# Run the Model Yourself: A Clinician's Guide to Local LLMs & Agents Transcript

(Transcripts are generated by AI and may contain errors)

## Podcast Introduction and Guest Setup

Good morning everyone and welcome again Good morning everyone and welcome again to Signals and Symptoms podcast. I'm to Signals and Symptoms podcast. I'm here with my two besties, Harvey Castro here with my two besties, Harvey Castro and Edward Marks, and Edward Marks, which are definitely not just my which are definitely not just my besties, but the co-host.

besties, but the co-host. Today we're going to feature our own Today we're going to feature our own Harvey Castro again because clearly Harvey Castro again because clearly future of health care is from Earth to future of health care is from Earth to orbit. But he just actually is working orbit. But he just actually is working on an amazing book. Uh on an amazing book. Uh is AI in emergency medicine. And then is AI in emergency medicine. And then we're going to go ahead and talk about we're going to go ahead and talk about all of this. He's just This is just a all of this. He's just This is just a summary of 11 chapters. But let let's summary of 11 chapters. But let let's talk about this book and then go from talk about this book and then go from there.

there. there.

## Harvey's AI in Emergency Medicine Book

One of the good readings and again, I One of the good readings and again, I don't know if you guys have subscribed don't know if you guys have subscribed to Medium, but the dark side of uh to Medium, but the dark side of uh uh AI of data is actually very uh AI of data is actually very important. And as we go along into this important. And as we go along into this process of flywheel, process of flywheel, data is going to become a significant data is going to become a significant problem. This was very good. Open problem. This was very good. Open evidence goes Hippocratic AI. Again, evidence goes Hippocratic AI. Again, they said that, you know, why use open they said that, you know, why use open evidence when you can use chat GPT.

evidence when you can use chat GPT. And there's a whole research paper that And there's a whole research paper that came out. All of this. Interesting read.

came out. All of this. Interesting read. Then of course Then of course uh I'm sorry. I didn't pull up the open uh I'm sorry. I didn't pull up the open evidence reply to it, but it's a very evidence reply to it, but it's a very good read of how things are progressing.

good read of how things are progressing. And then that's another because what is And then that's another because what is it? Is it the large language models it? Is it the large language models outperforms specialized clinical AI outperforms specialized clinical AI tools?

tools? tools? And then are the benchmarks okay? So And then are the benchmarks okay? So those are the kind of things that I've those are the kind of things that I've been thinking about and then going from been thinking about and then going from there. I'm going to share some of my there. I'm going to share some of my findings directly, but an another uh findings directly, but an another uh thing that we're trying to do over here thing that we're trying to do over here is to ensure this. Before I move into is to ensure this. Before I move into something else, I just want to make sure something else, I just want to make sure that you understand a massive big update that you understand a massive big update from SaveLife AI. We just got FDA from SaveLife AI. We just got FDA approval for our third one, which is approval for our third one, which is chest x-ray. So this is officially now chest x-ray. So this is officially now FDA approved. We can do actually chest FDA approved. We can do actually chest x-ray analysis through AI. So we now x-ray analysis through AI. So we now officially three FDAs. The fourth one is officially three FDAs. The fourth one is coming coming coming I think in a day or two. And then we I think in a day or two. And then we have five planned have five planned for this month.

for this month. Interestingly, the things that we want Interestingly, the things that we want to build is not really have an FDA to build is not really have an FDA pathway. We have talked about regulation pathway. We have talked about regulation and everything. So, we are now working and everything. So, we are now working as SaveLife AI as our first FDA as SaveLife AI as our first FDA breakthrough device.

breakthrough device. The application is going to be filed by The application is going to be filed by the end of August. I don't know what's the end of August. I don't know what's going to happen because it is how it is.

going to happen because it is how it is. Unfortunately, we need really need to Unfortunately, we need really need to move the needle. And I cannot just move the needle. And I cannot just confine myself with the 510k pathway confine myself with the 510k pathway because it's just too limiting. I mean, because it's just too limiting. I mean, unfortunately. And we're going to talk unfortunately.

## SaveLife AI FDA and RCM Expansion

And we're going to talk up to Harvey about this. But here's the up to Harvey about this. But here's the biggest update and Harvey, as an biggest update and Harvey, as an investor, you need to know.

investor, you need to know. We have, of course, the clinical We have, of course, the clinical platform, the radiological platform, the platform, the radiological platform, the data integration platform with EHRs and data integration platform with EHRs and everything. And then our AI suites.

everything. And then our AI suites. And then, as you know, we believe in And then, as you know, we believe in what we call clinico-radiological what we call clinico-radiological intelligence. But while I'm talking to intelligence. But while I'm talking to the clients, interestingly, they're very the clients, interestingly, they're very interested in our solutions. They're interested in our solutions. They're buying like crazy. But one of the things buying like crazy. But one of the things that they everyone is talking about is that they everyone is talking about is essentially the complete end-to-end essentially the complete end-to-end workflow, which we can do at Gentec.

workflow, which we can do at Gentec. And I'm going to That's why I'm going to And I'm going to That's why I'm going to show you some of my updates at the end show you some of my updates at the end after we talk about Harvey's book. So, after we talk about Harvey's book. So, we basically moved into what we call we basically moved into what we call AIRCM.

AIRCM. AIRCM. Basically, documentation to dollars Basically, documentation to dollars since we are already since we are already doing documentation implementation and doing documentation implementation and everything. So, this is going to be a everything. So, this is going to be a complete, you know, end-to-end AIRCM complete, you know, end-to-end AIRCM solution where you can have denial solution where you can have denial management, you can have management, you can have charge management, claims management, charge management, claims management, also as a voice agent, also as a browser also as a voice agent, also as a browser agent, everything end-to-end. Prior agent, everything end-to-end. Prior authorization, front desk, coding engine authorization, front desk, coding engine that is ICD-10 coding auto coding coding that is ICD-10 coding auto coding coding we can do. How to create super bill we can do. How to create super bill directly from ISMND. All of this. This directly from ISMND. All of this. This is going to be the last leg that is is going to be the last leg that is going to just come the not just the going to just come the not just the clinical radiological and I think I've I clinical radiological and I think I've I need your help with the actually helping need your help with the actually helping Oh, am I not sharing this? Oh, I'm Oh, am I not sharing this? Oh, I'm sorry.

sorry. sorry. Oh, so I'm sorry. The AI RCM and this is Oh, so I'm sorry. The AI RCM and this is actually going to go ahead and actually going to go ahead and completely so just to remind you again completely so just to remind you again AI suite now has uh you know, officially AI suite now has uh you know, officially chest x-ray fully FDA cleared. That's chest x-ray fully FDA cleared. That's third FDA, fourth is coming, etc. And third FDA, fourth is coming, etc. And then if you go into AI RCM solutions, we then if you go into AI RCM solutions, we are basically building complete are basically building complete end-to-end workflow from prior end-to-end workflow from prior authorization, front desk, front desk authorization, front desk, front desk voice, voice, voice, point of service. This is

something that point of service. This is something that we are this is the last leg before I we are this is the last leg before I release it. Coding agents, coding space, release it. Coding agents, coding space, super bill, charge audit, everything.

super bill, charge audit, everything. Claims engine, patient billing and Claims engine, patient billing and and then of course it connects directly and then of course it connects directly to our operational intelligence to our operational intelligence completely fed again from a real completely fed again from a real freaking company that really uses the freaking company that really uses the AI. You know what? I have like freaking AI. You know what? I have like freaking seen three demos so that I want to buy a seen three demos so that I want to buy a a software like a software like you know, for my own company from like you know, for my own company from like quality management systems and all of quality management systems and all of that. They say AI. I mean, this is like that. They say AI. I mean, this is like but they have no AI.

but they have no AI. >> [laughter] >> [laughter] >> [laughter] >> So, this is coming from a company that >> So, this is coming from a company that is FDA approved. They come over once a is FDA approved. They come over once a year. Then I bend over, take off my year. Then I bend over, take off my shorts, and they do a colonoscopy on me.

shorts, and they do a colonoscopy on me. So, just to remind you that So, just to remind you that >> [laughter] >> [laughter] >> [laughter] >> that this is it. So, we have complete >> that this is it.

## Why Books Still Matter

So, we have complete This is the last leg. We of course This is the last leg. We of course initially had clinical and radiological initially had clinical and radiological intelligence. And I don't know, I need a intelligence. And I don't know, I need a better line from you Harvey because you better line from you Harvey because you are amazing at coming up with this that are amazing at coming up with this that how do I put clinical radiological how do I put clinical radiological intelligence intelligence intelligence rather and then add RCM solutions as rather and then add RCM solutions as well into it. So, some well into it. So, some major massive updates from major massive updates from my side as far as that is concerned.

my side as far as that is concerned. Now, Now, Now, Harvey, going back to you, Harvey, going back to you, uh uh uh why do you keep writing books, man?

why do you keep writing books, man? Nobody reads them. I'm just messing with Nobody reads them. I'm just messing with you. So, what what you. So, what what >> [laughter] >> [laughter] >> [laughter] >> So, that's why we do partner with >> So, that's why we do partner with >> Ed and I don't make good writers. No, >> Ed and I don't make good writers. No, you're right. On the joke side side, you're right. On the joke side side, yeah, but you are right, actually. Less yeah, but you are right, actually. Less and less people are reading books, per and less people are reading books, per se, as opposed to 5, 10, 20 years ago.

se, as opposed to 5, 10, 20 years ago. We're accumulating knowledge a lot We're accumulating knowledge a lot quicker. Uh with that said, quicker. Uh with that said, there are still some people, including there are still some people, including myself, myself, myself, >> I do. And I'm actually learning a lot >> I do. And I'm actually learning a lot from you, so jokes aside, the one of the from you, so jokes aside, the one of the things that I wanted to say is so, things that I wanted to say is so, what's the difference between a book, a what's the difference between a book, a website, or a wiki article? Is that that website, or a wiki article? Is that that a book is an actual absolute snapshot.

a book is an actual absolute snapshot. And that snapshot in time is also very And that snapshot in time is also very important to encapsulate. Because he's important to encapsulate. Because he's going to update this book in a year, and going to update this book in a year, and then 2 years from now, etc. And that then 2 years from now, etc. And that that snapshot actually preserves the that snapshot actually preserves the historical evidence. And I'm going to be historical evidence. And I'm going to be very excited to see, at some point in very excited to see, at some point in time, that how fast he has to update the time, that how fast he has to update the book with book with with this AI now. So, it it it it is with this AI now. So, it it it it is this snapshot of intelligence, this snapshot of intelligence, comprehensive, combined, comprehensive, combined, is extremely important as we move is extremely

important as we move forward, not today, but also in future forward, not today, but also in future as well. Sorry. All right, go ahead.

as well. Sorry. All right, go ahead. >> Well, no, you're right. To that point, >> Well, no, you're right. To that point, you know, I feel like I can move a lot you know, I feel like I can move a lot quicker than the book publisher. So, I quicker than the book publisher.

## Patient-Focused AI in Emergency Care

So, I started this year a series called Pivot started this year a series called Pivot to Parish, and on my website, I have an to Parish, and on my website, I have an AI agent that takes that book and AI agent that takes that book and updates it every month. So, you buy the updates it every month. So, you buy the book, and then after that, you go on the book, and then after that, you go on the website, and it just updates you every website, and it just updates you every single month.

single month. So, it's a way of keeping up So, it's a way of keeping up the latest data, the latest information, the latest data, the latest information, without having to go out and buy a book, without having to go out and buy a book, right by the wait for the second right by the wait for the second edition. Eventually, I see the book edition. Eventually, I see the book publishers doing the same.

publishers doing the same. Mhm. Mhm. Mhm. >> I agree with you. So, tell me about this >> I agree with you. So, tell me about this why, as you said. I'm going to play why, as you said. I'm going to play Harvey now.

Harvey now. >> [laughter] >> [laughter] >> [laughter] >> Harvey, like always, the main question >> Harvey, like always, the main question is why. So, why write the book in the is why. So, why write the book in the first place, and then what is this book first place, and then what is this book about?

about? about? >> You know, it was a blessing. When I >> You know, it was a blessing. When I wrote ChatGPT in healthcare in the year wrote ChatGPT in healthcare in the year 2022, book publishers started reaching 2022, book publishers started reaching out for the second edition. And during out for the second edition. And during that period or during throughout the that period or during throughout the years, Wiley reached out. I was really years, Wiley reached out. I was really honored when they said, "Hey, Harvey, we honored when they said, "Hey, Harvey, we want to write a book in emergency want to write a book in emergency medicine AI. We want your personal medicine AI. We want your personal stories, how you went through it, and stories, how you went through it, and how you see it today." And I'm really how you see it today." And I'm really excited. I I've excited. I I've You know how busy I've been. I I feel You know how busy I've been. I I feel bad cuz I keep delaying, delaying, bad cuz I keep delaying, delaying, delaying, and then they're like, "Okay, delaying, and then they're like, "Okay, Harvey, the book's going to be out in Harvey, the book's going to be out in October." And then October." And then we've been proofreading it. So, I gave we've been proofreading it. So, I gave them the final proofread actually today.

them the final proofread actually today. So, hopefully here they said that it'll So, hopefully here they said that it'll be on the market. And so, why I did it, be on the market. And so, why I did it, two things. It's human plus AI, is what two things. It's human plus AI, is what I always say. Human, 20 years of I always say. Human, 20 years of emergency medicine experience, AI in emergency medicine experience, AI in that I've been leveraging, using, seeing that I've been leveraging, using, seeing where the future heads. And then I where the future heads. And then I divided the book in that format.

divided the book in that format. Too much to cover in a one sitting, so Too much to cover in a one sitting, so why not put it in a book? So, someone why not put it in a book? So, someone that can go through A to Z and and that can go through A to Z and and understand how it could be applied. The understand how it could be applied. The principles are general, so they apply principles are general, so they apply emergency medicine. Obviously, they emergency medicine. Obviously, they apply other parts of medicine. So, I'm apply other parts of medicine. So, I'm >> [snorts] >> [snorts] >> Love it. I'm looking forward to getting >> Love it. I'm looking forward to getting it and then reading it. Is it out it and then reading it. Is it out already on the Amazon and everything or already on the Amazon and everything or not yet?

not yet? >> It's all on pre-orders. It's on Amazon, >> It's all on pre-orders. It's on Amazon, and I didn't realize it's at Target, and I didn't realize it's at Target, too. So, too. So, so they're all having pre-sales right so they're all having pre-sales right now. So, I was kind of stoked about it.

now. So, I was kind of stoked about it. >> we'll >> we'll >> we'll We'll put the links here. I can you find We'll put the links here. I can you find it and put it in the comments.

it and put it in the comments. The pre-order links, and we'd love to do The pre-order links, and we'd love to do so. So, now we talked about the why, so. So, now we talked about the why, which is clearly there is a need and which is clearly there is a need and then of course we're looking at AI in then of course we're looking at AI in this. Actually, before I go, Ed, do you this. Actually, before I go, Ed, do you have your first question for Harvey have your first question for Harvey about the book?

about the book? >> Well, I was going to say we we should do >> Well, I was going to say we we should do a book signing like we did at my house a book signing like we did at my house when we did the Mayo Clinic book.

when we did the Mayo Clinic book. And we should bring Harvey's books and And we should bring Harvey's books and my books together and we'll do something my books together and we'll do something fun in the Dallas-Fort Worth region and fun in the Dallas-Fort Worth region and maybe we'll take it on the road, also.

maybe we'll take it on the road, also. >> I like that. >> I like that. >> of my >> of my thought. thought. thought. And you know, the thing I'll say about And you know, the thing I'll say about books, I I agree, you know, I'm I'm books, I I agree, you know, I'm I'm pretty new school, but the only place pretty new school, but the only place that I think I'm still old school is that I think I'm still old school is books. And and I'll tell you a couple of books. And and I'll tell you a couple of reasons why. There when you hand someone reasons why. There when you hand someone a book, you can't hand someone a a book, you can't hand someone a website. You can say Harvey's the website. You can say Harvey's the greatest dude in the world, he's got the greatest dude in the world, he's got the greatest content, you can't hand them greatest content, you can't hand them anything about Harvey except point it to anything about Harvey except point it to the website. And the likelihood, just the website. And the likelihood, just like when you prescribe someone a test, like when you prescribe someone a test, only 30% of tests actually get only 30% of tests actually get fulfilled. I forget what it is with fulfilled. I forget what it is with prescriptions, it's probably 50%. When prescriptions, it's probably 50%. When you hand someone a book, it's fulfilled.

you hand someone a book, it's fulfilled. Now, will will they read it? But they Now, will will they read it? But they have it, right? So, I think that's have it, right? So, I think that's that's one thing. And another thing is that's one thing. And another thing is if you're on the plane and yeah, there's if you're on the plane and yeah, there's Wi-Fi which only works half the time if Wi-Fi which only works half the time if you fly American Air like I do.

you fly American Air like I do. And you can get to the website or And you can get to the website or whatever or read the latest blog, but whatever or read the latest blog, but having that book, you don't have to having that book, you don't have to worry about stowing it away when you worry about stowing it away when you take off or whatever. You know what I'm take off or whatever. You know what I'm saying? There there's still something saying? There there's still something magical about a book and then you know, magical about a book and then you know, and then you can point someone to it.

and then you can point someone to it. And then here's my final thing on books. And then here's my final thing on books. I've thought about this a lot. Is I've thought about this a lot. Is it's like albums, right? I have a kid in it's like albums, right? I have a kid in the recording industry. Why do they put the recording industry. Why do they put out an album? They still do. They out an album? They still do. They they're still music out there and then they're still music out there and then and every what? Every year depending how and every what? Every year depending how active the artist is, a year and a half active the artist is, a year and a half they come out with something new. They they come out with something new. They come out with new music and they put out come out with new music and they put out a new album, um, a new album, um, a record. So, I I'm still old school a record. So, I I'm still old school that way. I still think there's that way. I still think there's something about it. And and the final something about it. And and the final thing like I'm about to give a thing like I'm about to give a talk and it's around one of the books talk

and it's around one of the books and you give people a copy of the book.

and you give people a copy of the book. It's it's a it's a keepsake. They'll It's it's a it's a keepsake. They'll remember, "Oh, I heard Harvey speak and remember, "Oh, I heard Harvey speak and I got his book and he signed it. It I got his book and he signed it. It means something." And you can't do that means something." And you can't do that with a website. You could put your And with a website. You could put your And we all do it, right? Every speaker does.

we all do it, right? Every speaker does. You put your little URL up there to your You put your little URL up there to your website or to your LinkedIn profile. How website or to your LinkedIn profile. How many people are actually going to hit many people are actually going to hit that link or take that picture and that link or take that picture and actually go and do it. But you give them actually go and do it. But you give them a book, it's a it's different. There's a book, it's a it's different. There's an there's a connection when you do an there's a connection when you do that. It's a it's a gift, if you will.

that. It's a it's a gift, if you will. So, anyways, rock on, Harvey. Keep So, anyways, rock on, Harvey. Keep writing those books. I'm with you. writing those books. I'm with you.

>> Love it. >> Love it. >> 100%. So, >> 100%. So, So, yeah, it's AI in in emergency So, yeah, it's AI in in emergency medicine and you you already mentioned medicine and you you already mentioned that you you don't have to be emergency that you you don't have to be emergency doc to get benefit. But who else who doc to get benefit. But who else who else is the else is the target like for the reader? Like target like for the reader? Like obviously, obviously, obviously, everyone in emergency medicine, urgent everyone in emergency medicine, urgent care, those sorts of things would be care, those sorts of things would be like, "Hey, you should get this book." like, "Hey, you should get this book." But but who else is it written for?

But but who else is it written for? >> You know, honestly, the main goal voice >> You know, honestly, the main goal voice is the patients. is the patients.

You know very well, patients are going You know very well, patients are going to the emergency room. A good percentage to the emergency room. A good percentage of them their primary care is the of them their primary care is the emergency room.

emergency room. You know, so that's who I wrote the book You know, so that's who I wrote the book for. It's really for patients. And so, for. It's really for patients. And so, being me being a doctor and getting a being me being a doctor and getting a lot of administrators and physicians lot of administrators and physicians calling me, I said, "You know what? Let calling me, I said, "You know what? Let me add some physician stuff to it." But me add some physician stuff to it." But the main audience is patients and it's the main audience is patients and it's written for patients.

written for patients. >> Oh. That's that's interesting. So, That's that's interesting. So, >> Yeah, tell us more then. Okay, so I'm >> Yeah, tell us more then. Okay, so I'm I'm I'm I'm I'm average Joe Jane citizen here in I'm average Joe Jane citizen here in Portland, Oregon and um Portland, Oregon and um and so I would I would get this book cuz and so I would I would get this book cuz it would help me it would help me navigate emergency room help me in my navigate emergency room help me in my own health care what exactly what?

own health care what exactly what? >> Yeah, so I broke it down into the just >> Yeah, so I broke it down into the just assuming you and I and all of us are assuming you and I and all of us are different level but assume someone goes different level but assume someone goes into health care has no idea about AI.

into health care has no idea about AI. So I go through the foundation like hey, So I go through the foundation like hey, what's going on with your doctor? Hey, what's going on with your doctor? Hey, they're burning out. You know what? We they're burning out. You know what? We have some clinical AI. Hey, what's the have some clinical AI. Hey, what's the future look like? Hey, we have wellness, future look like? Hey, we have wellness, we have wearables, we're using AI to we have wearables, we're using AI to predict the predictive predict the predictive

proactive care. And then obviously proactive care. And then obviously when you go into the doctor like who's when you go into the doctor like who's doing all this? Is it the hospital? Is doing all this? Is it the hospital? Is it the doctor? What's the rate limiting it the doctor? What's the rate limiting step? What is the leadership? So I kind step? What is the leadership? So I kind of go into that why you know what things of go into that why you know what things can be released and why but then I go can be released and why but then I go into the critical part the guardrails.

into the critical part the guardrails. Everybody's like why should I do this? Everybody's like why should I do this? What kind of what governance is going on What kind of what governance is going on and is there a consent and there and is there a consent and there algorithms for that? So then I go into algorithms for that? So then I go into that. So it's broken up into those units that. So it's broken up into those units so that a person can go through and be so that a person can go through and be like okay, I really understand this like okay, I really understand this health care space now.

health care space now. with AI. with AI. >> So cool. I mean I love that that you're >> So cool. I mean I love that that you're actually truly thinking of So first of actually truly thinking of So first of all comments again.

all comments again. I love that these two guys because I'm I love that these two guys because I'm becoming an AI becoming an AI idiot because the time like I have like idiot because the time like I have like 16 AI employees working and then you 16 AI employees working and then you know it's just getting I need to redo my know it's just getting I need to redo my human stuff.

human stuff. And I love it that these two bring it And I love it that these two bring it back to the patient and the human again back to the patient and the human again revising it that tactile feeling the the revising it that tactile feeling the the book feeling is still one of those many book feeling is still one of those many pleasures of life that I still sometimes pleasures of life that I still sometimes pick a book. And I never got you know, pick a book. And I never got you know, an e-ink reader or iPad reading. It's an e-ink reader or iPad reading. It's just you know, that tactile feeling to just you know, that tactile feeling to just curl up in a you know, a blanket just curl up in a you know, a blanket and then just read. Especially how and then just read. Especially how Harvey writes, by the way, he's a great Harvey writes, by the way, he's a great story story story teller. If you look at his previous teller. If you look at his previous book. It's not technical read. I don't book. It's not technical read. I don't know, is this a technical read or is it know, is this a technical read or is it a is it a a is it a >> simple >> simple >> simple >> Perfect. So, it's simple. Especially >> Perfect. So, it's simple. Especially when it's patient based. I would love when it's patient based. I would love to, you know, just sit in a sitting and to, you know, just sit in a sitting and then just read it. There's something to, then just read it. There's something to, you know, move the page that brings you you know, move the page that brings you you know, that dopamine hit too, by the you know, that dopamine hit too, by the way, at least me personally. So, thank way, at least me personally. So, thank you, Albert, to realizing that again, you, Albert, to realizing that again, the humanity of things in the age of AI.

the humanity of things in the age of AI. And of course, writing it for the And of course, writing it for the patients. Now, Harvey, have you included patients. Now, Harvey, have you included any patient stories and you want to any patient stories and you want to share some of them on this that how AI share some of them on this that how AI will go through that and then extend it will go through that and then extend it one more to phantom space as well.

one more to phantom space as well. >> Yeah, I'll keep it generic cuz I'm not >> Yeah, I'll keep it generic cuz I'm not sure what I can share or not. I just got sure what I can share or not. I just got the the the copyright portion of it. But the the the copyright portion of it. But this skinny the the big picture I always this skinny the the big picture I always talk about and and I share similar talk about and and I share similar stories is that whole proactive versus stories is that whole proactive versus reactive approach.

reactive approach. At the end of the day, guys, does At the end of the day, guys, does everyone out there go to their doctor? everyone out there go to their doctor?

## AI Infrastructure, Models, and Edge Computing

And just going by statistics, women are And just going by statistics, women are more likely to go to their doctor. Us

more likely to go to their doctor. Us men tend not to go to our doctor as men tend not to go to our doctor as often. Even me as a doctor, I have a often. Even me as a doctor, I have a doctor's appointment next month and I doctor's appointment next month and I don't know when it is. I just know when don't know when it is. I just know when it is sometime. But my point is this.

it is sometime. But my point is this. Why are we waiting to go to the doctor? Why are we waiting to go to the doctor? Why not leverage AI? And I go into the Why not leverage AI? And I go into the stories, why not leverage the aura, the stories, why not leverage the aura, the wellness, the information so that we can wellness, the information so that we can gather that information. And I'm really gather that information. And I'm really excited that I'm seeing Perplexity, excited that I'm seeing Perplexity, OpenAI, all of them create a platform to OpenAI, all of them create a platform to connect your personal healthcare data.

connect your personal healthcare data. Regardless of, you know, how you feel Regardless of, you know, how you feel about it, think of the core of that.

about it, think of the core of that. You're analyzing your data with AI and You're analyzing your data with AI and saying, "Hey, here's some of the things.

saying, "Hey, here's some of the things. Let me explain your your blood work. Let Let me explain your your blood work. Let me explain some of the things the me explain some of the things the doctors are going to say." Why does that doctors are going to say." Why does that matter? Twofold. One, because you can matter? Twofold. One, because you can take charge of your health. Two, we take charge of your health. Two, we don't know what we don't know, but if don't know what we don't know, but if the AI is introducing to us certain the AI is introducing to us certain information from our labs, and then when information from our labs, and then when we go to the doctor, we'll be like, we go to the doctor, we'll be like, "Hey, "Hey, "Hey, what is this A1C hemoglobin A1C that the what is this A1C hemoglobin A1C that the AI says is this? Explain it to me, or AI says is this? Explain it to me, or can we talk about it?" And it takes it can we talk about it?" And it takes it to that next level. I want everybody out to that next level. I want everybody out there to realize in the United States, there to realize in the United States, you're lucky to get 13 minutes is the you're lucky to get 13 minutes is the national average with your doctor. 13 national average with your doctor. 13 minutes of medical expertise. We need minutes of medical expertise. We need more, and I'm glad to see these tools more, and I'm glad to see these tools >> So cool. So cool.

>> So cool. So cool. Add your thoughts on that, because Add your thoughts on that, because again, as a physician, I'm going to come again, as a physician, I'm going to come back as a physician mind headset, but as back as a physician mind headset, but as as as a as a general consumer of health as as a as a general consumer of health care, and not only general consumer of care, and not only general consumer of health care, as a board member of health care, as a board member of Hospital Supplements, what do you think Hospital Supplements, what do you think >> Well, I I have a huge proponent of it.

>> Well, I I have a huge proponent of it. You know, I I believe in You know, I I believe in arming patients with as much knowledge arming patients with as much knowledge as we can, and to Harvey's point, you as we can, and to Harvey's point, you know, know, know, I've visits are down 16% was what I saw I've visits are down 16% was what I saw from Gallup and West recently.

from Gallup and West recently. Office visits are pretty Yeah, they're Office visits are pretty Yeah, they're they're down 16%, and and a lot of they're down 16%, and and a lot of people are going to AI, and maybe people are going to AI, and maybe they're not doing it in a very wise way, they're not doing it in a very wise way, right? Could just be willy-nilly, and a right? Could just be willy-nilly, and a book like this I think is going to help book like this I think is going to help give them a little bit more structure give them a little bit more structure and knowledge, like how to how to and knowledge, like how to how to navigate safely, and and make sure that navigate safely, and and make sure that they don't, you know, they don't, you know, go go wrong direction. So, I think it's go go wrong direction. So, I think it's a I think it's a great a I think it's a great a great thing. I can't wait to read it.

a great thing. I can't wait to read it. I'm I'm surprised, Junaid, we haven't I'm I'm surprised, Junaid, we haven't gotten like pre-read copies or anything.

gotten like pre-read copies or anything. >> We Actually, I'm not so sure. >> We Actually, I'm not so sure. I do I don't know. Did we do this for I do I don't know. Did we do this for Harvey?

Harvey? Harvey? Did I interview you? No. Did I interview you? No. >> No, I I should have I >> No, I I should have I >> He's just freaking too busy. You know >> He's just freaking too busy. You know what?

what? what? You know, we we have to like penalize You know, we we have to like penalize him. So, here's the here's your penalty, him. So, here's the here's your penalty, Harvey.

Harvey. Harvey. You're taking us both on a steak dinner You're taking us both on a steak dinner tonight. Well, actually not tonight. tonight. Well, actually not tonight.

Harvey, and where are you by the way? I Harvey, and where are you by the way? I see in the background you're walking see in the background you're walking around in a in a in a around in a in a in a Where Where are you?

Where Where are you? >> Yeah, it's Portland. Portland. >> Yeah, it's Portland. Portland. >> Ah, Portland. >> Ah, Portland. >> Oh, it's early. It's really really Yeah, >> Oh, it's early. It's really really Yeah, I had to run first at 5:00 so I could I had to run first at 5:00 so I could start with you all at 6:00. So, I'm just start with you all at 6:00. So, I'm just doing my cool down with you all.

doing my cool down with you all. So, pardon my informality, but I had to So, pardon my informality, but I had to >> So, Harvey Harvey is in his house. So, I >> So, Harvey Harvey is in his house. So, I know he's in Dallas. So, when you're know he's in Dallas. So, when you're coming back, let's let's let's give coming back, let's let's let's give penalty to Harvey and then go from penalty to Harvey and then go from there.

there. there. Um Um Um And in the meanwhile, I just want to And in the meanwhile, I just want to make sure that I just want to cover a make sure that I just want to cover a very small topic of people are bringing very small topic of people are bringing it up to me and everything. And there it up to me and everything. And there are two of them and I need actually are two of them and I need actually Harvey's opinion. So, Harvey, remember Harvey's opinion. So, Harvey, remember we talked about how we're going to scale we talked about how we're going to scale AI. And this is Tailscale. I don't know.

AI. And this is Tailscale. I don't know. Are you able to see my screen? Are you able to see my screen? >> Yeah, I see it. >> Yeah, I see it. >> So, this is >> So, this is >> Tailscale.

>> Tailscale. >> Tailscale. >> You do, right? That This is So, this is >> You do, right? That This is So, this is now connected to my three Macs at the now connected to my three Macs at the moment in time and this is the latest moment in time and this is the latest one that I got for the company, by the one that I got for the company, by the way.

way. way. Which has 128 GB of unified memory. Which has 128 GB of unified memory. And what I was able to do was load so And what I was able to do was load so many models and tested. Now, um I'm just many models and tested. Now, um I'm just not going to give you the results, of not going to give you the results, of course. I'm going to give you course. I'm going to give you >> [laughter] >> [laughter] >> [laughter] >> the amount of research that goes into >> the amount of research that goes into it. Uh this is my research platform, by it. Uh this is my research platform, by the way, Corpus and then you go into, the way, Corpus and then you go into, you know, AI and infrastructure. We look you know, AI and infrastructure. We look at, you know, infrastructure providers, at, you know, infrastructure providers, if I'm going to go with Cerebras, all of if I'm going to go with Cerebras, all of this, price comparison, speed this, price comparison, speed comparison, accuracy comparison. It's comparison, accuracy comparison. It's just an insane amount of work just to just an insane amount of work just to understand that. And then of course we understand that. And then of course we have model cards, you know, what are the have model cards, you know, what are the agentic that we can use, what is the agentic that we can use, what is the Gemma? And by the way, Gemma? And by the way, since I'm going to forget, I'm going to since I'm going to forget, I'm going to talk about this QAT quantization. And talk about this QAT quantization. And then of course then of course the issue of Qwen, which has all the way the issue of Qwen, which has all the way from 2B, 4B, 9B, 27B,

and now 36B. So I from 2B, 4B, 9B, 27B, and now 36B. So I actually tested this out on a local actually tested this out on a local model, and it is giving me 100 tokens model, and it is giving me 100 tokens per second speed. Just to let you know per second speed. Just to let you know by comparison, 100 tokens per second is by comparison, 100 tokens per second is considered production grade.

considered production grade. So So So and then there is an equivalent by the and then there is an equivalent by the way Gemma model which came out, uh which way Gemma model which came out, uh which we are potentially going to move forward we are potentially going to move forward with, but after testing. But here's my with, but after testing. But here's my problem, Harvey.

problem, Harvey. Uh except for Google that released Uh except for Google that released Gemma, Gemma, Gemma, it is an onslaught. Like it's insane it is an onslaught. Like it's insane that how By the sorry, just to be fair that how By the sorry, just to be fair to US, Inkling is another one, we can to US, Inkling is another one, we can talk about this. But anyway, it is an talk about this. But anyway, it is an onslaught of models that are coming onslaught of models that are coming from Chinese companies, which includes from Chinese companies, which includes these small models that I use as the these small models that I use as the router models and tool calling models.

router models and tool calling models. Then we have these middle-sized models Then we have these middle-sized models which we have Qwen 3.6, 27B, etc., which which we have Qwen 3.6, 27B, etc., which honestly, at this level in time, is way honestly, at this level in time, is way better than Gemini Flash 2.5, which was better than Gemini Flash 2.5, which was considered a workhorse. And then lastly, considered a workhorse. And then lastly, frontier level models, especially Kimik frontier level models, especially Kimik 83, 83, 83, which like blew the waters off for which like blew the waters off for people.

people. people. So So So first step back, first step back, which is very interesting now, what the which is very interesting now, what the hell are you going to do with space?

hell are you going to do with space? Because, you know, you if you every Because, you know, you if you every single GPU squeeze, you can do so much single GPU squeeze, you can do so much more in space with these, you know, more in space with these, you know, small, medium, frontier level models.

small, medium, frontier level models. That's That's going to be the last one. That's That's going to be the last one. So, first of all, A, what are your So, first of all, A, what are your thoughts in general how the industry is thoughts in general how the industry is moving and how US companies are going to moving and how US companies are going to move in terms of model development and move in terms of model development and all of that and cheap human consumption?

all of that and cheap human consumption? Then the second one would be that should Then the second one would be that should we even use Chinese models? Because we even use Chinese models? Because that's a big debate. Should we actually that's a big debate. Should we actually ban them? Is there a significant ban them? Is there a significant security risk?

security risk? I think I need your opinion on that. And I think I need your opinion on that. And then lastly, come back to how do you then lastly, come back to how do you project it forward in the next 3 to 5 project it forward in the next 3 to 5 years?

years? years? I'm sorry, I'm going to take a call, but I'm sorry, I'm going to take a call, but go ahead. I'm listening. go ahead. I'm listening. >> Yeah.

>> Yeah. >> Yeah. All right, buddy. So, and that is All right, buddy. So, and that is This is a good question. So, the way I This is a good question. So, the way I personally see it, personally see it, the pendulum is moving all over the the pendulum is moving all over the place and I see the following.

place and I see the following. Think of it from small, middle, large. Think of it from small, middle, large. On the small scale, we have our phones, On the small scale, we have our phones, we're using them daily, and we are able we're

using them daily, and we are able to start putting models here. Why does to start putting models here. Why does that model Why does that matter? It's that model Why does that matter? It's because if the model's on the phone, now because if the model's on the phone, now instead of me asking a question, waiting instead of me asking a question, waiting for the server to hit, going to AI, for the server to hit, going to AI, ChatGPT servers as an example, process, ChatGPT servers as an example, process, and then coming back, even though to us and then coming back, even though to us that may not mean much time, but as that may not mean much time, but as pointed he's on the plane, there's bad pointed he's on the plane, there's bad Wi-Fi, there's Wi-Fi, there's nothing that's reliable. Now, it's nothing that's reliable. Now, it's processing on your phone. So, with that processing on your phone. So, with that principle, I feel like several things principle, I feel like several things are happening. One, if you can use your are happening. One, if you can use your energy and you're not using energy and you're not using the data center in the sense that you're the data center in the sense that you're decreasing that bandwidth cuz think of decreasing that bandwidth cuz think of it this way. Why use a limousine to go it this way. Why use a limousine to go across the street? You wouldn't. You across the street? You wouldn't. You would just walk. Same idea. If you would just walk. Same idea. If you needed something to go on your phone and needed something to go on your phone and you need to process now, why run through you need to process now, why run through that system? So, the I'm giving you that that system? So, the I'm giving you that tail end story because here's the deal.

tail end story because here's the deal. With the structure point from a phone With the structure point from a phone point of view, we'll have the small point of view, we'll have the small language model, we'll be able to language model, we'll be able to process. We need a network. What does process. We need a network. What does that network look like? That's where that network look like? That's where Phantom Space I I like we'll have a Phantom Space I I like we'll have a leverage. We'll be able to put that leverage. We'll be able to put that network in outer space so that when Ed's network in outer space so that when Ed's flying around the world, we could could flying around the world, we could could just call us from his phone and he'll just call us from his phone and he'll connect with our satellites. Two, connect with our satellites. Two, the processing, think of it again, if the processing, think of it again, if it's running on here, why not put those it's running on here, why not put those models on a small level in the sense models on a small level in the sense that I'm putting it inside the chip, that I'm putting it inside the chip, that small language model model on the that small language model model on the satellites and now all of the satellites satellites and now all of the satellites are speaking. Now, you may all say, are speaking. Now, you may all say, "Well, Harvey, what happens 6 months "Well, Harvey, what happens 6 months later when the next model comes out?" later when the next model comes out?" Well, that's a good point. There'll be a Well, that's a good point. There'll be a system that will be able to system that will be able to redo some of those chips on the redo some of those chips on the satellite, but the thing that you don't satellite, but the thing that you don't realize is realize is these satellites are going up daily.

these satellites are going up daily. And so, with time, And so, with time, systems that are coming up will be systems that are coming up will be updating and then the ones from a year updating and then the ones from a year to years that yes, they're outdated, but to years that yes, they're outdated, but they're talking to these systems. So, they're talking to these systems. So, think of it as a network. Some of those think of it as a network. Some of those will be running really quick, some of will be running really quick, some of them will be running slow, but with them will be running slow, but with time, it's always going to get better time, it's always going to get better and better. So, that's kind of what I and better. So, that's kind of what I see the future of going. I know that was see the future of going. I know that was kind of long, but I was kind of long, but I was >> I'm I'm going to rephrase this into a >> I'm I'm going to rephrase this into a more technical term. The one that he more technical term. The one that he uses on the mobile is considered what we uses on the mobile is considered what we use the word edge.

use the word edge. That is another term we use is near edge That is another term we use is near edge is basically where my laptop right here is basically where my laptop right here sitting is working. So, there's edge on sitting is working. So, there's edge on literally on your device, then there's literally on your device, then there's near edge, basically a closed source near edge, basically a closed source system in which you can have this. Then system in which you can have this. Then there's far edge, which is private there's far edge, which is private clouds and the reason I'm telling you clouds and the reason I'm telling you this is because say like and talk to this is because say like and talk to hospital systems, they don't want to hospital systems, they don't want to even send any more data to cloud code even send any more data to cloud code and open AI, which is fantastic news for and open AI, which is fantastic news for us because what what what's happening is us because what what what's happening is

because they are really saying, "Oh, because they are really saying, "Oh, they're going to train the model on my they're going to train the model on my data." All of that data." All of that And then lastly, that is what we call And then lastly, that is what we call essentially, what he is saying is essentially, what he is saying is basically basically basically cloud operators. And cloud operators can cloud operators. And cloud operators can work in both ways. Cloud operators can work in both ways. Cloud operators can work as a private cloud.

work as a private cloud. I mean, Azure and all of them can I mean, Azure and all of them can basically build a complete section for basically build a complete section for you, and that's yours, etc. So, you're you, and that's yours, etc. So, you're not worried about infrastructure costs not worried about infrastructure costs and everything.

and everything. So, a few updates. I don't know if So, a few updates. I don't know if you've seen Apple has released iOS next. you've seen Apple has released iOS next.

It's going to come in September. They It's going to come in September. They are releasing what we call Core ML. And are releasing what we call Core ML. And that you're already all on the beta that you're already all on the beta server.

server. server. >> [laughter] >> [laughter] >> [laughter] >> I don't do that on my production phone. >> I don't do that on my production phone. I'm getting another phone just to try I'm getting another phone just to try their beta services.

their beta services. >> I I regret it sometimes, and I'm like, >> I I regret it sometimes, and I'm like, "Ah, I don't care." So, it's on there. "Ah, I don't care." So, it's on there. >> Yeah.

>> Yeah. >> Yeah. So, they actually have a newer So, they actually have a newer technology called Core ML. And then, of technology called Core ML. And then, of course, when I use it on one of my course, when I use it on one of my laptops, again, remember Apple silicon laptops, again, remember Apple silicon is different than CUDA on NVIDI and is different than CUDA on NVIDI and Arcom and AMD. Again, we test Arcom and AMD. Again, we test everything. So, CUDA has basically the everything. So, CUDA has basically the best way is VLLM right now. But, Apple best way is VLLM right now. But, Apple silicon has MLX, and with the addition silicon has MLX, and with the addition of Core ML, all my models became 30% of Core ML, all my models became 30% faster with the same GPU.

faster with the same GPU. So, again, that I that it's getting So, again, that I that it's getting insane. So, one was insane. So, one was software upgrade, like we do. And by the software upgrade, like we do. And by the way, just to let you know, my Tesla way, just to let you know, my Tesla works so much better now. I mean, I works so much better now.

## Healthcare Access and Frontier Care

I mean, I click self-supervising. It literally click self-supervising. It literally freaking parks itself. So, again, that freaking parks itself. So, again, that flywheel is getting insane. I was flywheel is getting insane. I was telling you that, you know, telling you that, you know, like in me, I mean, I was building a like in me, I mean, I was building a whole freaking module in 2 months.

whole freaking module in 2 months. Which took me previously 2 years to Which took me previously 2 years to build. build. build. So, [snorts] that flywheel is is getting So, [snorts] that flywheel is is getting in there. So, edge, then there is near in there. So, edge, then there is near edge. And I will just use the private edge. And I will just use the private cloud / cloud as the three major things cloud / cloud as the three major things as one of those things. Now, coming back as one of those things. Now, coming back to how to how hardware devices are improving. I don't hardware devices are improving. I don't know what the next model of M5 Ultra or know what the next model of M5 Ultra or M6 or M7 is going to come from Mac, M6 or M7 is going to come from Mac, which I believe at this point in time. I which I believe at this point in time. I don't know what your thoughts are, don't know what your thoughts are, Harvey. Is Apple going to start Harvey. Is Apple going to start competing with NVIDIA and AMD for competing with NVIDIA and AMD for actually GPU inference Uh, not? That's actually GPU inference Uh, not? That's question

number one. And then, question question number one. And then, question number two, I want to go back that edges number two, I want to go back that edges of course has still will still have its of course has still will still have its limitation at some point, at least for limitation at some point, at least for the next two three years.

the next two three years. Uh, it's already we are I'm already at a Uh, it's already we are I'm already at a two nanometer process and I think I two nanometer process and I think I don't know if you can squeeze the don't know if you can squeeze the molecules enough, but we have 3D molecules enough, but we have 3D refracting. So they're actually putting refracting. So they're actually putting the chip on top of the chip on top of the chip on top of the chip on top of the chip. So they're stacking.

the chip. So they're stacking. Anyway, it's getting insane again. By Anyway, it's getting insane again. By the way, designed fully by AI. the way, designed fully by AI.

That's another problem. AI is designing That's another problem. AI is designing chips today. We're going to get We're chips today. We're going to get We're We're going to have you know, Antichrist We're going to have you know, Antichrist soon. I mean, we have doctors for this soon. I mean, we have doctors for this before. So having all this sort of before. So having all this sort of information that I wanted to give to my information that I wanted to give to my audience, audience, audience, uh, uh, uh, what what what are your thoughts of what what what are your thoughts of Apple as a GPU provider in the future?

Apple as a GPU provider in the future? Number two is what is the role of the Number two is what is the role of the middle the section this 3.6 27B models middle the section this 3.6 27B models 3.6 this models. And then lastly, 3.6 this models. And then lastly, comment why am I not seeing this?

comment why am I not seeing this? Because my problem is that let's just be Because my problem is that let's just be honest. I mean, it is getting cheaper.

honest. I mean, it is getting cheaper. I'm getting a two or $2,000 Mac which is I'm getting a two or $2,000 Mac which is giving me 80 tokens per second. I can giving me 80 tokens per second. I can get 10 of them which is much better than get 10 of them which is much better than giving money to Azure. And then I am the giving money to Azure. And then I am the only problem right now is that I'm only problem right now is that I'm forced to use Chinese models which I forced to use Chinese models which I don't want to use in production don't want to use in production because of the security risk. So are because of the security risk. So are there security risks? Am I just being there security risks? Am I just being Let's just do it. And then lastly, Let's just do it. And then lastly, why why am American companies are losing why why am American companies are losing because you didn't talk about Kimik 83 because you didn't talk about Kimik 83 either which is by the way, I've been either which is by the way, I've been using it and it's god damn better than using it and it's god damn better than Opus 4.8. I haven't tried it with Opus Opus 4.8. I haven't tried it with Opus 5.

5. 5. A lot of information for you, Harvey. A lot of information for you, Harvey. I need your opinion. And And you need to I need your opinion. And And you need to learn this, man. I know you don't test learn this, man. I know you don't test this on a regular basis, but it is this on a regular basis, but it is you know, and that's what I was telling you know, and that's what I was telling you before the meeting. It is so hard you before the meeting. It is so hard for me to even present all of this.

for me to even present all of this. Do Do you get it? Like, I cannot sit Do Do you get it? Like, I cannot sit there and make presentations now. there and make presentations now.

>> Okay, buddy. Uh gosh, you got me like >> Okay, buddy. Uh gosh, you got me like you [laughter] were speaking. I'm like, you [laughter] were speaking. I'm like, uh that's how I'm drinking my coffee, uh that's how I'm drinking my coffee, but No, no. All seriousness the but No, no. All seriousness the following.

following. following. These models are getting better and These models are getting better and better, and I see two things to your better, and I see two things to your point.

point. point. The models are becoming better, so you The models are becoming better, so you don't to my analogy earlier, you don't don't to my analogy earlier, you don't need the limousine because now the need the limousine because now the limousine is the Ford, but the Ford has limousine is the Ford, but the Ford has gotten so much better that it's like gotten so much better that it's like riding the limousine, my my analogy. And riding the limousine, my my analogy. And so, with these models dropping and so, with these models dropping and dropping and these chips getting better dropping and these chips getting better and better to your first part of the and better to your first part of the question, what's going to happen?

question, what's going to happen? I call it the bell curve. We're going to I call it the bell curve. We're going to have everyday users like, let's not say have everyday users like, let's not say you and I, but the majority of the you and I, but the majority of the population that uses an iPhone, population that uses an iPhone, but the processing power that's going to but the processing power that's going to have is going to be so superior, and have is going to be so superior, and that's going to give the Apple a that's going to give the Apple a competitive advantage in that they have competitive advantage in that they have the numbers, they have so many people the numbers, they have so many people using the devices, and they'll be able using the devices, and they'll be able to create a network quicker. And another to create a network quicker. And another concept is we don't know what we don't concept is we don't know what we don't know, but by Apple having this leverage, know, but by Apple having this leverage, they're going to start releasing they're going to start releasing applications that the average person applications that the average person will be like, "Oh, I didn't know my will be like, "Oh, I didn't know my phone could do that." But to Janae and I phone could do that." But to Janae and I and Ed, we're like, "Oh, no, this is AI, and Ed, we're like, "Oh, no, this is AI, and this is how you do it." But for and this is how you do it." But for everybody else, they'll just get it to everybody else, they'll just get it to that point. And so, the reason it that point. And so, the reason it matters that their chips happens is cuz matters that their chips happens is cuz it starts bringing that people to their it starts bringing that people to their ecosystem. Now, what's interesting is ecosystem. Now, what's interesting is you have uh you have uh uh uh uh Elon Musk uh doing chips now. And when Elon Musk uh doing chips now. And when that starts hitting, that's going to that starts hitting, that's going to really be interesting because what's really be interesting because what's going to happen to Nvidia? If Nvidia and going to happen to Nvidia? If Nvidia and and uh and uh Elon Musk start fighting, you know, what Elon Musk start fighting, you know, what what will it be? It'll be interesting what will it be? It'll be interesting because because because he has the power, he has the AI, and he has the power, he has the AI, and when I say power, solar energy and all when I say power, solar energy and all that, and you put it all together, and that, and you put it all together, and soon he'll have the data centers in soon he'll have the data centers in space. He's going to have a crazy space. He's going to have a crazy ecosystem to give him a competitive ecosystem to give him a competitive advantage.

advantage. advantage. And so, to your point, what's going to And so, to your point, what's going to happen is these chips will start going happen is these chips will start going down. The amount of data that we need, down. The amount of data that we need, everybody's saying, "Oh, these edge everybody's saying, "Oh, these edge technology, all this stuff, we really technology, all this stuff, we really won't be able to We don't need data won't be able to We don't need data centers as much." I My argument is you centers as much." I My argument is you guys are not thinking about this truly.

guys are not thinking about this truly. If the price goes down, the demand is If the price goes down, the demand is going to start going up because more and going to start going up because more and more people will realize, "Let's use more people will realize, "Let's use this more and more." So, then that's this more and more." So, then that's going to actually going to increase the going to actually going to increase the demand. You and I didn't even know about demand. You and I didn't even know about an iPhone when we grew up, most of us.

an iPhone when we grew up, most of us. And so, now, look how much we're using And so, now, look how much we're using this thing. I'm freaking programming on this thing. I'm freaking programming on this when I'm on trips now, on vacation.

this when I'm on trips now, on vacation. I'm like like on this like left and I'm like like on this like left and right. I couldn't do that even a couple right. I couldn't do that even a couple years ago. So, my point is we don't know years ago. So, my point is we don't know what we don't know. The you The chips what we don't know. The you The chips are going to get better. We're going to are going to get better. We're going to start having applications that we didn't

start having applications that we didn't thought of. Some I'm trying to patent thought of. Some I'm trying to patent for space. I'll just put the put it on for space. I'll just put the put it on the side cuz I feel like this is coming.

the side cuz I feel like this is coming. Um where are we headed with all of this Um where are we headed with all of this is is is a a a a bridge where a a a a bridge where do we go with American models? Do we go do we go with American models? Do we go with outside models? And here's the risk with outside models? And here's the risk from a US point of view.

from a US point of view. If China gives away all these models, to If China gives away all these models, to Jana's point, where DeepSeek is pennies Jana's point, where DeepSeek is pennies compared to these others, and they're compared to these others, and they're literally just giving it away, what will literally just giving it away, what will happen is what?

happen is what? The standard of will become DeepSeek, The standard of will become DeepSeek, where people are like, "Dude, I'm just where people are like, "Dude, I'm just going to DeepSeek. It's cheaper, going to DeepSeek. It's cheaper, cheaper, cheaper." And then everybody cheaper, cheaper." And then everybody builds an infrastructure around these builds an infrastructure around these Chinese models. And then later they're Chinese models. And then later they're like, "You know what? We own 92% of the like, "You know what? We own 92% of the market right now." So, for that point, market right now." So, for that point, it makes it a risk for the United it makes it a risk for the United States. And if the US says, "No, uh States. And if the US says, "No, uh United States is going to cut all these United States is going to cut all these models out. You cannot use them." Then models out. You cannot use them." Then that's going to be an issue because that's going to be an issue because these models are so easy to download and these models are so easy to download and use yourself that it's going to be hard use yourself that it's going to be hard to enforce. Like, are you really using to enforce. Like, are you really using them or not? And then from a competitive them or not? And then from a competitive point of view, if the United States is point of view, if the United States is like, "Hey, this is a capitalist market.

like, "Hey, this is a capitalist market. We allow everybody to compete." Then We allow everybody to compete." Then that's going to be hard. How do you let that's going to be hard. How do you let How do you now How do you now say you're not going to allow others to say you're not going to allow others to compete? And so, what I foresee compete? And so, what I foresee happening, sorry, is I foresee them happening, sorry, is I foresee them getting This is just me, my biased getting This is just me, my biased opinion uh pushing hard also for the opinion uh pushing hard also for the open models to be just as good as those.

open models to be just as good as those. The reason I'm biased to say this is the The reason I'm biased to say this is the reason I believe some of these Chinese reason I believe some of these Chinese models are so strong is because they models are so strong is because they have the government behind it and it's have the government behind it and it's helping the companies out. Whereas, the helping the companies out. Whereas, the US is like, "No, OpenAI, Anthropic, you US is like, "No, OpenAI, Anthropic, you do your thing." And I think that is why do your thing." And I think that is why we're starting to see this merge where we're starting to see this merge where uh US government owns part of Intel.

uh US government owns part of Intel. Like, who would have thought of that? Like, who would have thought of that? And so, we're And so, we're >> 10% Not even a Not even a small portion, >> 10% Not even a Not even a small portion, 10%.

10%. 10%. >> I'll take 10% of that. >> I'll take 10% of that. >> [laughter] >> [laughter] >> [laughter] >> And so, think about what's going to >> And so, think about what's going to happen. If they're doing that are we happen. If they're doing that are we going to start pivoting towards these going to start pivoting towards these models where it's US and Anthropic and models where it's US and Anthropic and own together, US and OpenAI own own together, US and OpenAI own together? Where now Now it becomes a together? Where now Now it becomes a competitive thing where the US is like, competitive thing where the US is like, "We don't want to lose money to China." "We don't want to lose money to China." So, it'll be quite interesting. Where So, it'll be quite interesting. Where are we going to be? And this is today.

are we going to be? And this is today. And I think things are moving so fast And I think things are moving so fast that I'd love to have the same that I'd love to have the same conversation in 6 months because I think conversation in 6 months because I think the whole game will be different. Real the whole game will be different. Real fascinating.

>> Yep. So cool. >> Yep. So cool. Uh Uh Uh Ed, your thoughts as a general person Ed, your thoughts as a general person when you see the news in two things.

when you see the news in two things. Number one, security risk of Chinese Number one, security risk of Chinese models models models or non-Chinese models, I mean, coming or non-Chinese models, I mean, coming from anywhere. And then the second from anywhere. And then the second problem of what he said, uh the stack problem of what he said, uh the stack building. And I just to let you know, it building. And I just to let you know, it is an insane amount of work for me just is an insane amount of work for me just to give you a little bit of context. If to give you a little bit of context. If I need to fine-tune a model, then I have I need to fine-tune a model, then I have to choose a model, then I have to make to choose a model, then I have to make great fine-tuned versions of it, then I great fine-tuned versions of it, then I have to check Laura heads, then I have have to check Laura heads, then I have to check inference times, then I have to to check inference times, then I have to check if the model is going to be check if the model is going to be supplied on an Nvidia hardware, AMD supplied on an Nvidia hardware, AMD hardware, or Apple silicon. Is it giving hardware, or Apple silicon. Is it giving consistent results? That very Making a consistent results? That very Making a fine-tuned model is joke for me. This fine-tuned model is joke for me. This whole testing takes 3 months.

whole testing takes 3 months. So, when I choose something in terms of So, when I choose something in terms of a stack, in terms of a GPU, a model, or a stack, in terms of a GPU, a model, or anything, I'm sticking to it for the anything, I'm sticking to it for the next 3 years. Because it's not not a next 3 years. Because it's not not a small amount of work to go through. So, small amount of work to go through. So, just to let everyone understand what just to let everyone understand what Harvey's meaning is that what every Harvey's meaning is that what every means by stack, it is not It is insane means by stack, it is not It is insane amount of work to And I'm telling you as amount of work to And I'm telling you as as as as a I have a company with AI. So, as as as a I have a company with AI. So, Ed, uh Ed, uh thoughts your thoughts on everything we thoughts your thoughts on everything we just talked about.

just talked about. >> Well, it's freezing. Like I I I I was >> Well, it's freezing. Like I I I I was just in all these other hot states or just in all these other hot states or Texas, and it's like cold, like really Texas, and it's like cold, like really cold in cold in in Oregon, and it's late July.

in Oregon, and it's late July. Yeah, so if I shiver, you'll understand Yeah, so if I shiver, you'll understand So, well, this is, you know, the tech So, well, this is, you know, the tech things, and you kind of alluded to it. I things, and you kind of alluded to it. I I really leave that to you all. You I really leave that to you all. You You're You're actually both much more You're You're actually both much more tech-savvy when it comes to all this tech-savvy when it comes to all this AI type of capability.

AI type of capability. But, on the Yeah, on the on the China But, on the Yeah, on the on the China side, those sort of things, you got to side, those sort of things, you got to think about like Airbus. You know, we think about like Airbus. You know, we have Boeing and American companies have Boeing and American companies competing with Airbus, which is, you competing with Airbus, which is, you know, know, know, government-funded, EU-funded. And it's government-funded, EU-funded. And it's the same with any country that we the same with any country that we compete with. They're A lot of them are compete with. They're A lot of them are government-funded. And government-funded. And so, the but the big difference here so, the but the big difference here obviously is, you know, the LLM and and obviously is, you know, the LLM and and the security issues and all that kind of the security issues and all that kind of stuff. So, I do think there needs to be stuff. So, I do think there needs to be a lot of caution um when it comes to a lot of caution um when it comes to that sort of thing. So, I don't want to that sort of thing. So, I don't want to opine on the technical, cuz you I think opine on the technical, cuz you I think you already covered it uh pretty well.

you already covered it uh pretty well. And And again, it's it's a little bit And And again, it's it's a little bit out of my out of my knowledge frame.

out of my knowledge frame. >> So, let me let me just give you a little >> So, let me let me just give you a little bit of background on the model. The bit of background on the model. The model is essentially still a black box.

model is essentially still a black box. So, they have training and they have So, they have training and they have this. Now, this. Now, it is a big issue, right? I mean, it is a big issue, right? I mean, everything is built on top of everything is built on top of everything, right? I mean, I'm not everything, right? I mean, I'm not writing a calculator app. There are tons writing a calculator app. There are tons of it on the of it on the available open source. So, I'm going to available open source. So, I'm going to use that. I mean, it's just not use use that. I mean, it's just not use feasible. So, there's a whole thing.

feasible. So, there's a whole thing. This is called S-bomb. It is actually This is called S-bomb. It is actually software bill of material. software bill of material.

That I have to submit to FDA. What that That I have to submit to FDA. What that means is that that I need to tell them means is that that I need to tell them that exactly any open source libraries that exactly any open source libraries I've used and I have actually created a I've used and I have actually created a whole process in which I have literally whole process in which I have literally at the library level at the library level any security risks.

any security risks. By the at this is how invasive in FDA By the at this is how invasive in FDA is. When I was saying bend over, you is. When I was saying bend over, you know, they were really bending me over.

know, they were really bending me over. So, when I say that, that's why I'm So, when I say that, that's why I'm telling you. telling you. Now, the problem with the model is, at Now, the problem with the model is, at least in this one, I know this is a least in this one, I know this is a Python code, this is JSON code. We have Python code, this is JSON code. We have AI reviewers, we have human reviewers, AI reviewers, we have human reviewers, we have security checks, tons of we have security checks, tons of How do I deal with a black box?

How do I deal with a black box? Do you do you do you guys understand? Do you do you do you guys understand? How do I go ahead and convince anyone How do I go ahead and convince anyone that okay, I've checked the internals of that okay, I've checked the internals of it? Literally, software bill of it? Literally, software bill of materials. So, think of it you build a materials. So, think of it you build a house, which electrical wire you use, house, which electrical wire you use, which, you know, kitchen counter you which, you know, kitchen counter you use, which door you use. That is the use, which door you use. That is the level that we need. And then the models level that we need. And then the models do not give you because they cannot give do not give you because they cannot give you. And who knows what they have you. And who knows what they have injected? And believe it or not, in in injected? And believe it or not, in in prompt injection, security risk is in prompt injection, security risk is much harder to identify, first of all.

much harder to identify, first of all. And then also even harder to fix because And then also even harder to fix because they can have layers of prompt they can have layers of prompt injections that can go through. Anyway, injections that can go through. Anyway, so my thought is that in that case, I so my thought is that in that case, I don't think so this is ready for don't think so this is ready for production, in my opinion.

production, in my opinion. Uh we are uh always interested. We are Uh we are uh always interested. We are thankful to Inkling, which is by Maria thankful to Inkling, which is by Maria Moradi, which actually did release a Moradi, which actually did release a very amazing very amazing model, model, model, uh, and as a completely new business uh, and as a completely new business process. And I think that is where the process. And I think that is where the winning process will be. She basically winning process will be. She basically provides the infrastructure, the model, provides the infrastructure, the model, and the pre-training and fine-tuning and the pre-training and fine-tuning services. And then basically, "Hey, services. And then basically, "Hey, bring your business, bring your data.

bring your business, bring your data. Here's your model, fine-tuned, and I'll Here's your model, fine-tuned, and I'll take care of the rest." And I think take care of the rest." And I think that's my in my thinking, Harvey, is that's my in my thinking, Harvey, is where the industry is leading to.

where the industry is leading to. We're going to cut it today short. Any We're going to cut it today short. Any other comments, Ellen Harvey, that you other comments, Ellen Harvey, that you want to add? Then we can join next week.

want to add? Then we can join next week. >> Yeah, yeah, let me let me change the >> Yeah, yeah, let me let me change the topic completely cuz I I'm about to do topic completely cuz I I'm about to do this talk for, uh, this talk for, uh, you know, the the country is divided in you know, the the country is divided in four or five regions by the federal four or five regions by the federal government in terms of community health government in terms of community health centers. These would be FQHCs, migrant centers. These would be FQHCs, migrant care facilities, those sorts of things.

care facilities, those sorts of things. And and, uh, And and, uh, and so I'm speaking and so I'm speaking today to the Northwest. That's why is today to the Northwest. That's why is this is Alaska, uh, Oregon, Washington, this is Alaska, uh, Oregon, Washington, Idaho.

Idaho. Idaho. And so I did a lot of pre-work to And so I did a lot of pre-work to prepare. And I was shocked by two by one prepare. And I was shocked by two by one thing and then one thing lends really thing and then one thing lends really well into the book that Harvey is about well into the book that Harvey is about to release.

to release. So the first thing is when I I So the first thing is when I I interviewed four CEOs of these various interviewed four CEOs of these various clinics. One of them is more advanced clinics. One of them is more advanced than any academic medical center I've than any academic medical center I've I've ever been to.

I've ever been to. I was shocked and I was like, "Wow, what I was shocked and I was like, "Wow, what a lucky community." Uh, she is like a lucky community." Uh, she is like totally on it, totally with it, totally on it, totally with it, deploying all sorts of AI capabilities deploying all sorts of AI capabilities that again, I've not seen anywhere else.

that again, I've not seen anywhere else. >> Let's bring her on to this podcast, too. >> Let's bring her on to this podcast, too. >> We We should. I will see her in a >> We We should. I will see her in a second. And then, uh, and then but but second. And then, uh, and then but but the majority of them, if there's a scale the majority of them, if there's a scale from one to five, uh, from one to five, uh, you know, and she's the five and let's you know, and she's the five and let's just say our academic medical centers just say our academic medical centers are fours. You know, the majority are are fours. You know, the majority are going to be around 1, 1.5.

going to be around 1, 1.5. And but they're looking to close that And but they're looking to close that gap very, very quickly. Uh they have gap very, very quickly. Uh they have they're going to have plans by the end they're going to have plans by the end of this event uh to close that gap in of this event uh to close that gap in 12, 24, 36 months. So, uh but think 12, 24, 36 months. So, uh but think about it. Uh about it. Uh you know, cuz cuz again, a lot of our you know, cuz cuz again, a lot of our audience and ourselves, we come from big audience and ourselves, we come from big cities, but a lot of care is done on the cities, but a lot of care is done on the frontier. And even Texas, 85% frontier.

frontier. And even Texas, 85% frontier. 85% of Texas is rural. 85% of Texas is rural. So, so think about the United States. I So, so think about the United States. I mean, it is a still a very rural mean, it is a still a very rural country. And a lot of health care is country. And a lot of health care is done out there where there's no, you done out there where there's no, you know, there's no academic medical center know, there's no academic medical center like there is in Dallas or Fort Worth.

like there is in Dallas or Fort Worth. So, uh so so I I find it uh super, super So, uh so so I I find it uh super, super interesting and their desire for AI and interesting and their desire for AI and just how AI can help bridge that gap, just how AI can help bridge that gap, right? Um on, "Hey, we don't need to go right? Um on, "Hey, we don't need to

go we we can do um we we can do um ultrasound. Anyone can do the ultrasound. Anyone can do the ultrasound. We can have someone else ultrasound. We can have someone else read it somewhere else, right? Uh and read it somewhere else, right? Uh and help guide them, guided ultrasound." help guide them, guided ultrasound." That saves so much money, saves uh That saves so much money, saves uh patient experience. Anyways, there patient experience. Anyways, there there's so much to unpack and I'm super there's so much to unpack and I'm super excited for for these people who are excited for for these people who are really doing health care. One one last really doing health care. One one last one last really quick example cuz it one last really quick example cuz it just blew my mind. There There's some, just blew my mind. There There's some, right? And then I know it's going to right? And then I know it's going to sound silly saying it cuz it's kind of sound silly saying it cuz it's kind of obvious, but until you really meet obvious, but until you really meet someone from an area like this, there's someone from an area like this, there's many parts, many communities in our many parts, many communities in our country, United States, country, United States, there's no road there's no road to that community.

to that community. Literally no road to that community. Literally no road to that community. They fly. Like if they need something, They fly. Like if they need something, they have to fly. They have to get into they have to fly. They have to get into an uncle's plane and and fly.

an uncle's plane and and fly. So, so this is so important what we're So, so this is so important what we're talking about. The need for AI and these talking about. The need for AI and these AI capable technologies to bring the AI capable technologies to bring the same level of care to them as we enjoy same level of care to them as we enjoy being in a big metroplex. And I think I being in a big metroplex. And I think I think we we I think we owe it to them to think we we I think we owe it to them to do it. That's why I'm so glad to be part do it. That's why I'm so glad to be part of of of Singles and Symptoms and and and be with Singles and Symptoms and and and be with you all because you know, we're we'll you all because you know, we're we'll help those people. So, anyways, that's help those people. So, anyways, that's my rant for the morning.

my rant for the morning. Yeah. Yeah. Yeah. >> And then of course everything is going >> And then of course everything is going to go through Harvey's intellectual to go through Harvey's intellectual property protected billion dollar property protected billion dollar applications that are going to be on applications that are going to be on hosted on Phantom Space and more. So, hosted on Phantom Space and more. So, Harvey, last words and then we'll close Harvey, last words and then we'll close for dinner.

for dinner. >> All right, guys, last words. Let's have >> All right, guys, last words. Let's have fun with this stuff. fun with this stuff. I'm going to three calls of actions.

I'm going to three calls of actions. Physicians, health care professionals, I Physicians, health care professionals, I want you to learn AI and figure out how want you to learn AI and figure out how can you use it today to help your can you use it today to help your patients. And just even if it's just patients. And just even if it's just something minor.

something minor. Second, developers, I want you to like Second, developers, I want you to like someone like Junaid, how can we pass someone like Junaid, how can we pass these things through the FDA and really these things through the FDA and really perplex, how can we use these other perplex, how can we use these other models, maybe the Chinese models, but models, maybe the Chinese models, but how can we use some of them? Let's think how can we use some of them? Let's think long-term American models so that if long-term American models so that if there is a stop, we can still continue there is a stop, we can still continue working and not worry. And then the last working and not worry. And then the last for the consumers out there, how can I for the consumers out there, how can I use this today? We know that we have use this today? We know that we have some wearables, we know we have some some wearables, we know we have some stuff. It's the bell curve for people stuff. It's the bell curve for people listening. I know there's somebody out listening. I know there's somebody out there that's listening that can program there that's listening that can program some of this stuff themselves. Just on some of this stuff themselves. Just on the side, I'm having fun with this the side, I'm having fun with this camera. It's literally giving me a check camera. It's literally giving me a check every day. When I speak, it tells me my every day. When I speak, it tells me my if I'm stressed, how much power's in my if I'm stressed, how much power's in my voice, if I'm ready to present, and it's voice, if I'm ready to present, and it's giving me grades and it sometimes I talk

giving me grades and it sometimes I talk to it really soft and like I'm relaxed to it really soft and like I'm relaxed and I know I'm stressed and it'll call and I know I'm stressed and it'll call me out and say, "Here's your stress me out and say, "Here's your stress number." I'm like, "Damn, I'm like number." I'm like, "Damn, I'm like faking it with what I'm saying and faking it with what I'm saying and everything." And it's the algorithm everything." And it's the algorithm catching it. And so my point is this, catching it. And so my point is this, let's leverage these tools to help let's leverage these tools to help humanity. Let's leverage these tools to humanity. Let's leverage these tools to help us in health care. And I just want help us in health care. And I just want >> Love it. Leverage AI to help humanity.

>> Love it. Leverage AI to help humanity. And then also hopefully in the same And then also hopefully in the same breath breath breath do not over expect that it's going to do not over expect that it's going to take over. People have been talking take over. People have been talking about, you know, doom scenarios for all about, you know, doom scenarios for all the other technologies in the past.

the other technologies in the past. It is okay. Have fun with it. It is okay. Have fun with it. >> [sighs and gasps] >> [sighs and gasps]