**Itamar:** 0:00

In the brain. As we humans or mammals really learn new tasks, we tend to invoke parts of the brain so as to learn this new task, while other parts of the brain remain dormant or inactive, and that makes a lot of sense. That's what allows us to learn how to ride a bike without forgetting how to walk. What it involves is sort of processing the information, looking at the geometry. Essentially, that's being represented inside the network. The network is this deeply layered billions of parameters, and so if you have a way of looking at which parts get activated as that new information comes in and you have statistics or information about which parts were called upon when the original data set was trained on, then you have sort of a chance at figuring out which neurons, which weights should be used, as in fine tuning, and sort of mask off or deactivate the rest. This is a super high dimensional representation, so it's hard to visualise. The underlying thing I would say is that these networks are really linearly partitioning the space. If you're able to understand how that happens and analyse these, what are called the fine transforms, these patches of partitions, then again you have a chance of figuring out which partitions should remain as is and kind of not move those boundaries, whereas in contrast to the ones that should be adjusted so as to learn this new information.

**Craig:** 1:15

Hi, I'm Craig Smith and this is Eye on AI. My guest today is Itamar Arel, a former academic and now a tech entrepreneur, who's been at the forefront of Voice AI Customer Service solutions. Itmar shares his journey from academia to the tech industry, detailing his work at TENYX, a company specialising in human-like voice AI experiences for customer service. Itmar delves into the intricacies of developing robust voice AI solutions, addressing the challenges of understanding and responding to human speech and the nuances of voice dynamics. I hope you find the conversation as enlightening as I did. Hi, good tech solves problems that you've thought about. Great tech solves problems that you haven't even thought of. What can the commerce platform trusted by millions of merchants do for you? It's time for Shopify, the commerce platform, revolutionising millions of businesses worldwide. Whether you're a garage entrepreneur or IPO ready, shopify is the only tool you need to start, run and grow your business without the struggle. Shopify puts you in control of every sales channel. So whether you're selling satin sheets from Shopify's in-person point of sale system or offering organic olive oil on Shopify's all-in-one e-commerce platform, you're covered. Shopify powers 10% of all e-commerce in the United States and Shopify's truly a global force powering all birds, rothes and Brooklyn and millions of other entrepreneurs of every size across over 170 countries Plus. Shopify's award-winning help is there to support your success every step of the way. Sign up for a $1 per month trial period at Shopifycom. Give them a try. They support us, so let's support them. Edomar, introduce yourself some of your background and what you got to Tenix and what Tenix does and what you guys are working on.

**Itamar:** 3:35

Yeah well, first, thanks for having me. I guess a little bit about my background. I still consider myself a recovering academic on most days. After getting a PhD in computer engineering, I, in a short postdoc at Stanford, took a faculty position at the University of Tennessee in the electrical engineering and computer science department, where I ended up doing 10 years of research in machine learning and AI, in areas specifically pertaining to reinforcement learning and what later became the field of deep learning. Mind you, this is just before the 2012 deep learning revolution, so a very different time. Much, much smaller community. And then, after that, eventually took a couple years of sabbatical back at Stanford, had a courtesy appointment, visited professorship at the AI lab and worked with Silvio Sovereza in a fairly large DARPA project. We had Silvio, of course, at the time was a CS professor and in the role of chief scientist at Salesforce and enjoyed all that. But at some point I realised that what I was really passionate about is taking research outcomes and maybe pushing them a little further and seeing what products and services they can empower to really impact a lot of people. And when you kind of reached out for recognition, you realised that academia might not be the optimal setting for you. So the uncommon you gave up tenure and left academia, moved to the Bay Area. I was an EIR, an entrepreneur in residence for a while at Ame Cloud Ventures, which is Jerry I think it's fun the guy who started Yahoo, of course, and then started a company called Apprentice in 2017, where the vision very early on became building these robust voice AI agents to automate the order taking process at drive through. So think Starbucks, Mcdonald's, all these chains that we know and love. The idea was to fully automate that drive through the ordering process. Interestingly enough, across what's called the quick service restaurant chain world, over 70% of the revenue comes from the drive through, which is a little mind blowing for most people that hear that first, but it was an interesting use case. Of course, this is pre LLM, pre a lot of the revolution or the advancements of the last two, three years. We ended up working with a fair amount of these, these well known chains, and in October of 2019, we were acquired by McDonald's Corporation. The CEO at the time believe that we on McDonald's views itself rightfully, as an industry leader, so they believe in this, this solution, this project, and they were interested in the team and the tech, I suppose, but also they wanted to establish a Silicon Valley Center of Excellence around intelligent automation AI. That was kind of the drive division of the CEO at the time and so that was sort of the pitch to us. How do you feel about being the founding team around which we'll build this centre? And it did unfold that way. We were about 20 at the acquisition and grew to about 100, about a third PhD is very focused on this voice platform solution. Really interesting and fascinating journey. This solution is now deployed in hundreds of stores. I think the plan is still to deploy this across 14,000 stores in the US and then go global to maybe complete that chapter Two. Three years later, under the leadership of the new CEO, there was a sense that maybe they're still very bullish about this initiative, but felt like maybe the right house for that organisation, which again was about 100 people at the time, was at a tech giant rather than McDonald's. So at the end of 2021, we were reacquired. The whole organisation joined IBM, became part of IBM Watson and it's still part of IBM Watson Now. Ibm, of course, offers this to all these chains, so spend some time there, but at that point if you're hit with a startup bug, then you kind of miss it and some of us felt like we missed being in a small setting, a small group that wants to do something new and change the world. So that sort of led me, or led us, to 10x about almost two years ago.

**Craig:** 7:39

Are you now competing with IBM and McDonald's, or is your solution different from what you were doing there?

**Itamar:** 7:47

Yeah, totally not competing. Totally different solutions. As I mentioned, really the last two, three years have been transformational in the field, right? So LLMs weren't around. Much of the advances around ASR, the speech recognition side, TTS, as well as really every part of the pipeline, is new and unrelated to that. I just showed you how fast the field is moving, of course, because it's only been a few years. So, yeah, we don't operate in the restaurant space and it's based on totally new technology.

**Craig:** 8:18

Yeah, before we go on to the new technology, what happened to the order taking with a conversational agent? I mean, I've never run across that in a drive-through.

**Itamar:** 8:33

Yeah, so McDonald's is based in Chicago and so deployments started from the Midwest that's sort of the FD centre and then it's expanding from there. It is deployed in hundreds of stores. I believe the plan I think it's not 100% sure, but the plan by early 2025 to be across 14,000 stores, which is the McDonald's stores in the US, and then explore going global. Mcdonald's has almost 40,000 stores worldwide and, as I mentioned, a good amount of the revenue of the order volume actually does come from the drive-through. So it makes perfect sense from a business standpoint. Yeah, it's live and kicking.

**Craig:** 9:10

So how does TenX's technology differ from what you were doing before?

**Itamar:** 9:16

Yeah. So, as I mentioned, there's been a lot that's happened in the ML world since then. Of course, the elephant in the room is the introduction of large language models, and what that really allows to do in a much easier way than it was before is understanding, introducing robustness to the NLU part of the pipeline, the natural language understanding. I always give this example. People can often go to the drive-through and say something like can I get a number two? Yeah, number two, a diet coke. So you and I know that was not two orders, that was just a completion of thought, that was one order. But in order back in the day, back in the day, quote, unquote 2020 and before, to try to capture all these million different ways people have of asking or ordering their food or asking for information, you really needed to either templatize the various ways in which they can ask it or simulate it. It was quite challenging, in fact, to get the long tail of ways in which people communicate. Llms really changed all that in the sense that you can actually try it on chat, gpt. If you say a customer said quote can I get a number two? Dot, dot, dot. Yeah, number two is diet coke. What did the customer ask for and in all likelihood you're gonna get one number two with diet coke right. And so, first and foremost, I think NLU has contributed to making the understanding, the robustness in the understanding, much more, much easier and quicker, and we did eventually get to automate the vast majority of orders at McDonald's, but frankly, it just involved a lot of manual work, a lot of laborious designing, when now LLMs can really help out with that.

**Craig:** 11:00

Yeah, so Tenix, when did you start at Tenix? Again?

**Itamar:** 11:06

So Tenix started in early 2022, almost two years ago.

**Craig:** 11:09

So you started with generative AI with LLMs, is that right?

**Itamar:** 11:15

Right, exactly, we knew about it slightly before it was fashionable and felt like that was one of the big things that can really change the field in a sense. So, Tenyx, we offer our customers what we feel as truly human-like voice AI customer experience, and customers tend to be mid-sized to larger customers that are interested in either completely automating the customer service or voice-based customer service or just the vast majority of it, and basically there's really a bunch of things you have to do. Ties into your previous question to do that. Well, I mean, I'm sure we've all had the experience of calling a hotel reservation or airline and you're usually greeted. If it is an automated system, you're greeted by a conventional system called IPRs, and these voice response systems tend to be very brittle. It's the ones that tell you you know, press one for that, press two for that. I think you said representative, and what winds up happening most of the time I'm sure you and I both experiences is 90% of the people actually press zero, zero, zero and just ask for talk to a person, and the strong sense is that now, in 2023, soon 2024, with the introduction, particularly with all these new technologies, we can finally build systems that can feel natural and flowing and robustly understand us and provide accurate responses to the point where it's just a much better experience and you don't press zero, zero. 90% of the time, maybe 90% of the time you're serviced well by these systems and that's sort of the vision.

**Craig:** 12:48

Yeah, a couple of questions. Are you guys, do you have your own Model or are you pinging a commercial model through an API?

**Itamar:** 13:02

Yeah, so we have to own our own models, partly because of latency issues in voice. Voice conversation is very sensitive to latency. You can't have like five seconds between Turns. It really needs to be below a second and a half, ideally even second, to feel natural. And so we end up taking Open source, cutting edge open source models and customising them in many ways to make them work all for us. So, yeah, we own our models.

**Craig:** 13:29

Yeah, and the other obvious issue is hallucinations. Keeping the model on point, do you guys use routine retrieval, augmented Generation or some strategy? Vector database strategy?

**Itamar:** 13:49

Yeah, so absolutely. I think the conventional wisdom is that to make the most of LMS, you do need rag, but you also need fine-tuning to make this thing work really well in any specific domain. I think that one of the relaxing aspects of the problem we're tackling is that, on the enterprise side, almost every company, the conversations that people call and talk about, tend to be within this restricted domain, right? 99% of the people that called an airline either book a flight or change a flight or cancel and so forth. So that makes it slightly easier, you know, in contrast to consumer facing products like Siri or Alexa, they're just super open-ended, right. You can ask about music, can ask about information or Of the web or the weather in San Francisco tomorrow. So you touched on hallucinations in our world, you're absolutely right. There's just almost zero tolerance to a solution. It doesn't adhere to business logic and rules. You can't say things that are not grounded in reality. You certainly can't say things that are offensive or otherwise inaccurate, and so we developed an approach that has multiple guardrails to mitigate that problem. Some of them rely on traditional rule-based schemes and others are deeply rooted in machine learning techniques, and and the Spirit of things is to try and characterise a normalcy profile, like a distribution, if you will, of what conversations in any particular domain look like, so that you can then use sort of anomaly detection style techniques to detect whether the conversation somehow beard off yeah, it is now off topic or something is, is, you know, not within the scope of what it should be. And so the multi guardrail approach, in our humble opinion, is the way to reach. You're never gonna reach zero, but you're gonna be arbitrarily close to a hundred percent accurate or grounded conversation that does not hallucinate. But this is an ongoing challenge and a problem for just the community at large.

**Craig:** 15:52

Yeah, and you specialise in One or two verticals, or is this a horizontal solution?

**Itamar:** 16:00

Yeah, so we started with travel and hospitality, hotel reservations and things of that nature, and then expanded to real estate, insurance, fintech, that. The nice thing we think about the particular approach We've taken is that it can fairly easily be applied across multiple domains and and that's that's exciting part of it it's powered, of course, by this architecture that incorporates LLM other has to do with, you know, the other pieces of the puzzle. I'll say one thing about fine tuning, because I think you mentioned that Part of the work we've done and we just had a piece and venture beat about. That is, you know, fine tuning is, as it sounds, a way of taking Some amount of information within a restricted domain, say hotel reservations. They take a bunch of conversations, of real-world conversations between an agent and a customer and what you'd like to do is take that so as to Polish, to customise an off-the-shelf sort of foundational LLM to be better in that domain and find you any just involves Changing the weights and, in effect, sort of continuing the learning so as to change the way so that again it's a little better in that domain, it captures the right nomenclature, responds in the prototypical way. The challenge we noted almost a year and a half ago when we started playing with these things, is that it's. It's not as easy or as simple as it sounds. When you change the weights and in the process of fine tuning you oftentimes end up distorting things that were there before. So you suddenly forget, forget knowledge, forget reasoning, and that's called. That relates to a well-known problem, machine learning, called catastrophic forgetting, which has been around for 30, 40 years. How do you continue to train while making sure you retain what was there before? And so we have a research team that sort of took that on because we realised that was a key challenge, something we had to solve, and came up with with a novel scheme that was that's able to do that, to Allow for fine-tuning while making strict guarantees that the knowledge and the reasoning and, more importantly, the RLA chef protection that was there before remains. Because one of the other things we we saw is that oftentimes when you find you, even on benign in domain Conversation, suddenly the protection that was there introduced by RLA chef Can be, can be removed or cracked to the point where either harmful or bias comments can be produced, and of course that's highly undesirable.

**Craig:** 18:27

Can you talk about that, that novel Approach you have published on the outer, because that's certainly a hot topic?

**Itamar:** 18:35

I think it is rightfully a hot topic. Um, I know we haven't published it yet. We are offering it as a beta service which people can sign up on our website. But to share a little bit about the sort of the spirit of the approach we've taken, it's a bit of a neuroscience inspired approach in the sense that there's Fair amount of consensus that in the brain, as we humans, or mammals, really learn new tasks, we tend to invoke parts of the brain, so it's to learn this new task while other parts of the brain so remain dormant or inactive, and that makes a lot of sense. That's what allows us to learn how to ride a bike without forgetting how to walk, and that's not been the traditional approach to continue learning or fine-tuning. And and we were sort of inspired by that In the sense that when fine-tuning, we have an approach that allows the network to look at the data coming in, figure out which parts, which neurons, which weights should be included, should be invoked, you know, tune those so as to learn as much as possible from this data. In this, in the specific domain, well, kind of keeping everything else as is, and that allows you to fine-tune while mitigating forgetting to get knowledge, reasoning and the RHF out of stuff. I'm sure there's more than one approach to do it, but we've come up with a framework that we think is very scalable, and that's been pretty exciting.

**Craig:** 19:56

Yeah, can you talk about the framework? I mean, what, what the yeah, what it involves? Is it offloading Memory to a knowledge base, or so not quite it.

**Itamar:** 20:09

What it involves is Sort of processing the information, the fine-tuned, the information to be fine-tuned on, and looking at the Activations, looking at the geometry. Essentially that's being represented inside the red network. This is deeply layered, you know now, through the billions of parameters, and so if you have a way of looking at which parts get sort of activated as that information, this new information, comes in and you have Statistics or information about which parts were called upon when the original data set was trained on, then you have sort of a chance at figuring out which neurons, which weights should be used as In fine-tuning and sort of mask off or deactivate the rest so as to retain what was there before and in the sense that's that's why there's a deeply mathematical approach to to look at the geometry of what's worth being represented, what was represented before and what's being called upon or represented as this new data comes in.

**Craig:** 21:05

Well, it's classic, so it's mathematical. You're not. You don't have a Visualisation of the network that no, that's a choppy.

**Itamar:** 21:12

This is a super high-dimensional representation, so it's hard to visualise. But the underlying thing I would say is that you know, these networks are really linearly partitioning the space, right? So if you're able to understand how that happens and analyse these, our work all works fine, transforming these patches of Partitions. Then again you have a chance of figuring out which partitions should remain as is and kind of not move those boundaries, whereas, as in contrast to the ones that should be adjusted so as to learn this new information and that's maybe a little technical, but that's what the yeah.

**Craig:** 21:52

Yeah, that's fascinating and the fine tuning of the data that you use for the fine tuning Is is yeah, where are you getting the data sets? Are there data sets or customer interaction data sets out there that you can License or that are open source, or do you have how you build up your owner to use synthetic data?

**Itamar:** 22:21

all of the above. I Think the best approach is to use real data, even a little bit of real data, and have a way of taking that real data and using a high-performing LLM like GPT-4 to generate synthetic data that covers that space, that distribution, right? So some way of taking real data which you always have less than you'd like, and having that be the, the, the core, to Generate additional synthetic data and then train on all trains and evaluate on all that. We think that's probably the best approach at the time and that's generally the approach we've been taking.

**Craig:** 22:58

Yeah, and that's the training data. Is text, not voice, right? I mean that the voice to text and the text to voice on either end are Not part of the inference.

**Itamar:** 23:14

Right. We partner with other companies to do the speech to text, the ASR, and the text to speech. There's a lot of companies that specialise in that. I guess the one thing I would say is, in additional to the transcription to the text, which of course is the core of what we're Processing, we do extract non-fanatic, non textual information that oftentimes gives you some just an additional cues as to what the customer was saying, and in particular, it pertains to the other thing I think that's key to getting this kind of system to work Well is to handle the just the dynamics of the voice Exchange well. We speak very differently than we do. Then we write or we text. We tend to use poor grammar and broken English and ums and ems and repetitions, but somehow, you know, six or seven year olds can handle all that almost naturally, and traditionally it's been challenging to build machines that do that, that do that well, right. But, and so we really focused on that because we think that's key to making people engage and having a flowing conversation, as opposed to having people press 0000 and ask to talk to a person. Some of it has to do for example, you should have alluded to that to detect when is the person talking, kind of finished saying what he or she wants to say Versus when? Are they just midthought, right? So somebody can imagine a hotel reservation use case where the customer says, well, I'll be checking in on you know. Then he or she kind of checks their calendar and Three seconds later so on the 15th and departing on the 18th these are things we say every day, right, but understanding or detecting that's called end-pointing prediction. The world of voice processing, that becomes critical because if you somehow start processing half that statement, you're likely to produce garbage and the whole conversation can go the wrong way. Other times, when somebody just asks, well, what's your checkout date or a checkup time, you do want to detect that really quickly and say check out, is it noon, right? So those kind of things Dealing with situations when one side speaks over another, we humans again have this tendency to To figure out how to resolve that, say I'm sorry, go ahead, or understand whether the other side Finished hearing what I was saying, or do I need to repeat myself? All these nuances of the voice dynamics are really really critical to To get these things to not be brittle and really be engaging and flowing, and that's a big part of what we're building, hmm, the the market of.

**Craig:** 25:41

I mean, since GPT-4 certainly has kind of filled up with all kinds of solutions and I would guess Tenix's market is as crowded as any of them, how do you differentiate yourselves? How do you carve out market share? Are you? Is it largely on the strength of your past relationships? Is it through marketing? Just curious, for a solution like this, because I mean, I'm very, I mean I'm a very different and much less sophisticated Activity, but I get pitched daily. There must be At least five oftentimes more pitches from LLM-powered marketing companies of one sort or another. We can grow your podcast to 10X and it's just, you know, and I'm sure each one has its merits, but there are just so many. So how do you handle that?

**Itamar:** 26:56

Yeah, you're right. I mean, even in the customer service automation space there are a lot of companies. The majority of them are actually in the text realm like chatbots for e-commerce and like that's sort of the natural immediate use case Voice, as I mentioned, voice is more nuanced, voice is challenging. We speak very differently than we write, and getting to manage voice conversations well is a non-trivial task. Maybe that's part of the reason why that subspace, like voice AI, customer service automation, is far less crowded. I'm sure it will get more crowded in the future. And, to your point, we do feel like we bring five years of experience in building voice AI solutions in the real world with our McDonald's adventures, and so a lot of those learnings do crossover in, particularly around making the conversation feel natural and recovering well. I guess I would say there's four things you gotta get right to build an efficient and engaging voice AI solution. One is robustly understanding what the customer said, and that has to do with everything we talked about, the broken English and the poor grammar and so forth. The other, of course, is producing accurate responses. You do want to respond when you respond, you want to respond accurately, and that has to do with incorporating RAG and fine tuning, and I would even say fine tuning in the context of RAG. Rag involves retrieval. Augmented generation involves having a semantic embedding space that allows you to then search a knowledge base and dig up the right pieces so as to respond to the customer. But that relies on that embedding space being accurate and differentiating. And so there's work, and so we certainly do that in fine tuning even the RAG piece, so that we really make sure we capture the right representation, dig up the right pieces out of the knowledge base to respond. So that was number two. Three, as I mentioned, you really need fine tuning, end to end, to do well in any restricted domain and to mitigate things like hallucinations and adhere to the business logic and rules which is critical for any enterprise. And finally, everything has to do with optimising the speech dynamics, so barging and end pointing and dealing with acoustic distortions. All of that is non-trivial and luckily yeah, specifically the last two, three years have been transformational in offering these new pieces of the puzzle on the technology side that we strongly feel now you can build the solutions that we've been promised for 20 years that are robust, engaging, flowing and create the right customer experience. We're obsessed with that. I think it all starts and ends with really delivering a customer experience that feels like you're talking to a person.

**Craig:** 29:44

And the different industries that you're looking at, do you fine tune the same model? So let's say you're in hospitality. You fine tune the model for that industry. You have a customer in that industry. You take all of their data documentation if it's a restaurant, menus and that sort of thing, put it into a vector database and use retrieval, augmented generation for the output. And then when you move on to I don't know another industry, any other industry, do you go back and fine tune the same model, being careful not to catastrophically forget the previous fine tuning, and then plug in a new vector database for whatever customer there is in that industry. Or do you have to build, create a new instance of the model and fine tune it for each industry? I mean, can you fine tune, is it one model that keeps getting fine tuned as you add industries, or do you have a row of models, one for hospitality, one for car repair, one for whatever?

**Itamar:** 31:27

Yeah, so it's more of the latter. Every industry, we found that you could fine tune for that industry and the small differences between, say, Hyatt and Hilton and Marriott. I really usually don't like complete fine tuning, right. But across industries, yes, you do need to specialise. Each has its own terminology, nomenclature, just dynamics of how typical conversations unfold, and so you do need to do that per industry fine tuning. Having said that, one of the things that, rightfully, any customer is very sensitive to is contamination I think you alluded to that contamination of their private little long PII information, and so we take extra steps to make sure that anything that is unique or private to any customer is never used across even customers in that same industry.

**Craig:** 32:17

But conceptually, yeah, every industry would have its own sort of fine tuned macro model and you talked about fine tuning the vector database or the RAG system or mechanism or whatever. I haven't heard anyone talk about fine tuning on that side. What does that involve?

**Itamar:** 32:40

So it involves the same ingredients, if you will, you want to have typical or prototypical conversations in that specific domain. So that usually means you take whatever actual real world conversations you have and use high performance LLMs to enrich that and, with synthetic, to augment that with synthetic data. And ideally you want to take all that and use your embedding model if it's a bird or a similar model and essentially fine tune like, continue to train that model on that data. So it's really kind of a polishing. It's not that you're training everything this whole embedding from scratch. You want to leverage everything that that model knows about general language and semantics and so forth, but you want to refine it such that that vector database becomes more accurate, more differentiating for that particular domain. It's not always needed, but a lot of times it is and we found that to be beneficial.

**Craig:** 33:42

Yeah, how do you evaluate these models once they've been fine tuned and you've got your rag working, what's the process for evaluation? Evaluation, because you can't just turn it on and try it out. You know, even with a Dozen people in the office and oh yeah, it worked great. I mean, if you're gonna put it in production, you need some robust evaluation to make sure that it's catching the long tail, or it's, or you know where in the long tail that drops off that sort of thing?

**Itamar:** 34:28

And the other reason why you don't want to just test it in books in the office is because you naturally become very biassed and you're not an objective tester, right? The answer to that always has to be some combination of anything you could do in terms of automated Systems, scripting and dynamic testing. And then to your point we do have external Testors companies. There are companies that you can hire that Sort of call upon people to test this and generate reports of things that work and things that don't work. So you, to your point, you have to do that, and you have to do it with People that have not played around with it a hundred times. Otherwise they tend to be too biassed and not really reflect what a typical customer would interact with like on the automated side. That is all another very active space of applied research and engineering. Obviously, the general approach is to take again high-performing LLMs that have been engineered right to play essentially both parts the virtual customer and the virtual agent and Try to capture as much as you can through that interaction. Of course, voice again injects another layer of complexity, because text only Is very polite and clean, turn taking and so forth, and voice, you have interrupts, you have everything else we talked about so that we have some ways of automating that. But again, always rely on a complimenting, actual Sort of mechanical Turk, like, if you will, testing that's independent. Otherwise you're right, you're bound to Discover the bugs when it hits the customer which you don't know and and so on, the external testing firms.

**Craig:** 36:12

They have armies of Freelancers, I imagine, out there, sort of like a mechanical Turk, as you said who you know, ask, interact with the LLM, and then all of those interactions are are captured and collated and and then you can See where you're weak or you're strong and that sort of thing. Right Do you? Do you have an LLM that's playing the customer?

**Itamar:** 36:44

Yeah, yeah, you have both, and you Ask them to interrupt each other, at times with some probability, and you ask them to Maybe ask questions off topic. And, yeah, there's just the whole rich world of how do you design these simulated customer and and virtual agent interactions, which again helps you oftentimes discover many of the the bugs or the issues, but it's never a hundred percent. It's always highly recommended to.

**Craig:** 37:14

And is that? Are those third-party companies or their companies that provide that service? They have conversational LLMs, train for evaluation of other LLMs and I would imagine you set it up and let it run for two days or however long, and it just accumulates this body of Data that then you can analyse. Is that third party or done in-house?

**Itamar:** 37:47

So it's done in-house. I wish there was a third party that might be a good idea for a business down the road. It requires linguists and ML engineers to put that together and at this point we did all that in-house and how long would you let a system run Against itself like that or in that evaluation? Yeah, usually would entail tens of thousands of virtual Conversations to get a good signal as to whether yeah so what are the issues that you're able to capture that way? and then Real conversations at that point could be in the order of hundreds, which doesn't take that long to run through. Yeah as you mentioned, you have recordings, you have transcripts and you can learn what worked and what did that way.

**Craig:** 38:31

Yeah, and is the system once it's live? Is it Training on on the interactions as it goes forward, so that it's continually refining itself, discovering New corner cases?

**Itamar:** 38:50

Yes, but not in real time. So you do run it and you collect information and you periodically use that to improve the system, after after running exhaustive Evaluations and testing to make sure that you've improved what you thought you've improved, but you didn't break anything that was working before, and then yeah, and then you sort of have the next release. So it's kind of a cadence of these updates that does build on, of course, real conversations that have been collected. The other thing may be worth mentioning is that Obviously there's a cost reduction value proposition to all that, but beyond that, I remember that was strongly the case with McDonald's. What is really exciting about this kind of solution is that it can really improve the customer experience. I mean not only are you not going to wait 23 minutes on the line sometimes until the person picks up, which is not great for the customer, not great for the brand perception, but you can do things like a B. Testing companies sometimes want to say, well, we want to respond this way to a question, not that way, and it's a challenge to convince 5,000 people somewhere in the world to do so tomorrow. It's much easier to do that with machine Analytics. You know, these are the kind of things that with humans, just objectively very challenging to do and these solutions offer so much to Our clients, not only to the enterprise, but also to the customer him or herself, so it's really transformative.

**Craig:** 40:10

Yeah, so you guys started in 2022 and you know you're not very far in this, but we're not very far into generative AI, you know. Revolution revolution, if you want to go. Yeah. And in looking back, I mean to what you were doing before Gen AI, you know a lot of those solutions I imagine look Kind of crude or primitive from today's point of view. What's your expectation in, in how your business and products will evolve in the next five years and what do you think will change?

**Itamar:** 41:05

Yeah, so we're very bullish and optimistic about ISIS, but yeah, it's not quite clear whether it'll continue to advance at the pace that it has in the last couple years, but there's so many people working on this stuff that it's hard to believe it'll slow down anytime soon. I think all the pieces will get better. Not sure how familiar with the latest text to speech, the speech synthesisers. There's many companies, including, of course, some of the big ones, that have offerings and it's gotten to the point where it's really indistinguishable. You hear somebody, somebody you know that, whose voice has been cloned, saying things that they've never said, that the Untrained ear certainly can't pick up on. So I think things like that will continue to improve and become really indistinguishable. The speech to tech or sorry, yeah speech to text has continuously improved and that's that's gonna continue here from our end. There's interesting challenges, interesting challenges. I mean. Frankly, some of what it takes to get voice AI to work well is almost AGI complete problems, but we think that there'll be improvements. One of the things we're working on is trying to train LLMs to not just predict the next word or the next token, which is generally how it's been done so far, but try to look over some horizon and optimise over that. So slightly less myopic optimization, if you will, we're not the only ones working on that. Humans do that very well. And then some of the criticism that's been levelled at LLMs is that they do a lot with this trick of predicting the next token. But what if you're able to, like we do, kind of look a little bit further out and produce a token or an action that maximises some return over some horizon? So obviously, reinforcement learning is all about that. It's about solving what's known as the credit assignment problem. If you have some horizon, something good or bad happened, can you go back and attribute credit to it so as to try to repeat things that work well and avoid things that didn't. So I think all that's going to enrich LLMs and it's not just going to be language. As you probably know, there's a lot of these multimodal models coming out now and they're all going to be stronger and better. I think some of the challenges have to do with implementation, the GPU shortage. There's been interesting papers recently that ask the question: can some of this be implemented in CPUs, which will cost a lot less and reduce the shortage problem around GPUs? So I think those are some of the exciting things that are over the horizon.

**Craig:** 43:42

Yeah, yeah, yeah. It's really a remarkable time, hi. Good tech solves problems that you've thought about. Great tech solves problems that you haven't even thought of. What can the commerce platform trusted by millions of merchants do for you? It's time for Shopify, the commerce platform, revolutionising millions of businesses worldwide. Whether you're a garage entrepreneur or IPO ready, shopify is the only tool you need to start, run and grow your business without the struggle. Shopify puts you in control of every sales channel. So whether you're selling satin sheets from Shopify's in-person point of sale system or offering organic olive oil on Shopify's all-in-one e-commerce platform, you're covered. Shopify powers 10% of all e-homers in the United States and Shopify's truly a global force powering all birds, rothes and Brooklyn and millions of other entrepreneurs of every size across over 170 countries. Plus. Shopify's award-winning help is there to support your success every step of the way. Sign up for a $1 per month trial period at Shopifycom slash IonAI. Give them a try. They support us, so let's support them. That's it for this episode. I want to thank Itmar for his time. If you want to read a transcript of today's conversation, you can find one on our website IonAI. That's E-Y-E-O-N-AI. In the meantime, remember the singularity may not be near, but AI is changing our world, so pay attention.