SDS PODCAST EPISODE 1031: TOKENOMICS: WHY YOUR AGENTIC AI BILL IS EXPLODING (AND HOW TO FIX IT), WITH TYLER COX AND ISH SHAH

Show Notes: http://www.superdatascience.com/1031

Jon Krohn: 00:00:00 Over a single weekend, one of today's guests burned through two billion tokens building a video game for his wife. He and his colleague are here to explain why agentic AI bills are exploding and how a box under your desk can cut them by up to 93%. Welcome to another episode of the Super Data Science Podcast. I'm your host, Jon Krohn. Today I've got two returning guests, but they are on the show together for the first time. Those guys are Ish Shah and Tyler Cox. Both are distinguished engineers in the office of the CTO for the client group at Dell Technologies, where they work out how to run powerful AI models on the machines closest to you. In this episode, they dig into tokenomics, why agents and their subagents devour so many more tokens than chatbots ever did, how to pick the right model for the job, and how moving agentic workloads off pay per token cloud APIs and onto your own hardware can pay for itself in as little as two months.

00:00:56 Enjoy. This episode of Super Data Science is made possible by Anthropic, Origin, Palo Alto Networks, and the Open Data Science Conference. Ish and Tyler, welcome both of you. Back to the Super Data Science podcast you've been on separately with different guests, never together. I think this is a dangerous combination. And I think everyone knows why. Let's start with Ish on why this is so dangerous to have both of you together.

Ish Shah: 00:01:23 Good to be back, Jon. Thank you for having us back. I think that Tyler and I have spent the last year out of our CTO office at Dell working on a lot of the things that we talked about the last time I was on with you, Jon, with our friend Shrish, and Tyler was also on with Shrish. And our job is to think about the client device. And I remember the first time we talked to you about client devices, you're like, "Hmm, laptops?" Yes, yes. All end user compute. And I think the role of those devices has changed a lot over the last, call it 12 months since we last talked to you. So just as a refresher, Tyler and I are both

distinguished engineers in the office of the CTO for the client group at Dell Technologies, and we've got a lot to talk about.

00:02:14 So I hope you're ready. I hope the audience is ready, Jon. I think

Jon Krohn: 00:02:18 We've got

Ish Shah: 00:02:19 A lot to cover.

Jon Krohn: 00:02:20 Tyler can confirm this or maybe Ish can confirm this about Tyler, but we should let Tyler speak, which is that Tyler is full of detailed facts about things. And then Ish provides great color commentary.

Ish Shah: 00:02:33 That's what they pay me the big bucks for. And by big bucks, I mean not that.

Tyler Cox: 00:02:37 Yeah, we do a little bit of play by play. I think last time we were talking about linear attention and the rise of hybrid and state space models and we've only gone from there.

Jon Krohn: 00:02:44 Oh yeah. That's right. That was fun. That was great. We'll have links in the show notes to previous episodes that Ish and Tyler have been on separately and they are both exceptional. This one, multiplying them together, as I said, dangerous combination. It might be too much for one podcast, but we're going to try to make it happen. We're going to do our best to contain the danger.

Ish Shah: 00:03:05 What was it you used last time? The ishiness of it all?

Jon Krohn: 00:03:09 Ishinish. Yeah. Ishinish. It's a tongue twister. All right, let's get into the technical content here. So this episode is about agentic AI, tokenomics, solutions to get you better results faster, cheaper. But for those of our listeners who aren't totally sure, or maybe it doesn't even hurt to get this definition back to you every once in a while, what

makes an AI system agentic versus a traditional model or a standard chatbot?

Ish Shah: 00:03:42 I think it's the ability to use tools. And we actually have this conversation a lot. I mean, even within a company like Dell, there's a very broad range of folks and how deep they've gone on this technology. And I would actually say as a company, we're pretty far ahead of a lot of others in terms of the baseline AI literacy. That phrase agent still confuses the heck out of people. If I'm using ChatGPT on my app, is that an agent? If I'm doing something like what Tyler does day in and day out, which looks like the matrix flying across his screen, is that an agent? To me, and I think to a lot of people, an agent is when you take the LLM and you give it abilities that allow it to kind of break out of the tab, so to speak. It can act beyond the tab.

00:04:32 It can do things with other data that lives elsewhere. It can do things on your computer and move your mouse around and click on things for you. That to me is the line of agentic versus, hey, I'm having a conversation with an LLM and it's like writing a poem for me versus go check my inbox for this esoteric piece of data that I saw once six years ago and then do this other thing with it on this other website. That's the difference.

Jon Krohn: 00:04:57 Yeah. I like the definition of tool use in a loop to achieve a goal for agents because I assume probably most listeners have a vague idea in their own minds of what an agent is. So we can probably move swiftly on from that to how widespread are agents these days. Let's say in enterprises, how many enterprises are using agents?

Ish Shah: 00:05:21 Over 90% in our experience, and according to Signal65, which is a research firm that Dell does a lot of work with, we are seeing over 90% of enterprises in some way, shape or form have deployed agents into day-to-day use or production use. And they define these things differently, obviously, enterprise to enterprise. But at this point,

unless you have a literal reason that you wouldn't or couldn't, I think the thread that's been pulled through is a lot of folks have been using this stuff in their personal lives. And I think they show up at work and they bang on the door of Mr. and Mrs. IT decision maker and they say, "Hello, I would like this thing please for work purposes." And I think that that's how that number got so high so quickly, Jon. But Tyler also spends a lot of time with customers, Tyler similar vibes with some of our big accounts.

Tyler Cox: 00:06:17 Yeah, definitely. And what's interesting is the number of different ways we see agents getting used. Software development is obviously a huge use case for agents having the ability to go off and build new tools and software. But we're seeing through our customer engagements, a lot of really interesting and exciting different uses of agents. You're doing lots of research tasks, you're doing lots of knowledge work, aggregation of data, creation of reports and things like that, but you're also seeing more domain centric pieces in healthcare and finance and other pieces.

Jon Krohn: 00:06:54 Really exciting. There is a huge amount of potential. I feel like we're only scratching the surface of what can be improved in organizations, whether they're enterprises or not. And I think a lot of our listeners will have that experience. When you're using Claude Cowork, CloudCode, lots of desktop tools, OpenAI products, Grok products, there's tons of different options out there for you to be using these kinds of tools personally. But as you alluded to there, you can run into a lot of roadblocks within organizations from in particular the IT department. There's a joke that I heard a while ago, which is probably rude to IT folks out there, sorry, IT folks listening to the show, but there's this quip that if IT could remove your keyboard, they would.

Ish Shah: 00:07:42 And it makes sense, right Jon? These are the people you ask to protect these systems and keep them running at 99.999% uptime. They got to keep the org moving, they got to keep the org going. And it's a risk first mentality because it has to be. So there's this very organic push and pull happening inside of enterprises right now where from the top down, take a CEO who read a thing in the Wall Street Journal to the bottom up, take a independent contributor who just finished their internship. Those two folks are, "Holy cow, look at this thing that I did." And they're sort of converging on the IT person in the middle. So the IT person from the top down and the bottom up, there's signal in the noise and they're like, "Okay, how in a world where I am being told to do more with less every single year, smaller budget, fewer people, but keep my 99.999% bulletproof uptime.

00:08:42 I don't want anyone to lose a day of productivity." How am I supposed to reconcile that as an IT leader with the reality that is AI in the enterprise? How do I do those two things? And for us, we spend a lot of time talking to our customers because this is what keeps them up at night. And the answer is that you have to do a really good job of explaining the types of things we're talking about on this podcast here to your superiors, to the folks in the C-suite of this is not your grandpa's IT. The name of the game is changing and the costs that come with that sometimes are not ones that people have been used to for the IT department spending. So IT procurement, pick your poison where it lives within your org. For

Jon Krohn: 00:09:29 Sure. Speaking of spending, we're going to talk about tokens in just one moment, but before we get there, apologies to Tyler that I'm letting Ish talk way too much here and I'm going to give him even more floor. Though there might be some aspect of this that you have to add to this, Tyler, that we didn't talk about before hitting the record button, but just before we hit the record button, we were talking about how. Well, so Ish brought up on

screen, he said, "Oh, I need to bring up my buddies." And then he said, "Okay, here's my Grok buddy, my codex buddy, my Claude buddy." And I said, "What? What was going on on your side, on your monitor when you were saying those words, Ish?"

Ish Shah: 00:10:05 I mean, a workflow is such a sensitive thing. It's such a personal thing. When it's time for Jon to do work or Tyler to do work or Ish to do work, there are things we reach for. You reach for your favorite pen, you reach for your favorite mouse, you reach for. If you're my wife, you reach for your favorite snack, right? It's like depending on where you are and what industry you're in, your workflow's going to look different. But my workflow has changed so much over the last 12 months because it's almost like I'm on a construction site and I'm constantly reaching for what is your handiest, handiest tool? And there's a very nice, these are called multiplexers, but to normal people, they take lots of screens and they put them into one screen. It allows you to run these different agents side by side because what my workflow now looks like a lot is I'll kick a task off in one of these panes and my little buddy, so to speak, will go run off in the world and try to do the thing I've asked it to do.

00:11:06 Well, I don't just want to sit there and twiddle my thumbs. I want to move on to the next thing. So I'll spin up the second session and set it off on the second task and the third and the fourth and so on and so on. And your workflow is so important because what really gets you is when one of these sessions hangs and you kicked it off in the morning and you come back at noon and you look at it and you launched it at 9:00 AM and it's been stuck since 9:01 asking you to approve this itty bitty prompt or like, "Hey, are you sure you want me to do blank?" That's why it's so important to get your workflow set up right. It's so you don't burn. There's a lot of productivity that happens when humans are away from keyboard now and my setup is a big part of it.

	00:11:52	So hence the little buddy, but it turns out when you've got four of them running at once, that starts to put new kinds of loads onto your system and there are all these downstream effects of what a modern workflow for a person looks like.
Jon Krohn:	00:12:05	We're going to get into those downstream effects in a second. And I did also now think of a great question, Tyler, to ask you related to this conversation that is going to lead us to tokens momentarily and token usage. But really quickly before we get there, one last one for Ish, which is you alluded to before we started recording that you have rules of thumb for why you choose a particular buddy for a particular task. So why do you, for your usage on your personal, well, I guess your personal work machine, if that makes sense, or your personal machine, whatever, how do you choose when you're going to use Claude or Grok or Codex?
Ish Shah:	00:12:45	So they're all good. It's like if you're our boss, Tyler and my boss, hello Mr. Rob Bruckner. If you are Tyler and my boss and something crosses your desk, you're going to think to yourself for a sec, is this a Tyler job? Is this an ish job? A Shiresh job? Who does this go to? It's the same for me. I have opinions and preferences on if I want speed, for example, of late. Grok seems to be fastest. If I want depth and thoroughness, for example, right now Codex is one I trust a lot to go tackle these. It's like a Rottweiler. It just goes after the thing until I tell it to stop, sometimes to a fault and it burns all my tokens and I have to sit quietly for a couple hours until it resets. And then Claude, which arguably Claude is. If ChatGPT was responsible for that first inflection point in bringing people into this world of LLMs and AI, I would argue Claude and Claude Code over the last 12 to 18 months are, it's one of the biggest drivers of adoption into the enterprise.

	00:13:57	So for the things Claude is good at, I turn to Claude and Fable 5.1, which is their latest model, Anthropic's latest model, tends to be at the top end of a lot of the leaderboards for various kinds of tasks. So artificial analysis is a very good website. If people haven't checked it out, they should. Very easy to understand rankings and benchmarks. And if you have an image task, you might want to consider this one, but people also need to be mindful of benchmaxing, which is all of these labs are now training their models to beat the benchmarks against which they are tested. So it's a lived experience question and that's how I determine. For
Jon Krohn:	00:14:35	Sure. And as an enterprise at least, this would be probably, it might be overkill for an individual, but to get around that artificial analysis or just general benchmaxing that all of the vendors are doing, all of the model creators are doing, the best thing to be doing, if you have some repeated task that you're going to want to be automating in your organization, you've got to have a great eval for that. And then that's something that's internal to you and you can be pretty confident the big labs aren't trying to optimize for your particular internal task. And
Ish Shah:	00:15:09	Jon, to that point, I'm going to turn this one to Tyler because Tyler has built along with several very talented technologists within Dell, a model evaluation protocol framework that we use internally that when these models come out, particularly open source models, open weight models that are capable of running on Dell devices locally. Tyler used to, in the beginning, be up late at night trying to benchmark all of these like, "Hey, how does this work? Does this work well?" Tyler, you got to tell the folks about MEP because I think this is Dell's secret sauce, Jon. This is how we make sure models are going to run great on our stuff out of the box. There's a whole army of people and agents working on it. Yeah.

Tyler Cox: 00:15:51 And so we hit the three million model mark of public models on Hugging Face within the last month or so, and that's a big milestone. Each of those models are well suited for different things. And so we decided 18 months ago that we needed a more systematic way to do that. Today, that capability for us means that I can tell an agent, "Hey, there's this new model out there. Go run it on these 10 platforms and tell me where it fits on the Purrito curve so that when I go talk to a customer in financial industry or in a healthcare industry, or should I buy this system or this system?" I'm coming in with, "Well, here's what I see from the data. If you're operating point, you need to put 10 users on this thing, here's what you need. You need this model on this platform for this use case." And so what Ish is talking about, that is what we are bringing in across the board.

00:16:41 It's not just how smart is the model, it's how well does it fit for your use case?

Jon Krohn: 00:16:46 I love it. What's a Pareto curve, Tyler?

Tyler Cox: 00:16:48 Pareto Curve is looking at what's the best thing to run if I care about a couple different KPIs, right? So what we typically will do for a Pareto curve is look at how fast is it versus how intelligent is it? And so if I need a model that has at least this intelligence, then I probably want to pick one that's there and fast. If I need something that's at least of particular speed, then I want to pick the smartest model that I can do. And what Ish was mentioning earlier on the artificial analysis site, they have a bunch of Pareto curves that are looking at other types of comparisons. It's cost per task for intelligence on different capabilities. It's a really interesting way to look at it. I think the closer that you can tune in your use case for it, that's a great way to pick when to move, when to upgrade.

Jon Krohn:	00:17:43	Fantastic. Really well defined there. All right, we should probably get back on track, although Ish, it did look like you might have just inhaled. Did you inhale? I
Ish Shah:	00:17:51	Did inhale. Is that President Obama who said that on the campaign trail?
Jon Krohn:	00:17:56	No. Oh, yes.
Ish Shah:	00:18:00	Yes.
Jon Krohn:	00:18:00	Yeah,
Ish Shah:	00:18:00	Right? You know, you know.
Jon Krohn:	00:18:03	And he was talking about breathing in to say something really important on a podcast.
Ish Shah:	00:18:06	Absolutely. That's what he was talking about. Pareto curves and fundamentally what is at the heart of this platform Tyler has built, and what is at the heart of every decision that every enterprise is having to make right now are trade-offs, cost versus speed, speed versus performance, performance versus cost. You can pick your X-axis and your Y-axis and plot to your heart's content and that curve that pops up, you're looking for the part where the curve starts to flatten out and it's probably your best bang for your buck, so to speak. I think that it's hard to understate the world of tokenomics and this economics of tokens and this economics of what model you pick, why you pick it, where it runs, what size is the model, how many NVIDIA GPUs are in your PC. All of these things are part of this decision matrix of how to get to the end state that your C-suite is demanding.
	00:19:06	And Shirish and I talked to you about this last time, Jon, the hardware is one dimension of that. Over the last 12 months, it's become clear the software is another dimension. Now you've got hardware, you've got software, you've got use case, you've got the literacy of the people

using the production stack. All of these things matter. So the reason I inhaled was to say trade-offs are super important and defining what trade-offs you are and are not willing to make as an organization, that's going to be the ballgame. So you got to stay on top of it.

Jon Krohn: 00:19:38 Love it. So many great pieces of information for us to work with practically already in this episode. The next one is the long promised tokens and token economics or tokenomics to make a port monteau. I think port monteau is the right word. Might have to look that up when you guys are speaking. Oh, I got some head nods. Great. So probably 95% of our listeners know what tokens are, but we can really quickly define that and then talk about how token consumption differs between the AI of 12 months ago or more that was this kind of generative or conversational only tokenomics relative to the tokenomics of today in 2026, which is OEGentech.

Ish Shah: 00:20:23 Yeah. Ty, I'll leave the what is a token to you and then I'll take the second half of the question. And

Jon Krohn: 00:20:30 Tyler, with you speaking about this, I understand that in your lab in Round Rock, Texas at Dell Technologies, you have a big screen showing token usage and somehow you're getting tons of free token usage. That seems to be something prominent on the screen.

Tyler Cox: 00:20:46 Yeah. So tokens are basically, it's the atomic unit of compute for a AI system. You can break up a paragraph into tokens, you can break up an image to tokens, so you can break up video to tokens, audio. All of it is just the mathematical representation of a particular chunk of information. And at every unit of compute, every cycle of a model, it's producing the next one. And so what we've got in the lab, we've got a bunch of the systems that we were talking about earlier, and we're going to talk a little bit more about what exactly those are that we're doing in the AI space with our Dell platforms. But we've got a

bunch of them hooked up with a variety of the latest and greatest models on there, and we're running workloads on. We're doing it for the lab infrastructure development. We're doing it for tools and analysis and reporting and visualization.

00:21:45 We're doing it for customer pilots of, "Hey, here's the use case that I need to size for you on our hardware." It's not a leaderboard, we're not token maxing here, but we're taking all that work and just visualizing what is the cost deflection from that. If I went and ran that on the equivalent cloud frontier model, what is the cost of that? We're doing hundreds of millions of tokens worth of volume a day inside the lab. When we say free, what we're really saying is none of that is incremental cost. We're not being billed for any of that. That is a system we purchased one time that we are operating with open software, open models for free and perpetuity. And right now my run rate, I'll tell you, is about $160,000 off of a lab of five or 10 people hitting this thing with just normal usage.

Ish Shah: 00:22:41 And what I'll add to what Tyler just said, heuristically, a token you can consider three-fourths of an English word. Take an average English word, consider three-fourths of it. And tokens include spaces and dashes and commas, and to Tyler's point, pictures can be converted into tokens, and that conversion is what allows you to have a conversation with ChatGPT. "Hello, ChatGPT, good morning, good morning-ish. Hey, I'm going to give you a picture. I need you to take a look at it. Here's the picture, hit send. "What's happening on the back end? That picture's getting tokenized. It has to be converted before a model, this black box engine thing that someone has made and trained and tied a bow on and handed to you can intake it, process it, figure out how it wants to respond to it, spit those tokens out on the backend. Now what's really important here is not only is it the atomic unit, like Tyler said, it's how you get billed for frontier

models, a million tokens of output, 50 US dollars per million.

00:23:46 Now to give you an example, over the weekend, I burned about two billion tokens working on a side project. And again, there was a brief moment a couple months ago, Jon, where the token maxing news cycle really picked up and it was like all these companies had these leaderboards and productivity equals token burden, right? It's an incredibly crude heuristic to use, but it's what we had, and to a certain extent, it's what we still have. So early in the adoption curve of a company in AI, how many tokens people are using is a good heuristic for our people using your AI tools at all, right? But the key here is, and this is kind of the drug you get hooked on, it's $50 per million tokens of output. Very, very smart listener base. I don't have to tell them what two and a half billion dollars would've cost me, right?

00:24:37 So it's important to understand that link because it's the gas, it's how it gets measured, and the gas is expensive. Who knew? So

Jon Krohn: 00:24:46 How do we. I guess we'll get to that later in the episode. It seems like we have a solution obviously involving hardware so that we can be churning through billions of tokens. There's a stat you said, Tyler, that I didn't quite understand if it was dollars or tokens you're talking about 160,000, what were the units? 160,000 what per day?

Tyler Cox: 00:25:08 Well, our run rate in the lab is about $160,000 per year of what we would spend that we are using the devices we got in the lab that would cost a heck of a lot less than that to - It's

Ish Shah: 00:25:24 The cost avoidance. Right. Yeah.

Tyler Cox: 00:25:25 I see.

Jon Krohn: 00:25:26 I see. And what is roughly back of the envelope, orders of magnitude, how much do you think the equipment costs, like 10 grand kind of thing

Tyler Cox: 00:25:34 To be

Jon Krohn: 00:25:35 Doing that 160?

Tyler Cox: 00:25:36 So the one that we're using the most right now, and we'll talk about the lineup here in a minute, but it's the Dell Pro Max with GB300, which is basically the biggest, baddest thing you can plug into a wall in an office space. It's got a Blackwell Ultra GPU from NVIDIA. It's got 1300 watts of GPU capacity, 252 gigabytes of HBM 3E. I can go through the spec sheet on it, but

Ish Shah: 00:26:03 You

Tyler Cox: 00:26:03 Can.

Ish Shah: 00:26:04 This

Tyler Cox: 00:26:04 Is the

Ish Shah: 00:26:04 Naturally aspirated V12 of AI computers is probably the best way to put it. Except

Tyler Cox: 00:26:11 It's liquid cool. But yeah, right now on dell.com, it's somewhere over $100,000. So it's a serious piece of iron.

Ish Shah: 00:26:23 He got me on the natural aspiration job.

Jon Krohn: 00:26:25 You're still saving. So it's roughly $100,000 piece of hardware, but your team is spending $160,000 per year and you can have that hardware for multiple years and it will still be current. So pretty obvious how that is major cost savings. Before we get into reeling off tons and tons of stats, which Tyler just did from memory, and I guess it's your job, but it still was impressive. Why does token consumption go up so much with agentic AI? Ish, how did

you burn through billions of tokens on the weekend on a side project? And can you tell us what it is?

Ish Shah: 00:27:02 I can. You may have to bleep out a word if I commit some sort of IP issue. Okay. So it's my one-year anniversary this Sunday, and as my wedding gift to my wife last year, what I did was I took. Everybody played Pokemon as a kid. Pokemon's making a comeback. It's cool again. Everything old is new again. I basically built a fan game in the art of Pokemon where the map is my area that we live in Atlanta, where my wife and I have met, where we got engaged, and I have these little pixel art maps, and I had her caricature done as pixel art, and I replaced the Pokemon with my dogs. That's the project. Every year, every major life event that we have, I build a chapter into the game, and that's my get out of jail free card on the present part of things.

00:27:59 And so what I've been working on is these models and their capabilities over the last couple months have shot through the roof. The artwork has gotten considerably better. The game mechanics and how much I need to supervise my little buddies as they go off and work. I can go have a cup of coffee and when I come back, the chapter is built. The reason the burn was so high is because what these agents are doing in order to achieve the task, just like humans, they're divvying up the work and they're spawning subagents. So now you've got an agent in charge of a bunch of other agents, and yes, the pie of work is finite. You have your finite pie of work, but because you've got all these subagents in action, are the subagents doing things to the Nth level of token efficiency that a single agent would've done or a single.

00:28:55 It's the same thing anthropologically as when you think about humans in a workplace, right? If one person says, "Everybody get out of my way, I'm going to own this task single-handedly. I'm going to do it as efficiently as possible, but I'm one person." This is like queuing theory.

How much throughput do you have? Multiple subagents means that you go faster, means the work gets divvied up, but the pie of work might get a little bit bigger because those subagents are at liberty to do certain things. The point of this is best probably articulated by something that has almost nothing to do with what we've talked about, although I'm sure it'll come up. It's this organization called METR, M-E-T-R, Model Evaluation for Research. Jon, you're nodding, so I'm not sure if they've been on the pod or.

Jon Krohn: 00:29:41 I talk about Meter probably more than any other single thing on the podcast, and then almost every talk that I've given for a year or two now, near the beginning, I show Meter charts. Ah,

Ish Shah: 00:29:54 So our presentation and your presentation are basically starting the same way. And then yours continue to be smart and mine kind of plateau. METER, Model Evaluation Threat Research, and SDS listeners are going to be familiar with this at this point, has a chart which when you land on their website, maybe we can put it in the show notes here, it shows on one dimension time, like 2021 until now. And then on the other dimension, it shows the ability of a model to operate unsupervised to achieve a certain goal at a certain fidelity of accuracy compared to a human given the same task. Now Meter, the reason they have this big scary name, which says threat research inside of it, their whole point was like, "Hey, at what point is AI going to cause harm to human beings? And we should probably be tracking that." And the heuristic they came up to track that with is this chart.

00:30:53 How much can it do by itself? And that chart is just like, not only is it up and to the right, it's gone vertical. And at a certain point they just kind of said, "I don't know, it just keeps going up."

Jon Krohn: 00:31:04 Since the release of Methos, they can't really track. It has gone off of the meter charts because in order to be able to benchmark the performance of models effectively on one of these charts, you have to have had humans doing these tasks and know how long it takes humans to do these tasks and supervised. And that was easy. 2021, when you're looking at GPT-3 level capability and the tasks are only seconds long or then minutes long with GPT-4 on average, it's very easy to come up with tasks that you can give humans to do. And it's not that expensive to pay them to do it and figure out how long it actually takes long hours to do it. But now that Methos is doing or Fable or Astra, GPT-6 from OpenAI, that class of models is now doing dozens of hours of work, work that would take a human dozens of hours.

00:31:57 It could take the AI model 30 minutes or whatever to do something that takes a human 16 hours or 24 hours or 36 hours. We don't know how long those tasks. We don't know how capable these models are because we don't have any human benchmarks. It's hard to even think of write a book chapter, write a book.

Ish Shah: 00:32:15 It is quite literally off the charts. Quite literally off the charts. And they accidentally invented a chart for one purpose is now the best visual we have for capabilities of models over time. But as these capabilities go up, it's Jevin's paradox here. Even if token costs get cheaper over time, the base is going to move on you because people are going to realize they could do things like. It took Nintendo how many years to develop a Pokemon game? They'd come out every two or three years when we were kids. Now it's like in a weekend someone can sit down and build a video game to the same level of fidelity. The token consumption is growing and it kind of doesn't matter how cheap you make the individual token if the order of magnitude of usage is just constantly chain reacting on itself to get bigger and bigger and bigger.

00:33:14 So that's kind of where we are.

Jon Krohn: 00:33:16 Yeah. And so that is why token use is exploding. Jovan's paradox, another thing I'll have in the show notes if people want to read more into that, but it's something we do also talk about on the show a fair bit. And it sounds like clearly instead of, like I do for the most part today, instead of buying tokens, instead of paying for tokens or running into token thresholds with one of the major cloud providers, we could be using our own hardware instead. That's the other big option. We're going to talk about specific examples, but just generally at a high level first, what is the big. If organizations move agentic workloads from paying per token to some cloud provider relative to doing it on their own desk side infrastructure, what are the potential kind of cost savings

Ish Shah: 00:34:10 There? So one thing I'll add before I answer the question, Jon, is that there's a middle that it's really important that we talk about the middle before we even get to desk side. That middle is on-prem compute. It's your own big computer as opposed to your own under desk computer. And obviously Dell, it's a very uniquely situated company because we do both. We have our infrastructure solutions and we have our client solutions. Dell is building the backbone for training and inference for massive companies all over the world, frontier labs and all, and they're doing so with that middle. So this decision of like, "Hey, I don't want to pay a cloud service for inference or I don't want to pay a cloud service for training or I don't want to pay a cloud service to host my deployed enterprise workload." You then sort of hit a fork in the decision chart, which is okay, how big are we talking?

00:35:06 How many people are we talking about? How much compute do you need? Oftentimes the answer's going to be an on-prem server, not an Underdesk GB 300, but the Underdesk GB300 is going to see that class of device and that class of ability closer and closer to the person. I

didn't need one of those devices a year ago. I mean, arguably I don't need one now, but I didn't need one of those a year ago. Now that my workflows are actively getting interrupted by token caps, I'm interested. So that middle is really, really important to acknowledge because. And Dell has papers on this that we've published which articulate literally the answer to your question, Jon, which is here's what it would cost in the cloud, here's what it would cost on-prem on your own server, and here's what it would cost on a T6 tower, which is one of our best AI devices, which you can cram full of Nvidia GPUs and you can host a little server.

00:36:05 You can put some models on this thing that have some serious capabilities. So the cost savings, I'm going to give you my recovering consultant answer on this, it depends. I know, I know, I know. My BCG bosses would be so proud of me. It depends and it highly is contingent on are you Tyler's team of five software engineers who are. They're driving the H2 Hummer. It's a gas guzzling pedal to the metal. How much code can I write? How many of these problems that I've been wrestling with for years can I try to solve quickly? They're going to experience the cost saving curve a lot faster than a more casual user or more casual work case or workload. This is why you're seeing enterprises adopt AI for software engineering faster than arguably any other function within the company.

Jon Krohn: 00:36:58 Nice. And after all that, which was very interesting, and thank you for the tour of the middle ground. I know that you have the it depends answer on the kind of cost savings thing, but I do also know that the research groups that you work with at Dell, like Signal 65 Solution Brief, I know that you do have some rough figures that you can give us.

Tyler Cox: 00:37:21 Yeah. So we worked with Signal 65 team across the Dell Technologies portfolio. We took devices like our Dell ProMax, the GV10, so compact form factor workstations.

Jon Krohn:	00:37:34	Tyler just held one up on the screen for those of you who aren't watching on YouTube. I
Tyler Cox:	00:37:38	Do.
Jon Krohn:	00:37:39	It looks like
Tyler Cox:	00:37:40	It's
Jon Krohn:	00:37:41	Kind of Kleenex box size. This is
Tyler Cox:	00:37:42	One of those where AI has gotten so powerful that you can unlock some really, really nice form factors. We also have scale up from there. We did a T2, we did the GB 300, we did GPU servers in here too. And so from that work, it depends answer, we've seen as high as up to 87% cost savings versus the equivalent workloads running in cloud. Those are all studies of am I modeling for knowledge workers? Am I modeling for sales workers? Am I knowledge for coding workers? How many agents am I running? How many times are they using it per day? What's the volume of work that they're doing there? And what that really translates to, and this is a really different way to think about PC buying, is that you're not looking at the CapEx, how much does the system cost, you're looking at how much does this system save me or make me?
	00:38:39	And so on some of these systems, for some of the workloads we've seen, they'll pay for themselves in two months versus running that same workload up in cloud. And over the lifetime of that system, you'll get over a million dollars worth of equivalent tokens spend.
Ish Shah:	00:38:57	I went and did the homework while Tyler was covering my rear, Jon.
Jon Krohn:	00:39:02	Which of your agents did the homework-ish?

Ish Shah: 00:39:04 I can't disclose that. I need to be on commission to disclose that. No. Okay, so we're talking a high complexity workload, AI agent, software assistant, software development assistant. A T2 workstation running an RTX Pro 6000 Blackwell achieved 93% cost savings for deployments supporting approximately 20 agents. So you're talking about orders of magnitude of potential savings, both in that middle layer, if you invest in an AI factory and you have these massive use cases, and for the under desk layer, which is if you buy T2 tower, which is much more accessible entry point into local AI and running your own inference than a T6 or a GP300, progressively those get more expensive as you move up the stack. But what used to prevent people from realizing that 93% cost savings, and this kind of moves into this concept we've been playing with around desk cytogenetic AI, it used to be really hard.

00:40:07 It used to be complicated and it used to be like Tyler and Ish over a weekend would spend hours getting set up on it and getting it all tuned and pecked out right so it would work, so I could use it from my phone, so I could do all kinds. Over the last 12 months, the strides in software, the strides in ease of setup, the strides. The ecosystem has come together to make it such that while it's not the same level of point and click as opening a website on your phone and just starting to talk to a model, the savings of up to 93% are certainly worth now the amount of effort it would take to set up a system this way. So you can hit it when you need it, so it can run the model you need.

Jon Krohn: 00:40:49 And I think some people might worry about not having the capabilities they need, but there's not that many use cases where you need a frontier fable or GPT-6 Astra capability, especially when there are open weight models, Kimi series, Quinn series that you could be using and getting so close to the frontier.

Ish Shah:	00:41:18	GLM is another one by ZAI. A couple of weeks ago, perhaps a month ago now, many, many, many organizations signed onto letters supporting open models. The Llama series of models from Meta back when all of this was getting started, it was the articulation of like, "Hey, we need open models because we need people to have choice and we need things that people can fine tune." The model layer of control in Meta's early opinion of all this was like, "We need people to have options and we're going to build the best option for people to use." Since then, many, many companies have entered the fray. Inkling is one of my favorite series of models right now, and it's actually Mira Murati who was at OpenAI and now I think it's -
Tyler Cox:	00:42:11	Thinking Machines. Thinking
Ish Shah:	00:42:12	Machines. Yeah, Thinking Machines is her company. They intentionally did not do a model to compete at the front. And you'll hear this term a lot, and I know Jon's heard this term and I know Tyler's heard this term, the jagged frontier of AI. It's not a clean frontier, it's a jagged frontier, which means that for different tasks, and this goes back to the trade-off discussion we were having, different models are going to be right or good enough. In order to do the artwork for my game, I have found that, yes, I need frontier model level image generation capability, otherwise it doesn't look how I want it. But the code underlying my game engine, I don't need Frontier for that. So I will have those agents running on a babier model with a lesser level of thinking and that'll save me some token burden. But right now I'm the human and I'm routing all that.
	00:43:10	Pretty soon, you're not going to have to do that either.
Jon Krohn:	00:43:12	Yeah. This Jagged Frontier thing is critical to mention because it also, even those crazy meter charts that we were talking about, that is specific to areas where we have a lot of training data. And the places that we have a

lot of training data are things like mathematics problems, computer science, machine learning, where we can simulate tons of data and know that it's accurate because the math works or the code runs or the machine learning model works. And so that is where the frontier is sharpest or furthest ahead. Whereas the jaggedness, if you try to. Good luck getting an AI model to clean bed sores off a hospital patient. Yeah. Right.

Ish Shah: 00:43:57 Although physical AI, man, I think 12 months from now, if you have a spec, Jon, we're going to be having a conversation that may not be that far away from that. So never say never.

Jon Krohn: 00:44:07 Yeah, I'd love to see it. But anyway, back to. I'm going to try to keep us on track a bit more so we get through everything we wanted to cover in this episode. It sounds like with these options of having our own hardware, whether it's desk side or on-prem, it sounds like we can get a return very quickly. It looks like I'm leading the witness here or actually I'm just going to. I know that you can get breakeven as quickly as three months after you buy that hardware. I don't know if you want to tell me more about that stat.

Ish Shah: 00:44:44 If you use Tyler's lab as the example, earlier on, we had fewer engineers running even more on it and then we were exploring concurrency. Our early math on the first GB 300 that we put in Tyler's lab is that it broke even in three months. We saw it. So that stat and that stat, they tie out, for me at least, because if you also know that you're not paying marginal token costs and empirically you're achieving the objectives that you sought out to achieve and you have evidence that like, "Hey, I'm not using the tip of the spear frontier model that costs $50 per million tokens of output. I'm using DeepSeq or I'm using I'm using Quen or I'm using GLM or I'm using one of these Nemotron or Poolside or Inkling, all of these folks

who make these models intended to run on smaller hardware than a full-blown data center.

00:45:43 If you do that math, you are very quickly going to come to very short breakeven periods, but you're inclined to use it more because it's empirically solving for your need. So you will realize very quickly, I don't need the bleeding edge to do this. I can do this this way. Therefore, your usage will go up so that breakeven time will get pulled in.

Jon Krohn: 00:46:03 Cool. Really cool. It is much faster than I would've anticipated. And so if you guys have. I think I'm going to start moving you to solutions, your specific solutions. I think we want to talk specifically in this episode about desk side agentic AI solutions from Dell to all the problems we've been talking about in this episode. But do you have one big takeaway from me on the tokenomics conversation that has enriched our conversation so far?

Tyler Cox: 00:46:32 So I think the really interesting thing that is definitely a challenge to how we think about IT is that we have this lived inherent assumption that the day that I put a device on the user's desk is its best day of life. There are lots of things that happen after that. You have policy updates, you have OS updates, you have the users doing crazy things on the systems. And what we've seen is with AI, and we've been talking about the trend of models getting better, that carries through for the platforms you're buying. The systems that we're talking about here, they can do way more things than they could a year ago. They'll be able to do way more in another year. So just as the frontier is increasing its capabilities on different models, even for the same size of hardware, because of what the industry's doing right now, you're going to be able to achieve harder and harder problems over time as well.

Jon Krohn: 00:47:34 Really cool. All right, let's move on to the solution part of the episode, which is Dell, desk side, Agentic AI. We

already talked a bit about on-prem and we already kind of got an introduction to these desk side solutions that y'all offer at Dell. So tell me what is included in one of these packages. I think it's more than just being a workstation, right?

Tyler Cox: 00:48:00 Yeah. So with our desk side agentic AI, what we're really saying is, "Hey, we have these set of platforms, this portion of our high-end AI portfolio that are agent ready. We know, we can tell you they run powerful enough models. They'll do them at scale, they'll do them efficiently. You buy one of these platforms, everything from the GB10 to the Dell Pro Precision nine series of scalable workstation towers, a T2, a T4, a T6, one to five GPUs, or the Dell Pro Max, the GB 300 we were talking about earlier, you buy one of those systems, you go put the NVIDIA agent toolkit, you put NVIDIA NemoClaw, go run OpenClaw or Hermes agent or your agent harness of choice on top of that. There's lots of very easy ways to get to value. And so with these solutions, we have a partner ecosystem in there.

00:48:52 We bring in security tooling for it, we bring in management tooling for it, which really take it from, I can do this thing as an exploration and to move it to, I can do this in my business. And that's where we've seen in a lot of customer conversations, that's where we're trying to help is how do I take this out of my lab and get it into my workflows broader than that? The other piece of the Deskside Agentik AI offering is we have this professional services team that will come in and help you get started. We have an adoption services to help you get started with local AI. Lots of people are using cloud and frontier models right now because it's easy. And what we're trying to do with Desktied Agentic AI is make it so that it's as easy to do the work with the value realization we've been talking around with the tokenomics piece.

Ish Shah: 00:49:49 Yeah. Tokenomics is this big theoretical thing where I can find the point on the curve that is best for me and my business, whether I'm a small business that has a couple of retail locations and I need inferencing happening at those locations, whatever the case might be, all the way up to I'm McDonald's and I have many retail locations and I need all of those things to work together. If a lot of what Tyler said out loud just now sounded complicated, that's what Deskside Agentik AI services from Dell and Nvidia, that's what we're trying to solve for. We are trying to make this as easy for people as what they're used to on the consumer software side, which is it just works. And we're solving that problem by forward deploying folks like Tyler intent to come help you out. And we're not just going to ship you this box and say, "Ta-da, here's this Dell PC.

00:50:44 Boy, do I have a solution for you?" The box comes with the Tyler, and that is, I think, a big part of the services offer, a big part of what turns this into, "We're going to sell you a PC. No, no, we're going to help you achieve a business outcome and we're going to leave you with the keys to a car that runs and it's not a DIY assemble it yourself unless you want that, in which case we're happy to provide that too.

Jon Krohn: 00:51:10 Nice. Yeah, so the offering here with, I guess this is the Dell AI factory with Nvidia, this kind of end-to-end offering of hardware, software, so things like the Nvidia Nemoclaw Stack that we' talking about. We'll get more into software again in a second and software options people have there. It includes security and it includes services like having a Tyler. Although I got to say, I don't know, does Tyler know that much? He hasn't impressed me that much in this case. I

Ish Shah: 00:51:33 Know. Underachiever.

Jon Krohn: 00:51:36 All right, so we're going to talk about hardware specifically and then software specifically after that. So what are the three specific different tiers of hardware that are available in this desk side Agentik solution?

Tyler Cox: 00:51:54 Yeah, so we have an exploration tier, right? It's where I may have a power user that I just want to get out of my inbox asking for more tokens that I want to go put a GB-10 or a T2 and say, "Hey, go nuts." Or I may have a team or a lab or a site where I just need dedicated intelligence at some small scale. We've got multi - GPU towers that you can go in there, deploy up to a 500 billion parameter model there and get to a better tier of intelligence. And then we have larger scale out solutions with the GB300 with the T6 and multiples where you're really looking at up to a trillion parameter models. You might be looking at hundreds of different agent instances. You might be doing things like the self-improving agents or self-optimizing problem sets that we've seen some buzz around the industry where there's hundreds or thousands or tens of thousands of agents working on one problem together and experimenting to try to find what the right answer is.

00:53:02 And so I think there's a lot of different problems that map well into those different pieces. And one of the reasons when we talk about where am I going for data center, where am I going for the edge, a lot of the customer use cases that are driving more towards these edge deployments with the desk side systems, it's because they want to bring intelligence into where the data is because it's IP, because it's sensitive data, because it's data that has legal agreements governing where exactly that can be moved around to, where it can be processed, what types of tools and systems. It's a really a wide variety of reasons of why you would use this, but it's a flexible operating model with a level of capability that we've never had before.

Jon Krohn: 00:53:48 Nice. And let's now move from hardware on to software. So we talked earlier about Nvidia's NemoClaw. What does that include?

Tyler Cox: 00:54:00 Yeah, so Nvidia NimoClaw has a couple major pieces. So Nvidia with a lot of the AI ecosystem software that they're promoting, they're doing some great open source contributions. For me, one of the key pieces of NemoClaw is OpenShell, which is a guardrail layer that wraps around your agent harness. You can put it into an OpenShell sandbox, you have fine grain permissions over the policies to really dictate what that agent can and can't do. You can read from these websites, from these data sets, you can use these tools, you can use these tools to access these sites. You can get input and post and patch or not for all these different pieces. And so with NemoClaw, they're really making it easy to deploy that consistently across different environments.

Ish Shah: 00:54:52 That manageability part, Jon, is super important. A lot of this stuff has been increasingly possible over time. We mentioned the difficulty part of it. Yes, there's a technical difficulty component to this problem. There's also a manageability and security component to this problem, manageability, security, and all of the observability that that entails, which for any of your listeners who are IT folks, this is going to be old hat to them. It's the classic problems of who's on my network, who's on my devices, what are they doing? How are they doing it? How do I track spend? How do I track if what they're claiming to be doing is actually what they're doing? All of that now are dimensions of a new order, like the AI question within the enterprise. So what NemoClaw allows an enterprise to do is bring some of those traditional IT guardrails into the AI context where you can have more control at a granular level over what is and is not happening within your IT environment.

Jon Krohn: 00:55:53 Love it. And so now we have a good understanding of what these Dell desk side agentic AI solutions have in terms of hardware, in terms of software. Let's, if you can, get into some specific use cases that illustrate for our listeners how that hardware, that software can work together to provide token efficient, highly accurate solutions.

Tyler Cox: 00:56:22 Yeah, so I think the number one use case that is common across most of our customer engagements is how do I get my developers to stop spending millions of dollars per month? How do I put caps on that? I like the productivity. I like what I'm getting out of it, but the line keeps going up. And so a lot of them will come in with, well, how do I get good enough models at a capitalized operating model so that I can move more of that volume down? But past that, there's a lot of really interesting places. We have healthcare researchers who are looking at how do I go off and scale out so that I can do these tests and tasks and look at different papers or pull in different research information because there's so much going on right now. How do I use an agent to give me extra hands?

00:57:17 We also have, with where we're at in the year and in the cycle, we have a lot of academic institutions coming in and saying, "Hey, how do I train my students, whether they're computer scientists or going into data science and data engineering or not, how do I train them to be able to take advantage of these new capabilities that are going out there?" And so we have a bunch of these lab deployments where it's, "Hey, I need 10 seats or 30 seats of these labs. How do I give them access to these great capabilities in a way that's friendly to my academic learning environment? Let them go play with different things, pull different tools, try different harnesses and models and all these capabilities." And so we've seen a lot of engagement around our desk side agentic AI systems for those types of use cases as well.

Ish Shah: 00:58:08 And it's not just big industrial use cases, Jon, right? These are of all sizes, shapes, scopes. Something that's really important to understand in the context of the question you asked earlier, what's with the token explosion? Why is this happening? Why is this happening the way it's happening? I'm going to steal a term a customer in London recently used with us, citizen development. It's such a nice way to say vibe code, but citizen development in the enterprise, right? So everyone, whether they know it or not, is a software engineer now. Everybody, whether they know it or not, you're writing code and code equals tokens, right? It's very simple A to B to C here. So as far as use cases go, even people who don't realize that what they're asking for is a desk side software assistant, that's what they're asking for because the types of things they're describing are achieved through code.

00:59:09 So it's a similar harness, it's a similar setup that we would come in and do for you, even if the use case you're describing, the qualitative words you're using to describe it might be something else.

Jon Krohn: 00:59:20 I love it. And this works, as you said, across different kinds of scales. Tyler was talking a few minutes ago about this explore tier where you've got Dell Pro Max with GP10 or Dell Pro Precision T2 tower. And so that handles up to eight agents running concurrently. You've got this orchestrate tier where you've got 40 agents, that's more like a T4 tower supporting 120 to half a trillion parameter models. And then obviously you've got the big heavy hitters, your GP300s, your Dell Pro Precision T6 tower where you're talking about trillion parameter models, 150 agents running. And it sounds, or from research that we did, Prior to recording this episode, Signal65, we've mentioned them already earlier, they're a third party that does research for Dell. They provided some pretty staggering stats where if you talk about common kind of workload types that you would be doing in this agentic

era that we now find ourselves in, someone who's like a knowledge worker doing low complexity tasks like email writing, text summarization, that kind of thing, you can see by having a desk side solution like Dell offers as opposed to paying per token with a proprietary API, you'd be seeing savings of like 28% to 71%, that kind of thing.

01:00:41 A sales agent, so we're talking increasing complexity here where you have a mix of that kind of email writing that the knowledge worker was doing, but also maybe research tasks happening, going out and finding out information about a prospective client or lead, gathering sales contacts, that kind of thing. So if you're talking tens of millions of tokens per agent per day, you could be in that kind of medium complexity scenario saving 76% to 91% versus cloud solutions. And then for a lot of our listeners doing software development, using agents to be increasingly looking more like that meter chart, using agents for handling tasks that take hours or days for a human to do, we're now talking about tens of millions of tokens per agent per day being consumed, many potentially dozens of agents, subagents running. And it's that as you move more, I mean, you're seeing savings even with a low complexity like knowledge work kind of agentic work.

01:01:43 But once we're getting to this high complexity code generation, code review, bug troubleshooting, testing, it starts to become a no-brainer where you're seeing 90% savings relative to using a cloud solution.

Ish Shah: 01:01:57 Yeah. And I would say the direction, and remember the same thing, we try to design workloads and benchmarks that are representative of what we think the complexity of a task looks like and Signal65 is independent in the way that they come up with things. So I think that the direction of the number is the most important thing and the magnitude of the number is the most important thing. Whether you're actually getting 25% savings, 75%

savings, 95, again, it depends. I think what's really important is that there is an opportunity here. It is a lot simpler than it used to be. And now with Dell desk cytogentic AI with Nvidia, it is push button to get to a point where you can use this stuff in a production environment. And I think that's really the big takeaway. And as much as I would love to think everybody knows what our desktop lineup acronyms are, I know that that's not the case, but don't let that drive you off.

01:02:53 There's a lot of people at Dell who can help you answer the question of, hey, what device needs to be part of this package? And we'll make sure you get the right one.

Jon Krohn: 01:03:01 All right. So all of that sounds great. Between all of us, we've given now a good run through of the tokenomics problem, the potential solution, the compelling solution in a lot of cases for having a desk side or on-prem solution to be saving money and getting the same kinds of results faster in a lot of cases. But if we have listeners out there wondering for them as individuals or particularly for organizations that they might work at, how hard is it to get up and running with the kinds of solutions that we've discussed today?

Ish Shah: 01:03:39 I would say it's not hard anymore. And I would say that that may not even have been true six months ago. So I never want to undermine that this is complicated and we're trying to reduce people's most difficult business problems into reusable kind of work blocks. But I would say that we are going to make this process as simple for you as possible. We're going to be partners in strategizing. We're going to be partners in procuring. We're going to be partners in deployment. We're going to be partners in sustain over the life cycle of that solution. So I would say that if this is something you are faced with, and this being exploding token bills within your enterprise, and if you do have an inkling of this being something that might

help you, you should give us a call because I think that that intuition is probably right.

Jon Krohn: 01:04:32 Love it. Great takeaway there from Ish. And we actually, while you were giving that response, we lost Tyler because he had a hard stop and I haven't been moderating this podcast recording session diligently enough. And so yeah, we ran out of time for him.

Ish Shah: 01:04:47 In truth, in truth, Tyler got tired of listening to me talk, so he left.

Jon Krohn: 01:04:52 So

Ish Shah: 01:04:52 Don't worry about that. That happens all the time, Jon.

Jon Krohn: 01:04:56 But so that means we're going to have to get a Tyler book recommendation in a future Tyler Cox appearance on the podcast. Ish, what do you have for us?

Ish Shah: 01:05:05 I think last time I cheated and I gave you two, and I think I'm going to do the same thing again this time.

Jon Krohn: 01:05:10 Oh man, you're going to run out someday at this rate.

Ish Shah: 01:05:12 I know, I know. There's just too much to consume right now. It's truly, truly difficult. And I can't keep up with the SDS podcast cadence at this point either. So books, podcasts, I got a lot of them. Two. So the first one is topical. It's Genius Makers, which I'm sure someone on the pod must have mentioned or you have read Jon at some point, which is kind of the origin story of the current lapse. Where were all these folks distributed across Silicon Valley? What companies incubated them? When did they choose to leave and start their own thing? How did they get folded back in sometimes? And it's got names of folks that at this point are canon for anyone who is interested in this industry. So Genius Makers was a great walk through the personalities, the humans who are responsible for the AI. And I think that that is going to

be so important over the next 12, 24, 36 months as we get into discussions about ethics and safety and things that we've never had to.

01:06:16 We always thought about them, but we never had to think about them. And now it's here. My fun book is I've continued to progress through the Jack Reacher series since the last time we talked and I'm almost done. I'm on book number 29 and I think book number 30 is coming out soon. In Too Deep is the last one I read and Exit Strategy is the next one. So Jack Reacher, nice fun airplane read, always a good time. Those are my two.

Jon Krohn: 01:06:41 Love it. Great recommendations, Ish. Thank you so much for those. And yeah, thanks to you and Wayward Tyler for taking so much time with us today, having such an information packed episode as we knew it would be. And in the end, I think we have circumvented all of the danger and we've survived to the end of the episode remarkably. For people who want to risk it, risk more danger in the future, how should they follow you, Tyler, Dell about what's going on? Yeah, right at the frontier.

Ish Shah: 01:07:16 So we've got all of our corporate social media handles and those are, if you punch in Dell, you punch up Nvidia, they'll come up. I'm on access@ishonshaw. It's the full name, not the ishiness, but maybe I'll get that handle one day. And we talk about these things and we talk about fun other stuff like Omarchi and fun other stuff that are frontier nerd kind of deals. So we welcome that engagement from everybody and looking forward to continuing the conversation and thank you Jon for having us back as always.

Jon Krohn: 01:07:46 Of course. I can't wait till the next time. It is always such a joy to have you on the show. Such invaluable minds to be able to get time with. I'm really lucky. I couldn't

Ish Shah: 01:07:57 Keep it. I'm sorry. No,

Jon Krohn:	01:07:59	It's true. I know. Both of you really -
Ish Shah:	01:08:04	Always a fun time.
Jon Krohn:	01:08:05	Such a joy to have on the air and I hope it won't be long till the next time.
Ish Shah:	01:08:09	Same here. Thanks Jon.
Jon Krohn:	01:08:12	What a fun episode. In it is Shah and Tyler Cox covered why token consumption is exploding. Agents spawn subagents that grow the total pie of work and thanks to Javon's paradox, cheaper tokens lead people to take on ever more ambitious projects. We talked about how Tyler's small app pushes hundreds of millions of tokens a day through local hardware, avoiding about $160,000 a year in cloud spend with their first GB300 breaking even in three months and how the jagged frontier means most tasks don't need a frontier model. So open weigh models like Quen, Kimmy, and GLM running locally are often good enough. 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 Ish and Tyler's social media profiles as well as my own at superdatascience.com/1031.
	01:09:03	Thanks of course to everyone on the Super Data Science podcast team, our podcast manager, Sonja Brajovic, media editor, Mario Pombo, our partnerships team Natalie Ziajski, our researcher, Serg Masís and our founder Kirill Eremenko.Thanks to all of them for producing another excellent episode for us today for enabling that super team to create this free podcast for you. We are deeply grateful to our sponsors. You, yes, you can support this show by heading to the show notes and clicking on sponsor's links, checking out what they're offering. And if you yourself are interested in sponsoring an episode, you can get the details on how at Joncrohn.com/podcast. Otherwise, please help us out by sharing this episode

with other folks who would love to learn about local AI hardware. Review this show on your favorite podcasting app or on YouTube. Subscribe, obviously, if you're not already a subscriber, but most importantly, I hope you'll just keep on tuning in.

01:09:59 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. Till 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.