can be relying on this in the flow assistant for helping you with tasks, which is something that, going back to episode number 730 with Kyle Daigle, the COO of GitHub, Kyle was highlighting how important it is for these generative AI tools to be in the flow of work in order for them to be adopted and in order for them to lead to the kinds of efficiencies that employers are hoping for in getting those subscriptions.

Lisa Cohen: Yes, definitely connect with that. The first product that I worked on was actually Visual Studio. It was back in my Microsoft days, and so we were working on things like code completion and just completing that next word. And so it's pretty inspiring I think with just the state of generative AI to be able to just expand creating a class, a complete program. I think there's definitely an inline experience, as we're discussing with Duet, now Gemini for Workspace, as well as coding experiences, I think that there's also a conversational experience that sometimes gives you then a higher altitude and where you can do multistep type of instructions as well. And so really, the combination I think of those two experiences is pretty powerful.

Jon Krohn: Research projects in machine learning and data science are becoming increasingly complex, and reproducibility is a major concern. Enter Ready Tensor, a groundbreaking platform developed specifically to meet the needs of AI researchers. With Ready Tensor, you gain more than just scalable computing, storage, model and data versioning, and automated experiment tracking. You also get advanced collaboration tools to share your research conveniently and securely with other researchers and the community. See why top AI researchers are joining Ready Tensor, a platform where research innovation meets reproducibility. Discover more at readytensor.ai. That's readytensor.ai.

All right, and so that gives us a clear sense of how the Gemini family of models can be used. But Lisa, something that I'm aware of because I've been following this Gemini story since the first announcement months ago, but what are all of the models in the family of Gemini models? And why would somebody want to use one over another?

Lisa Cohen: Yeah, so when you go to gemini.google.com, you can also use the Gemini Pro model through our free experience. And essentially, with the larger models like Ultra, which you get through Advanced, these are larger more parameters. They are our most powerful capable models for things like logic, reasoning, coding. And so I think you can also experiment between the two and see the different experiences. But yeah, I think that, depending on the task, you'll have our best response from Advanced. I think that there might also be cases for smaller devices or more real-time requirements where sometimes smaller models might also be ideal. And we're looking at how to

ensure that you get the best of both worlds and the way that we implement this for you as well.

Jon Krohn: So if I am remembering the terminology correctly off the top of my head, obviously the model that we've been talking about the most here is Ultra, which is the largest, most capable state-of-the-art model. And so that is available only through subscription like the Google One subscription that we were talking about earlier. And then there's Google Pro is the next tier down. Sorry, Gemini Pro-

Lisa Cohen: Gemini Pro.

Jon Krohn: ... is the next tier down.

Lisa Cohen: And then the next one down would be Nano.

Jon Krohn: Is Nano. And Pro, interestingly, despite the name, is actually available for free. You can just go to gemini.google.com and be using that. And that is similarly the generative model that if people were familiar with Bard over the past year, that Gemini Pro now by default I think replaces Bard in that search experience. And then it's Gemini Nano, which is intended for edge devices, so a phone could be running Gemini Nano. And when I say that, I mean running Gemini Nano locally, because I guess you get access the apps that you can get from the iOS store or from the Google Play Store. You could be not necessarily running locally on your phone the Gemini Nano. You could be running from a remote server and taking advantage of even up to the Gemini Ultra on your phone.

Cool, cool, cool. All right, so we are on the same page on that. You mentioned in another podcast being excited about large language models like ChatGPT and how they represent a category changer and huge business transformation. Obviously, you are now very much in the thick of it with Gemini. In what ways do you think these models, like the GPT series models, the Gemini models, will specifically transform the practice of data science and the role of data scientists on product teams?

Lisa Cohen: I think about a couple directions. One, I think that it's really going to help democratize data further. I think with these large language models, the new programming language is English, so it's very

accessible and I think broadens data access to more people to be able to reason over data for things like business intelligence insights, you can upload a table, CSV, and ask for trends or insights on that data. And so I think that it really democratizes data further. And then I think for data scientists, I think, just like we were discussing in the developer context, we're constantly looking to increase the level of abstraction, make some of the more routine tasks more efficient. I think there's been a trend on this in general as we have more BI tools available, things like AutoML available where a lot of those implementation tasks are just further automated for you and you can focus more on the more sophisticated tasks and the creative logic behind the solutions that you're working on.

Jon Krohn: Very cool. Yeah, it's been a game-changer for me, and I've even done hour-long keynotes on this idea of how, when you no longer are responsible for writing every line of code, it allows you as a data scientist or as a software developer to be able to think one level up. And you still need to understand code, you still need to be able to write code and be able to edit code, but you don't need to be typing every character anymore. You don't need to be typing every line or every function anymore. And over the coming years, that's going to get easier and easier and easier and more and more and more accurate such that whole functions, you can rely on them being written accurately first time by a generative AI tool.

And so this frees you up as the data scientist or as the software developer to be thinking about the user experience and how now all of this code that you can quickly generate, or you and your colleagues can quickly generate, can be integrated together to create a powerful user experience. And so that's exactly the argument that I've been making in keynotes is that you are now armed with these powerful large language models. You as a data scientist or a software developer are now more of a product person than ever before, or you should be thinking that way, I think.

Lisa Cohen: Yeah, no, that's right. I think to your earlier point on staying in the flow, I think it allows you to just stay at that creative level as well. And whether you're typing out your ideas and bullet points or in pseudocode, and then having the LLM complete and implement it rather than having to keep context shifting from that high-level design to the implementation lets you stay more at that level. And I think as

you mentioned as well, you're still accountable for the code that's being written, but there's also ways that you can leverage the LLMs too to reason over the code that you're writing. I think we've just been tasking ourselves to any task that you would do yourself, keep asking yourself first, is this something that I could leverage an LLM for? And it's pretty amazing the number of times that the answer is yes,

Jon Krohn: It's unreal. I recently had a colleague say to me that they are using ChatGPT in their case less than they were when it first came out. And we do see that with, there's public data on the OpenAI website being used somewhat less. But it's interesting for me to see those stats or hear a colleague say that, because I'm using it more and more and more. Because the more that I use it, the more that I see it's so incredibly useful. I'm constantly thinking of situations, like there's surely no way that something this nuanced could possibly be nailed by an LLM, but let's try it.

And then, despite me seemingly not even giving enough of the context that I would've thought, if I had to explain this to a person, I would've thought that I'd need to give more context. But I've gotten used to this fact that I could even seemingly due to the metric, I guess, of reinforcement learning from human feedback, even just my vague context ends up pointing the LLM in the right direction of the answer that I was looking for, the output that I was looking for. And I am on a daily basis still, nearly a year after the release of GPT-4, I'm still stunned on a daily basis by the brilliant things that LLMs are turning at.

Lisa Cohen: Yes, it's a pretty amazing and delightful experience just to see the fun surprises of what you can make up.

Jon Krohn: Yeah, and so on in that vein. Also, in a recent interview, you said that it's fun to work with new techniques like prompting and fine-tuning models. On a high level, tell us about how you leverage these kinds of things. I think it's safe to say that if you're listening to this show, please tell me that you're familiar with prompting, because if you're not already playing around with prompting and seeing what that can do for you, you're really missing out. Please do that today. Please mess around with these LLMs and see how they can be improving your life of writing code, of writing prose, at coming up with ideas, anything imaginable. So hopefully prompting is something that's well-

covered, but something that I suspect not all of our listeners are super familiar with is fine-tuning models. I guess you can talk about the prompting thing if you feel like it, but even more interesting for me, I think, and probably our audience, is this idea of fine-tuning models and what you find exciting or promising about these methods.

Lisa Cohen: Sure, we can walk through one by one. Yeah, the prompting. Prompt engineering is the practice of designing effective inputs or prompts to get the desired results that you're looking for from large language models. And so this might be a preamble that you share with the model, a style guide or some context that you give it, and so I think it's also just a really powerful tool to get the best result from the model. So it might be about giving it very clear instructions and being specific about what you want or giving a few examples that help the model understand the type of outfit that you're looking for, maybe the formatting that you'd like for the results. Again, context, whether it might be code or the values that you want to come through with what you're writing, or a style guide. So anyway, there are some thoughts on the prompting.

Supervised fine-tuning is the process of taking that pre-trained large language model and then adapting it further to a specific task by training it further on a labeled dataset, and so that dataset should be of the scenario that you're trying to optimize for. So there might be an amazing pre-train model that you're using, but then you have an application or use case that you're developing for a specific class of users, and so you want to optimize more for their type of use cases, preferences, and so it's that next stage of then adopting more to that, ensuring the best quality for that. It could also be a niche scenario.

And then maybe another just technique in this group would be the reinforcement learning human feedback. And so reinforcement learning is a machine learning approach where the model learns by interacting with environment and then getting either rewards or punishment, as if you were to give a dog a treat after they do a good job, that approach. And so the reinforcement human feedback is the approach where then you use human feedback to get the positive or negative feedback signals.

Jon Krohn: Last month, HPE and Intel together showcased the power of RAG, retrieval augmented generation, to bring relevant business data to

your LLMs. In this month's free workshop, you can learn about the art of fine-tuning, embedding models to deliver verifiable conversational chatbots using the HPE machine learning development environment, powered by Intel Xeon scalable processors. No matter your experience level, join us to learn practical techniques to build trustworthy chatbots with guaranteed behavior for enterprise applications. Visit hpe.com/ezmeral/chatbots to register today. We've got the link for you in the show notes.

And so that's the kind of thing that reinforcement loading from human feedback, that RHF that you just described, is what I was talking about. I don't know, a few minutes ago, maybe five minutes ago, with respect to that training, which is often, I understand with these big LLMs, near the end of the training sequence. And it aligns the input that people provide with the outputs that come out and you end up with, like I was describing earlier, me providing vague context and getting spot-on answers seemingly miraculously, because I guess my vague thoughts and my vague context end up being about what other people in that scenario are also expecting. And it provided feedback on it.

Related to this idea of fine-tuning, something that I often do when I have guests coming up is I post on social media that the guest will be coming on, I provide a little bit of information. And so in your case, I was talking about it actually happened to coincide with the release of Gemini Ultra and with the renaming of Bard to Gemini. So my social media post a week ago at the time of recording, I mentioned that, "Hey, Google's largest language model, Gemini Ultra, is now out as of today, and next week I'm interviewing Lisa Cohen, Director of Data Science and Engineering for this release. What questions do you have for her?" And we got a huge amount of engagement on LinkedIn. There's still going to be more, but we already have over 80 reactions there, though I was surprised to see there weren't any particular questions on LinkedIn. And you and I were talking about this before recording.

I think it might be related to, and listeners, please feel free to pop onto my social media and let me know what kind of happens in these scenarios. But I think with someone like you, Lisa, what you do is so interesting and complex that probably when somebody comes across that as they're swiping on their phone for a break during their

workday, it might not be immediately obvious what a good question would be for you. And they're like, "Oh, let me just take some time to think about that and then I'll come back." But of course, this was just a social media feed, not something you put in your to-do list, and so tons of reactions on LinkedIn. The one question that we got actually came on Twitter, which is for me somewhat unusual, because I have a relatively small following there. But Mariya Sha, who was my guest on this podcast in episode number 639, she is well-known for running the Python Simplified YouTube channel, which has over 200,000 subscribers. And so she asked you, "What are the ideal system specs for running Gemini Ultra?"

And she says, "I have an upcoming tutorial about fine-tuning LLMs," the topic that you and I Lisa were just talking about on this podcast. So she has an upcoming tutorial about fine-tuning LLMs on the cloud, and she's still not sure which model to feature. And so this was an interesting question for me. I asked for clarification, but Mariya, I just asked last night. I haven't gotten a response yet from Mariya, but my interpretation of this was that Mariya might've thought that this Gemini Ultra release was a model weight release, which would've meant that she would be able to run Gemini Ultra on her own infrastructure. I am guessing that Google hasn't released these details, and you certainly don't need to say anything, but I just suspect that Gemini Ultra is enormous and that the idea of training or even fine-tuning something like Gemini Ultra on your own infrastructure might be an enormous and potentially expensive undertaking to try to do on your own. But there is Google Vertex, which streamlines people being able to do this themselves, right?

Lisa Cohen: That's right. Yeah. So through gemini.google.com, you can, through the prompting that we discuss, give specific instructions for how you want the model to respond and have some adaptation in that way. But I think for what she's describing here where you're developing an application and you want to call through to the Gemini model through an API and then build more upon that, then you would go through Google Cloud.

Jon Krohn: Yeah. And so through Google Cloud, specifically the Vertex application or module, I don't know exactly how these things get... I guess it's an application.

Lisa Cohen: Vertex AI.

Jon Krohn: Vertex AI. Yeah, it allows you to be fine-tuning, I guess, the whole family of Gemini models including Ultra for your own particular use cases, and then you can call them as an API and so then deploy them as part of an application that you're building, which is an amazing thing to be able to do. This is something that we do at my software company, Nebula. For Prototyping LLM capabilities, for getting generative AI capabilities up and running in our application as quickly as possible, relying on proprietary APIs like this through the Vertex AI platform application through Google Cloud is a great way to go and can allow you to get a huge amount of power in an application in no time.

Lisa Cohen: Yeah. So at the time of recording, currently as you go through Vertex AI, you'll be accessing Gemini Pro through that API.

Jon Krohn: Nice. Nice, nice. Cool. Well, maybe someday soon maybe we will have access to Gemini Ultra through there as well, but even Gemini Pro is a hugely powerful model. And then if you're able to fine-tune that to some particular task, at least my experience with taking even very small large language models, like in my case at our company, Nebula, we will often use a relatively small model like a 7 billion-parameter Llama 2 model. And then fine-tuning that for specific tasks, we can get performance that exceeds the very largest, most powerful LLMs out there that might have orders of magnitude more parameters in them, but just for that one specific task that we fine-tune them for. So I'm not even sure how many circumstances being able to fine-tune Ultra matters. Being able to fine-tune Pro could get you everywhere with all the cutting-edge capabilities that you could need for a given application.

Lisa Cohen: Yeah, I think like you see in other machine learning cases, it's the garbage in, garbage out, but in the opposite way of if you get that high-quality data, complete data for the scenarios that you're looking for, it's really powerful to the quality of the results that you receive.

Jon Krohn: Yeah, with fine-tuning for sure. So another consideration that people might have is they're thinking about a model like Gemini, particularly if they're thinking about building an application like we've been discussing, is safety. So I suspect that Google has rigorously tested

Gemini for safety. I further suspect that that is related to the delta of a few months between the announcement of the Gemini model family. So Nano, Pro, Ultra were all announced several months ago, but it wasn't until last week at the time of recording that we got access to Ultra, for example. And I suspect that safety testing was a big part of that.

Lisa Cohen: That's right, so safety is paramount for us. We've taken a stance at Google of being bold yet responsible, so safety has been a top priority for us throughout this journey and something we take very seriously in our launch decisions. So we go about it in a number of ways. For one, we measure and evaluate the model responses with various safety metrics, so we're measuring and evaluating the safety with each release. We also have a ringed release process to catch issues before they can have an impact broadly potentially. So we start with our team food internally, then our trusted testers, we'll run a live experiment to a subset of users. And so again, it's a careful rollout in that way. And then other safety practices such as red teaming, going beyond our typical tests, especially if there's a new capability or use case that we're releasing and doing targeted testing there for any potential vulnerabilities.

Maybe another scenario where this came up as we released to older ages first and then we later added releasing to teens, and that was, again, to have additional care to recognize areas that would be inappropriate to younger users and implementing safety features for that, having guardrails to prevent unsafe content such as illegal or age-gated substances from appearing to teens. And so there's a variety of things that we look for when we're evaluating safety for the product, from making sure that the response is not harmful, not offensive, not biased. We want to make sure we're reflecting multiple points of view on a topic. We also want to make sure it's factual and not misleading. We want to make sure it's not taking on a personality that we wouldn't want the model to have. And so we test all of these aspects and also test on different use cases to train the model on how to respond or how not to in different situations.

Regarding the factuality, we also have a double-check feature. So if you click the Google button at the bottom of response, it'll check the content and show the available sources for a citation as well.

Jon Krohn: That is a super cool thing to be able to do. And it's interesting to me, factuality a year ago was something that we talked about all the time with respect to LLMs. The GPT-3.5 release, for example, it very frequently was hallucinating, making things up, and that's not something that we talk about as much anymore, but I'm glad that you brought it up. I didn't even have it prepared in my questions. For me, at least, you guys have figured out how to make this work in a way that I can just rely on the results most of the time. Now, if you're listening to this podcast and you're trying to make a big business decision or any kind of big decision, you should be double-checking the outputs of a generative AI model. You can't guarantee that they are going to be accurate.

But with GPT-3.5, I was constantly checking information that I got, and now I seldom do. And so yeah, I'm glad that you brought that up. One term that you brought up with respect to safety that probably some of our listeners know, but we might want to dig into just a little bit more detail on, is this idea of red teaming. So if I understand that correctly, that is when you have people internally who are acting like hackers, like adversaries.

Lisa Cohen: That's right. Yeah. We'll have an internal team that simulates a bad actor and is just looking for pressure testing the system to look for any potential vulnerabilities that might be there. And so essentially, if there's any holes or risk, we want to find them internally before they might go externally, and so just helps make our process more robust. And as you mentioned that the factuality is a big focus for LLMs. Some of these safety issues are things that working on other products, we always have an aspect of data science and the product team looking around safety. Especially in this machine learning space, we're training the model on data that's out there and really reflecting society. It can have biased content, offensive content, and then we want the output of the result to improve beyond the broad public data that it's ingesting. And so I think that's where each company really gets to take an approach in terms of what are the values they espouse, what do they want to appear with their brand name, and so something that for Google, we definitely take very seriously as well.

Jon Krohn: And it's an interesting, the etymology of red team is interesting, which is that it comes from Cold War simulations run for the United States

military. And so the red team was the Soviets and they were going against the blue team, which was the United States. Interesting. Anyway, I digress. So we've talked about safety, we've talked about capabilities. One other interesting twist with a model like Gemini that is going to be used by Google users all over the planet, obviously the number one search engine in most countries, most developed countries for sure. So there's complexities there around language diversity and cultural nuances with AI models. What steps do you take to ensure inclusivity and accessibility for, say, non-English speaking users?

Lisa Cohen: This is a huge priority for us for the global product, so we expanded to over 40 languages and over 230 countries and territories in July. And then last month, we made Gemini Pro available for all of those languages as well. And so from a data science perspective, it's important for each release that we're not just looking at the overall metrics, but we also look to see if there's a particular cohort that's particularly deviating from the mean. So this helps ensure that we have the necessary training data per languages. Sometimes the UI might disproportionately impact some languages more than others. For example, looking at how things display with longer words, maybe in German or in Chinese characters, and so we want to make sure that the experiences we're enabling feel great in each language.

And just like with machine learning, to remove bias, you want to make sure that your training set is representative. So when we train on data sets, we want to make sure that we have a broad set of language, geographies, cultural aspects, and check the quality and performance by language. That's something that we're checking before we release and then invest if appropriate as needed to get everything to be an amazing experience across the board. And I think it also comes through with different features that we launch. So for example, with image generation, which we just launched recently as well, we really wanted to make sure that we had diversity in the images that we create. And so it was important that we trained then on examples with different racial and gender diversity, and that's something that was really important to us and the type of results that we create.

Jon Krohn: It's fascinating. Up until this point in the conversation, I'd lost sight of the international breadth and complexity of all of these different products that you have to work on, which just adds a whole new

dimension of wow to me. Thinking about the responsibilities that you and the people that you work with have as you roll out these products, it's so wild. If you're working on this huge scale every day with so many different geographies, so many different languages, so many hundreds of millions of people, probably billions of people impacted with these product releases, do you just get used to that or do you every day be like, "Whoa, this is huge?"

Lisa Cohen: I think it's definitely a responsibility that the team that we all feel. I think we feel it's also really empowering just to see how this AI can be so powerful and I think democratizing across the board, to bring information and capabilities to such a broad swath of users. If you think about just educational opportunities that can come through this and how it brings that so broadly across the world, so yeah, it's definitely a big responsibility and also really empowering.

Jon Krohn: Large language models are revolutionizing how we interact with technology. With companies rapidly adopting models like the GPT and Llama series architectures, the demand for skilled LLM engineers is soaring. That's why Kirill and Hadelin, who have taught machine learning to millions of professionals, have created the Large Language Models A-Z course. Packed with deep insights on tokenization, input embedding, transformers, self-attention, and LLM tuning, this course will help you gain hands-on experience with LLMs and stay competitive in today's job market. Enroll at superdatascience.com/llmcourse for your free 14-day trial. This course is exclusively available in the SuperDataScience community. You won't find it anywhere else. Once again, the link is superdatascience.com/llmcourse.

And so on that note of empowering and something that you must spend or I suspect that you spend some time grappling with in these product launches with things like Gemini and search platforms, even Assistant, there's the potential for highly personalized user experiences. However, the more personalized that you try to make an experience, the more that things could go off the rails, so there could be privacy or security concerns. So how do you balance that, getting this balance of personalization as well as safety?

Lisa Cohen: So overall, we want to make sure that we give users transparency and control. And so we have documented in our privacy hub, for example,

the disclosures around what's the data that we're collecting. So for example, when you interact with Gemini, Google will collect data such as conversations, location feedback, and usage information, and this helps us improve the product. So you talked about how you can give a very vague or brief prompt, and you're getting this amazing response that's exactly what you're looking for, and so all of that really helps enable the response that you're looking for. At the same time, you can easily turn saving your Gemini activity off. That's a setting that you have available right there through the UI. You can delete your Gemini activity from your account at any time. So, those are capabilities that Google takes really seriously and implements robustly as well.

Jon Krohn: Nice. Very cool. Yeah, and again, I'm completely wowed by the kinds of things that you're doing. Another strand following from the capabilities of models like particularly Gemini Ultra and the models that will come next, next year or the year after, there's a potential to revolutionize the way that we learn, not just the way that we access information. So do you have any insights into how AI-driven assistants could be improving outcomes in, say, educational settings?

Lisa Cohen: Possibilities are really limitless here, and it's pretty exciting to see the potential of what we can do. So one example I'm really excited about is the democratization of personalized tutoring. So currently, this can be really prohibitive due to costs for different audiences, but with AI, it really broadens access to this type of educational support. The other day, my son was working on some new vocabulary and there's all these apps and flashcards and things that are made online, and within a few prompts, I could just give Gemini the list of words and definitions and it could start quizzing him in an interactive way, so it's really very accessible and easy to use. And of course you could do that more broadly, asking the AI to teach you history to create a lesson plan for linear algebra, developing homework assignments, answer guides.

You can also have it adapt depending on which answers you know versus not. I've seen examples of people using it write and illustrate a children's book or make a personalized story for a specific child, so there's just so many directions that you can go with it, and it's amazing what it can produce.

Jon Krohn: It's so cool. It is so limitless. It's mind-boggling how we could, and increasingly, and maybe even with interesting philosophical implications, we could entertain ourselves endlessly. On any imaginable topic, it's like you're speaking with the best expert on that topic, and you can go into so much depth and point you in the right direction and point out fascinating things about that area, or just tell you entertaining stories. It doesn't need to be educational. It could be entertaining. And increasingly, these things are multimodal. So generating images, generating video to entertain or to educate. We are entering into a vastly different world from the ones that you and I grew up in.

Something that even as a small example of this kind of change that I've seen... I don't have any kids yet myself, but when I see kids these days, they expect all screens to be touch screens, and there's this level of interactivity. And generative AI takes that orders of magnitude further, where it could be the case a decade from now that you expect most objects to interact with you and be helpful to you and be anticipating your needs. It's a wildly, wildly different world that we're going into. I don't know. I don't really have a question. I'm just rambling.

Lisa Cohen: No, I completely agree. I think that our user expectations are evolving with each of these waves. There was the wave around personalization and recommender models, and you just expect all products to be personalized and know what you're looking for, and you don't need to completely explain your preferences each time, and so I think there was the personalization wave. And now with this generative AI wave and how it's being incorporated into so many products and experiences, I do think that's going to become an expectation going forward is that you should just have that efficiency, productivity, and for the products to really just help envision what you're trying to produce.

Jon Krohn: Yeah, it's wild. Endless possibilities. And yeah, hopefully a lot of listeners out there thinking up ideas for how they can be leveraging tools like Gemini Ultra, embedding them into products or services and creating efficiencies for businesses or people, creating entertaining experiences for people. It has never been a more exciting time to be someone in the data science field or in software development, or even in product decisions related to these kinds of products. And it is

accelerated, if anything, as this ecosystem evolves and gets more and more powerful. So cool.

My understanding from public resources is that Google aims to eventually integrate in the near term, so you and I were just talking about maybe a decade from now most devices being interactive and efficiency producing, but even in the near term, Google aims to integrate Gemini widely into products like Gmail. And we already talked in this episode about through what was called Duet, you get access to it in Google Docs, and we already know that through what was called Bard, you're getting access to Gemini through your Google search experience, so more and more and more integrations happening. Are there any particular ways that you see AI transforming people's daily use of productivity software in the coming year, the next few years?

Lisa Cohen: Yeah, so I think we've talked about staying in the flow, increasing the level of abstraction, and just increasing your productivity across all these tools. I think for me personally, I've really enjoyed some of the capabilities for things like creative writing. I think the idea that it can really help you build the thing that you're trying to build, but it gives you that creative inspiration, those ideas to spark and really create the best end product that you can. And so I enjoy, whether it's for blogging or if I'm about to speak about a topic to my team and just getting ideas, you can use it as your creative collaborator to cure writer's block or, again, just get additional ideas. Maybe there's a topic that I'm going to be covering and I have an outline, but I'll still ask Gemini and maybe it'll give me a fourth topic that it's like, oh, yeah, actually, I really did want to talk about that. I just didn't think about it at the time, so it helps remind me of additional examples.

If you ever work with a really amazing editor, they take what you say and they really bring it to life and what you really were envisioning, so I think it's awesome to have these tools to just realize your potential and express something so beautifully, just how you want to. As a fun anecdote, my family has a tradition of writing poems for birthday cards and Gemini is really amazing at that too. You can give it just details about the person and then you can incorporate to make it shorter, more casual, et cetera. It always has a nice tone to it. And you could use it for your podcast to brainstorm interview questions. So Gemini can summarize YouTube videos, blogs and articles, and

then you can use it to propose questions or ideas and get some unique ideas from it there as well.

Jon Krohn: Yeah, that is something that blows my mind that I didn't know about that I could be doing with Gemini, is being able to put in URLs for, say, YouTube videos or blog posts. And that is indeed a very interesting use case for me as a podcast host. Yeah, suggesting topic ideas, questions. Nice. Well, yeah, so all of these amazing capabilities, very exciting. Reflecting on our perspective, both of us, loving the huge capabilities that are being unlocked by generative AI today, you made a comparison earlier to personalization being this first wave, and now as having this wave of generative AI. Are there other moments, are there other technological shifts, that you've experienced in your career that you could compare this generative AI moment to?

Lisa Cohen: Yeah, I think another one that comes to mind was just the start of cloud computing, when that really went big. And I just remember this visceral feeling that the way we were doing computing was changing, and I just needed to be a part of that, and I think it's the same with gen AI. It's a new paradigm, it's a new way of working. And I think there's other parallels from that experience where I remember when I was working on Azure and we were talking to customers about moving to the cloud, it was two steps. First, you needed to understand why you would move to the cloud and work through various aspects around being comfortable with privacy and those aspects on the cloud and the public cloud versus private cloud, and just understanding what capability to provide. And then it was the process of choosing the vendor that you were going to go with, which company or product.

And I feel similarly with Gen AI. I think we're still in a state where people are still learning what you can do with this tool and capability. Even just prompt sharing is a thing and a community, and you get, wowed, saying, "Wow, I didn't realize I could use it for this." And so I feel like there's that first step of just even knowing how to use these tools, and then the second on diving into your scenarios as well.

Jon Krohn: Yeah, that is a very cool comparison. Yeah, so these big shifts that we've seen in tech in our lifetimes. Obviously the internet, that's a big one. And I guess even just personal computing, at least for me. So yeah, personal computing, the internet. Personalization was a huge one that you mentioned earlier. Cloud compute that you just talked

about now, and generative AI, it seems to me like generative AI is as big as any of these things. And it builds on all of them. They're all stepping stones, and who knows what this will unlock next? Very cool.

All right, so moving on from generative AI-specific questions or Gemini-specific questions, which obviously we focused on most of this episode, you have hinted at tremendous experience beyond just Google. So you've talked about things like Azure with cloud just now, so obviously you've worked at Microsoft. You've also worked at Twitter. And so with this experience of data science leadership at all of these huge data-centric, super successful tech companies, what kinds of challenges have you faced in developing data-centric cultures, and how do you overcome them?

Lisa Cohen: Yeah, so I think for any culture change, you write the vision to get everybody on the same page of what does that final end state look like, but then it's all about living it in practice, so getting your leadership team aligned, having good role models, having quick wins that are demonstrating that. I think for data-driven cultures in particular, it's really about helping have a shared understanding of the power of what you can learn from the data. I think when you're in a state of, "Hey, I have this really challenging problem that I'm trying to tackle, and how can data help me get to that goal?" then that's a good state to be, as opposed to, "I'm already choosing this path. What's the data that will support it?" And so I think overall, we try to engage on those big strategic questions and see what we can learn through the data to see how to guide the strategy in that way.

Jon Krohn: Very cool. So using the data themselves, using the data itself, to be able to make cultural changes. Something that you've talked about in the past is using the 80/20 rule to help data scientists balance simplicity and sophistication when developing models. Can you elaborate on that a bit on the show?

Lisa Cohen: Yeah. I think that sometimes we're looking at the most performant model and what's going to be the best quality solution, but I think that when we're doing this craft in industry, we also have time to market to consider. And so working on Gemini for example, if there's some metrics or solution that we're trying to figure out, and it's not going to be ready for six months, then I've lost out on six months of so

much will change within that time. And so I think having some simple heuristics still helps move the conversation forward, helps us better decisions that we did prior, and it actually has this really nice side effect that it's very simple and explainable to the organization as well, while maybe in parallel you work on maybe the more sophisticated solution as you go. And it's a great way, just like with product development, where you have an MVP and you get feedback along the way to try these things and practice and move faster to market.

Jon Krohn: All right, so that gives us a sense. You're talking about the 80/20 rule there. It gives us a sense of what we can be doing as individual data scientists and making the most of simplicity versus sophistication, getting great models out. But as a data science leader yourself, you are often managing highly technical people who are developing models, but then you need to be ensuring that you're getting buy-in from the rest of the business. So how can you ensure that the data science research that you're overseeing is perceived as relevant and is well interpreted by commercial stakeholders in the business?

Lisa Cohen: I think it starts with having that context, building that context around both the domain, the product, the users, and the business goals that we're trying to accomplish together. Presenting the work within that context, so this is a lever that we can now use to achieve this goal that we're all working towards. Using similar terminology. I think sometimes I've seen a gotcha around DS teams seeming more academic or it's research that might not seem as aligned. So as much as we can use that common language together with our partners, I think that always helps land it in the context that's understandable and relevant as well.

And then I think just promoting a sense of curiosity and proactiveness and ownership within the DS team so that we're the experts on those data sets, we're coming across interesting, maybe unexpected trends and raising those to the group so that we can really see across the board how there might be surprising insights that we might be able to learn towards our goals. I think particularly as these systems get broader and larger, the data can be really unifying to come through and see how things are trending. I think sometimes when you're starting with a brand new product, your sample size is smaller, you can actually just through asking users get a good sense of the pulse and how things are going. But I think as you start to have these

broader scales, you really want to understand those trends more broadly.

And I've seen, especially with more complex systems, places where we learn about our systems through the data. I remember there was this interesting insight where we were looking at just what are the top referral traffic to the paid subscription for Azure? And it ended up being from this error page from users trying to get the trial who had already had a trial before, and then they were getting this very obscure error and it's like, well, if that's actually a primary path for how users are coming here, let's make it this amazing white glove experience of, "Oh, hi, you're here, and now come and sign up." So yeah, I think it's always fun. Again, we have the context. We're trying to help drive this funnel and make it a smooth experience and then kind of using those data and insights towards that goal I think always works well.

Jon Krohn: Very cool. I love how you use data as a plural term. That really works well for me. So yeah, you now have a great understanding of how you think about data scientists making the most of their workday, how to be conveying data science research across an organization to stakeholders. One last concept I have for you, given your rich experience over many tech cultures, many tech organizations, successful tech organizations, is this idea of centralized data science teams versus decentralized data science teams. So, what is the difference between those and what are the pros and cons of each approach?

Lisa Cohen: So I think if you start with the decentralized model, some pros that you get from that, and this is basically where you have maybe an engineering organization and they have maybe either a small team of data scientists embedded within that org or some singleton DS where they want to make their work more data-driven. I think one of the benefits is that those folks are going to have that great domain context because they're working directly within the team that they're supporting on that area. I think one challenge that happens in that case is that then the data scientists don't have a peer group to socialize ideas with, learn from other approaches, that efficiency of maybe some central libraries or experimentation type of infrastructure data platform you could build together. And then the career paths get

a little bit trickier as well. It's not like you have a team that you can naturally rise within.

So I think once you go to the centralized models, there, you get all of those benefits with the sharing of best practices, efficiencies of scale, and the natural growth for individuals. But there, you have to be really intentional about getting those other benefits from the embedding. And so I guess the way that I generally like to design the org is that we centralize to a certain critical mass, and then within myself and then each of my leads and all the way through the org, each data scientist has a counterpart and a leadership team that they're part of. And so when you have that product, engineering, ease of research, design counterpart, then you're working together towards a common goal, sharing insights along the way, and you can work on broader, sweeping, impactful changes.

Jon Krohn: Very cool. Great insight there on the pros and cons of centralized versus decentralized data science teams. Lisa, I have thoroughly enjoyed all of this conversation today across Gemini, generative AI more generally, your insights into data science team leadership. Before I let my guests go, and you might know this because you have been listening to some of our episodes, I always ask for a book recommendation. Do you have one for us?

Lisa Cohen: Several, probably. I love books, but I'll go with Leaders Eat Last by Simon Sinek. Great author. And yeah, it definitely gave me some new leadership perspectives.

Jon Krohn: Very cool. And for people who would like to have your insights after this podcast episode, how should they follow you?

Lisa Cohen: LinkedIn and Twitter.

Jon Krohn: Nice. Classic choices. The ones that I use as well.

Lisa Cohen: I like to write on Medium as well, but I always link it from those sources, so yeah.

Jon Krohn: Oh, nice. All right. Cool. Lisa, thank you so much for coming on the show. As I've been saying throughout this episode, it's so incredible the things that you are doing and that your teams are doing, bringing

these incredibly powerful technologies to people in so many cultures, all of the world, free access tiers to most of them. And yeah, you're playing a invaluable part in this massive shift. Like we've had in the past with cloud computing, the internet, personalization, personal computers, you are playing a huge role in this big generative AI wave today. And so, yeah, on behalf of our listeners, thank you. I'm super grateful myself and I look forward to continuing to use Gemini Ultra and other Gemini models as part of my regular workflow.

Lisa Cohen: Awesome. Well, thanks for having me and thanks for hosting this great series.

Jon Krohn: All right. I hope you found it as much of a treat as I did to be able to hear from someone responsible for the rollout of a transformative, cutting-edge AI model to billions of people just as that rollout is happening. In today's episode, Lisa filled us in on the Gemini model family, including Nano for edge devices, Pro for general free use, and Ultra for the most cutting-edge capabilities. She talked about how RLHS stunningly aligns LLM models with the outputs we'd like a generative model to have, how you yourself can fine-tune Gemini models through supervised learning by Google Cloud's Vertex AI, and these super cool emerging impacts of generative AI across work and personal life, including unbounded creative potential, it acting as a personal tutor, and devices all around us being interactive and efficiency inducing. As always, you can get all the show notes, including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Lisa's social media profiles, as well as my own, at superdatascience.com/761.

And if you'd like to meet in person as opposed to just through social media, I will be in person at the Data Universe conference at the massive Javits Center in New York City on April 10th and 11th. I'll be giving a talk on generative AI, and we'll also be walking around, interviewing attendees to capture what you think of the massive conference. All right, thanks to my colleagues at Nebula for supporting me while I create content like this Super Data Science episode for you. And thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara, and Kirill on the Super Data Science team for producing another exciting episode for us today.

For enabling that super team to create this free podcast for you, we are deeply grateful as ever to our sponsors. You can support this show by checking out our sponsors links, which are in the show notes. And if you yourself are interested in sponsoring an episode, you can get the details on how by making your way to jonkrohn.com/podcast. Otherwise, please share, review, subscribe and all that good stuff. But most importantly, just keep on tuning in. I'm so grateful to have you listening and I hope I can continue to make episodes you love for years and years to come. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.