You don't want any one party controlling a platform technology.
You want it to be open source, you want closed source solutions,
private company solutions, open source solutions.
You want a range of opportunities.
And remember the last time we had some founders say,
trust me, I'll I'll get us some regulation.
That was SBF and he was he was going to be the one who got us
federal regulation for crypto.
And he's on trial at this very moment for a bunch of Meshugana.
This week in startups is brought to you by LinkedIn Marketing.
To redeem a $100 LinkedIn ad credit and launch your first campaign,
go to LinkedIn.com slash this week in startups.
Vanta compliance and security shouldn't be a deal breaker
for startups to win new business.
Vanta makes it easy for companies to get a sock to report fast.
Twist listeners can get $1,000 off for a limited time at Vanta.com
slash twist and CLA.
Innovation takes balance.
CLA's CPA's consultants and wealth advisors can help you get from
startup to where you want to end up.
All right, everybody, welcome back to this week in startups.
It is Monday.
It is Madra Monday.
Yes, that's right.
Every Monday, Money Madra joins us.
Sunny Madra of Definitive AI and we do our AI demos.
This week, Sunny, you told me you were having some deep weekend discussions.
We know Silicon Valley and the tech industry, on the weekends,
that's when everybody goes deep.
They don't have meetings, but the back channels start to light up.
The back channel was lighting up this weekend talking about what topic.
Well, open source, right?
There was a couple of big threads that kicked up this weekend.
The tech start flying and then I think we have a special opportunity.
We have a guest with us today.
Oh, okay.
Who do we have?
Yeah.
We've got the legendary, the goat, Bill Gurley.
All right, Bill Gurley, I think maybe third time.
No, only second time on this week in startups.
You were last on 2017, episode 722 for folks.
Yeah, so welcome back.

Bill Gurley, of course, from Benchmark and many great companies.
You're particularly passionate about open source, why, in relation to AI, Bill?
Well, I mean, obviously, there's a lot of answers to that question,
but most recently, when I did the regulatory capture speech at your conference,
I was mentioning at the very end that, oddly, some of what you might call
the early incumbents in AI software were running out and promoting the idea
that open source should somehow be curtailed or knee-capped.
In particular, it's quite notable that all of the loud voices are either executives
at these companies and or large investors at these companies.
And because of the way these companies raise money, the largeness of those investments
are in the tens and hundreds of millions of dollars.
So quite a bit at stake.
And there now seems to be a growing populace.
And this is what led to the conversation Sonia were having that are worried
that these companies are basically at the very start of this AI movement,
trying to cut off their biggest competition whatsoever.
And the thing that would probably unleash the most innovation and the most prosperity,
which would be if open source models were prevalent.
And this is more Sonny's world than me, so I'll let him comment.
But I do believe that the the stats about the performance of these open source models
is actually quite compelling.
And what I'm hearing internally from our portfolio is that specifically Lama, too,
but Sonny sent me another one today, are really starting to gain market share
amongst the startups that are using these tools.
So it would be really unfortunate if they found.
And when I when I spoke at your conference, I mentioned, you know,
this guy, George Stigler, who had won a Nobel Prize in academician
who had talked about how companies use regulatory capture.
And he said the two things they try and do is store competition and protect pricing.
Clearly, if they were successful in getting governments to block open source AI,
they would achieve both of those goals.
Yeah. And so that great talk, by the way, if you haven't seen it,
just do a Google search for all and some at Bill Gurley and the name of the talk
two thousand eight hundred fifty one miles about regulatory capture.
So, Sonny, when we look at the space, the open source space,
let's get to, you know, maybe the the tail of the tape, some metrics here on
how these open source projects are faring against the large language models
that are available by API call, right?
And I guess, paradoxically, open AI, which is now the most
closed of all the companies, but has opened in the name and started under the mandate
that this technology was too powerful for it to be closed and needed to be open
to everybody has gone exactly closed and has gone exactly regulatory capture
with Sam Altman begging, literally begging for the government

to get involved and to regulate.
So maybe we could talk a little bit about, you know, the what are the leading
open source projects? Yeah.
So let's just kind of level set, like, you know, who the players are,
because, you know, to Bill's point, there's many different people that are showing up.
And also, this is kind of a good lay of the land, because you see here
some startups, some real incumbents and some new folks that really kind of risen
to the top. And so just level setting, both anthropic and open AI, you know,
I'd say kind of leading models in the space, definitely closed.
Microsoft is kind of in this unique straddle, they're open sourcing lots
of unique pieces of content, additional frameworks that you need around
like LLMs, but they're gone all in with open AI.
Google started this whole thing, obviously, you know, the reason others
can get there is the papers are open source and people have written about them.
But we haven't seen an open model from Google in a while,
although, you know, they do support open models in their vertex platform.
And then Meta and then Databricks, and there's another one
we'll bring up Mistrol as well today, these folks are fully open.
Commercial use cases different, commercial use differs based on their licenses.
But in general, they're sharing everything and we're seeing really fast innovation.
So that's the level set.
And by the way, there's two quick things to add to what Sonny just said.
It's it's remarkably ironic that open AI, when Elon backed it was open source
and moved away from it.
And then second, as Sonny mentioned, I've been reading a ton this weekend
to try and get a lot deeper in this world.
And everyone points to this Google paper attention is all you need.
As the thing that allowed these LLMs to become successful and to really progress.
So you did.
You do have an academic paper with all Google people on it
that actually led to the technology that's being used by anthropic and open AI.
So I think that's ironic also.
Anyway, sorry to interrupt, Sonny.
No, and going deeper into the playing field, what's your analysis of what's happening here?
What I hear often people say is when you're in the lead, you're closed.
And when you're behind, you go open.
So iOS was in the lead and they and they kept it very closed.
Google search, their algorithm, they're in the lead.
They have a dominant position.
They'll never open up the search algorithm.
But then you have somebody like maybe Metta.
They feel like they're pretty far behind on language models.
So they open source it.

And the report was they internally leaked it accidentally on purpose.

All right, listen, when you're selling to business to business buyers,

you really want to get your pitch in front of decision makers.

Why? Because upper level execs are usually the ones making purchasing decisions.

Duh.

The problem is high level folks can be really hard to find and target on most social media platforms.

But on LinkedIn, oh my God, they know all of the CTOs, all of the CFOs,

all of the VPs of finance, engineering, HR, recruiting,

all those titles are sitting there waiting for you.

And now let's just talk about the funnel.

LinkedIn is about to hit a billion members.

Did you know that 950 million members at this point in time?

There are 180 million of those 950 who are senior level execs.

There are 10 million C level executives in that 180 million senior level execs,

which are part of the 950 million members.

I am a C level executive.

I am on LinkedIn all day long because LinkedIn equals business, business equals LinkedIn.

And LinkedIn ads are built specifically for B2B marketers.

LinkedIn generates two to five times higher return on ad spend than other social media platforms.

LinkedIn equals business, business equals LinkedIn.

When people are on LinkedIn, they're ready to do business.

It's that simple.

So make business to business, marketing, everything it can be.

And get $100 credit on your next campaign.

From me, your boy Jay Cow.

I'm sending you the Hyundai.

LinkedIn.com slash this week in startups to claim your credit.

That's LinkedIn.com slash this week in startups.

Terms and conditions apply because they're giving you the Hyundai.

So maybe some historical information here on when do companies choose open versus closed?

And to your point, Google was behind in cloud services and led a movement to open source Kubernetes.

So there's a company that has played both sides of the aisle, depending on where they are.

When you look at this field, OpenAI and Anthropic, these are startups.

Have you ever seen startups ask for regulation this early and this often and be so opposed to open source?

Is this a new trend?

And what are you attributed to?

I mean, Bill, there's more your area.

You answer that one and then I can chime.

Well, it's funny because I have certainly never seen it, Jason.

I've never seen this early.

And maybe it just speaks to the wild success of open source.

I mean, the number of venture backed open source companies today versus 30 years ago is just amazing.

It is a very disruptive way to get your technology out there quick and fast.

One of the other things that Stigler talked about, a phrase I mentioned several times,

he said, when you have this type of blocking regulation, you end up with a net loss to society.

And I'm one that firmly believes that open source is amazing for society because when you have technology locked up with patents, it's harder for ideas to spread.

It's harder for ideas to spread across borders and other countries among all these different smart people that can get out and innovate.

And so I have never seen something this early.

Obviously, there's a ton at stake.

There's a ton, like I mentioned, these companies have raised money at an unprecedented level.

So yeah, you could call them early state startups, but you could also, I mean,

they've raised billions of dollars each and only maybe in the ridesharing market.

Did you have that happen so quickly?

And so there's a lot at stake.

There's clearly a lot at stake, but I've never seen startups proactively pitch governments and not just ours, but several governments around the world.

I've never seen that.

I also find it really suspect that there aren't like technologically,

like academicians out there leading this charge, the people that are leading the charge calling for the regulation and calling and some of them raising this question of whether open source should be allowed are the incumbents.

They are the one either the incumbents are their backers.

And some of them, you know, Mustafa from Inflection podcast said,

I know it looks odd me being the one that's asking.

And yeah, it does.

It is odd.

And it would, it reeks of protectionism of pulling up the ladder behind you.

And with Microsoft and Google, they seem to be maybe trying to dance along the line here, Bill.

They want to have cloud computing.

They both have major cloud computing services, Azure, Google Cloud.

They want to offer these things, but they also have, you know, a proprietary use for this.

Obviously, Google search and Bard and the chat interface are going to overlap.

Microsoft trying to get Bing to break out using AI in the Office suite.

So they see that as a competitive advantage.

What's your, how would you handicap Google and Microsoft's behavior?

Well, I mean, I think to a certain extent, it is interesting on the chart that I don't know who produced the chart that Sunny put up, but it had Microsoft's neutral.

I could find myself believing that.

I mean, they had to do a rather convoluted deal to get access to the technology that they're, you know, using with open AI.

And I wouldn't be shocked if they're comfortable with a hedge on that.

Primarily because they already control these creative products that they now believe will be enhanced with AI.

And, and I don't know that whether it's open source AI or someone else's AI,

that it really impacts either of them because it, it's the lock in they have on the product.

So I wouldn't be surprised by that.

And like I mentioned, Google's played this both ways.

So, you know, they, they pseudo open source Android because within open source,

there's different dimensions on how open it is and whether you've really committed to a

third party like the Linux foundation that, that runs the regulatory aspect of it or not

regulatory because I don't want to confuse it with the government that runs, you know, how it's,

many, the most open projects have a third party that, that keeps it independent.

Open source foundation.

Yeah. There's others other than the open source foundation, but that's the largest one

that, that manage the process and enjoy it's not like that.

So anyway, I had to do that quick aside, but Google's done, whereas Kubernetes wide open

Linux foundation manages it.

And so they've, they've been all over the map.

And as you mentioned, they once posed a paper, we love open source, but it's not right for search.

You know, yeah.

Well, once you get that lock in, that's when you don't want to open source it because

people can then build competitive products.

And as we've seen in search, that space has not seen any changes in 20 years.

Like that has been a locked, you know, box where nobody's innovated for 20 years.

A number of people have tried.

I tried myself.

It's very hard to get any kind of a foothold in search.

But I don't think, you know, those two aren't at the forefront of this.

As you mentioned, it seems to be a battle between these extremely well funded startups

and, and really more of a community.

I'd say the people that are on the other side of this, based on what I saw going around this

weekend, it's more of a community.

I mean, it involves, you know, like Jim Zimlin, who, who, who runs the Linux foundation, who's

been out talking at your conference, I spoke to Steven Wolfram, who, who, who told me he thought

it was ridiculous that someone would try and ban open source here on a safety reason.

And so that's what I say when, if, if I might be more open to listening, if I thought it were,

you know, some broad group of technologists and big thinkers that were making this argument.

But all of the arguments are coming from the people with the most to lose.

Yeah. And that seems crazy.

Um, Sunny, you want to just give us an idea of how.

Well, yeah, I mean, you know, we're going to lose Bill in a couple of minutes here.

But like one thing I'd love to get his thoughts on, because I think it really reinforces the

point here is that.

So, you know, last week, we saw a big funding announce around a Mistroll.

It's like a European based group that, you know, raised a hundred plus million dollars to Bill's point, a lot of money.

And, you know, they released their model open source, open for commercial use as well.

And you can see a couple of key points here.

One, you can see that there's seven billion parameter model uses again, half the amount of memory.

And down here, even just against other open source models, right there, there are seven billion parameter model is outperforming, you know, Lamma to 13 billion.

Again, this is just against open source.

But like the rate of innovation is moving so quickly here that if some of that regulation were to come into play and this, these folks couldn't put this out there and they had to do it, you know, through some kind of regulator, which has been, you know, like the FDA or some of those are the ideas that we've heard put out there.

I think we just wouldn't see that.

And I'll just add one more thing.

Like last week, Jake, you know, we even demo GPTV.

And, you know, a week later, we get lava and lava is an open source implementation of like a vision model in the same example that we did there.

You know, we have it.

And I think, you know, to Bill's point, I can't get my head around, you know, what it is around open source that's bothering people.

We've seen open source in operating systems.

We've seen it in databases, right?

We've seen it in mobile phone operating systems.

You know, you get things get patched quicker when the code is open, you can find vulnerabilities.

And so the arguments that around safety, no one is providing the sort of the background as to, you know, what is it that is not safe here?

It's fairly obvious what's happening here, Bill.

The people who have the lead are using job destruction and the fear of AI from science fiction and this, you know, could get out of control.

The demon could be unleashed.

They're using that in order to maintain their lead because they know full well that large numbers of startups or mid-sized companies embracing an open source project would lead to the demise and would absolutely evaporate the lead of open AI.

And this really is about open AI and Sam Altman.

Let's call it what it is.

Sam is the one who's leading the charge.

Although Mustafa and Reid Hoffman have been perhaps even more vocal or at least openly vocal on podcasts and whatnot.

So it's not just that.

But once again, hundreds of millions of billions in the state.

Yeah, I mean, I agree with Sunny.

I mean, it's ironic, but Linux is the most stable, most secure operating system that's ever existed.

And I think, you know, I go back to some of the original thesis of why open source would work and more eyes the better, the more the more transparency, the better.
And the notion that the people that it's just so ridiculous for someone to say this stuff, super scary, like you should be really afraid.
But let me do it, you know, like I'm, you should trust me, but it's super scary.
But it can do really good things, but it's scary.
Let me take care of it.
You know, help me be the only one that gets to take care of it.
And yeah, that's sad.
I hope there's quite a few people stepping up.
I hope this doesn't happen.
One last thing I would mention before I have to go.
I think the cat's out of the bag.
So you're not going to stop there from being open source in parts of the globe.
So if you shut it down in a particular region, that region is going to fail to innovate relative to the other regions that are out there.
There's a, there's a really cool piece of open source technology called risk five, which you guys have talked about a couple of times in the semiconductor space.
And I believe they intentionally moved the, the governing body outside the US because they were afraid of the restrictions the US was putting on semiconductor technology.
And I think it was smart that they did that.
And, and risk five is going to be wildly successful, you know, regardless of what happens to us regulation.
And so I think governments need to be particularly careful that if they take a first step move here, they're going to put their own, you know, society in a worse place in their own entrepreneurism in a worse place than others around the globe.
I mean, and we've seen this play itself out with social networks and the impact they have on, you know, society at large, the fact that so much power was consolidated in meta.
And, oh, trust us, we'll be the, we'll protect you.
It doesn't work.
The person who's making a profit and has a profit mode of the incentive is too great to act in the public's interest, whereas a group of people working on the project together, they keep each other in check.
Yeah, that's like, it's a governance thing, right, Bill?
Undoubtedly.
And this is, this is an important message to spread immediately.
So I appreciate, I appreciate you guys having me on.
All right, Bill.
We appreciate you checking in here and a great continuation of your talk from all in.
Everybody, that's Bill.
Girl, we follow him on Twitter where he's super active.
If you're a SaaS or services company that stores customer data in the cloud, then you need to be, uh, SOC2 compliant.

You knew that from a third party.

And you need that third party to close big deals.

And if you want to get compliant easier and faster, you need to use Vanta, V-A-N-T-A.

Vanta makes it so easy for you to get and renew your SOC2.

On average, Vanta customers are SOC2 compliant in just two to four weeks.

Prepare that to three to five months without Vanta.

And Vanta can save you hundreds of hours of manual work and up to 85% of compliance costs.

This is a total no-brainer.

And Vanta does more than just SOC2 compliance.

They also automate up to 90% compliance for GDPR, HIPAA, and more.

You can't afford to lose out on major customers.

We all know that.

Listen, it's a hard year.

Last year was hard.

You can't lose those major customers because you don't have your compliance dialed in.

Just work with Vanta.

Get your compliance automated and tight and tight is right.

Lock down those big deals.

Here's the best part.

Vanta is going to give you $1,000 off.

That's 10 hundies.

Get $1,000 off at vanta.com.twist.

That's vanta.com.twist for $1,000 off your SOC2.

Okay, so let's go deeper into the actual models here because the demos are great.

You glossed right over the demo of what you called Lava, which I think.

I'm going to pull that back up.

I think we should stop.

We should start there because this is, I think, where the rubber meets the road.

If you're wondering why a friend of the pod, Sam Altman, might not want competition,

well, if you look at open source, last week we had the multi-modal chat GPT-4.

We were playing with it.

I have it on my phone now.

It's extraordinary.

It's amazing.

You take a picture, you upload a picture, and you ask it to do things with the picture.

One of the things we had to do was to make a hamburger recipe based on a hamburger that Sunny was interested in.

But there is an open source project called Lava, large language and vision assistant.

Got it.

So this is built on top of Lama, I take it, or it's its own project?

It's its own project and it's built to mimic the spirits of multi-modal GPT-4 and just kind of touching a little bit on the Microsoft being yellow in that chart that we talked about earlier.

You can see here, it's put together by researchers at University of Wisconsin in Madison, Microsoft, and Columbia.

And so the kind of really interesting group, and I think it does justify that,

Microsoft is still supporting open source, which is important here.

And it's a great paper.

We won't spend too much time looking at it, but I suggest people to go look at the GitHub URL, which is here.

But this doesn't have the multi...

Just type in LLAVA, large language and vision assistant.

And so this interface looks very similar to the chat GPT-4 multi-modal interface.

She said, give me instructions on how to prepare this.

This being a brioche bun, weak egg yolk hamburger that looks absolutely delicious.

And this is the same image we used last week, correct?

And so how did it do versus chat GPT-4?

Yeah.

So I'll say, look, chat GPT-V did a better job in terms of describing what it was and then the instructions.

But for me, what I'll say here is the fact that this is available and it's open and we can build from it and it's one week later, that's that race that's starting to collapse.

Now, we're not like 100% equal a week later, but being, my grade on this would probably be in comparison to that.

It's like maybe a b because it doesn't...

Here's the results right here.

Yeah.

If you gave them an a, you would give this student a b, which means, hey, this student applies themselves.

And if you were, let's say, a startup, would you tie your wagon to chat GPT-4,

a closed system with deep ties into Microsoft?

Or would you fork this and start building with your own hooks into it and wrapping?

What would you advise a startup you've invested in?

What would you advise them to do, Sunny?

Or would you have them split the baby and do both?

It would be like on a use case by use case, right?

I think it really depends on what value you're trying to build.

If you really want to have the whole stack in your control because that's how you can provide value, I think I would do this.

Like I would use Lava.

But if it's just a small feature within my product, then I would use chat GPT because I'm going to get there quicker and I don't have to worry about the infrastructure and scaling costs.

Right.

So the reason I didn't do this one live, it probably takes like two or three minutes to run.

Sure.

Because they don't...
And this goes back to the funding.
They don't have the billions of dollars behind it to have huge farms for inference.
And so this thing, it runs a bit slower.
And you can try it out live yourself as well.
J. Calla dropped the link for us here.
And I think...
Yeah, they don't have the 10,000 GPU cluster that OpenAI has.
Exactly.
But there's going to be cases where for your business, it's important that you build this and you build some IP around it.
And so I'll go back to a point that I made a couple of maybe episodes ago, which was the API for GPS, right?
The API for GPS on a cell phone been around for a long time.
Apple made it slightly better.
But companies were built when they built a lot of infrastructure around that, right?
And saying, hey, I'm going to build my app there and I'm not...
This is just one part of what I'm trying to do.
So I think you have to kind of look at in some places, how important is this to the core of what you're trying to build.
But I do want to reiterate, look how fast, how quickly this is happening on the backs of it.
And I do think, as I was great to have Bill here earlier, this is why there's so much noise around slowing this down.
Because if you're this company being valued at whatever 20 billion, 50 billion, 100 billion, these ridiculous valuations right now to have an open source model right on your heels, it's got to be really challenging.
I'll put that back to you as an investor.
Anybody who wants to buy shares in open AI at 90 billion for common shares,
I take it with a cap upside, you would probably want to monitor these open source solutions and say, well, if they're getting 90 times revenue or 100 times revenue, whatever it is, for the current valuation of open AI, is that a good bet?
Or are these open source projects going to create downward pressure?
And the downward pressure would provide is the API calls.
And what CHAT GPT-4 or open AI could charge for access to their language model is going to go down
because you could just fire up your own open source solution on your own servers.
And this is where Microsoft, Google, and of course, we didn't talk about Amazon's position here, which is I think they also invested in Anthropic.
Amazon's been very clear that they would like to be a neutral third party in all of this.
And so Amazon will make money no matter what.
They're going to have every language model on AWS.
And their interest is in continuing to lower the costs continually and just make up for it in scale.
But this could be a road to nowhere.

This could make open AI be like Seagate.

It's like a hard drive provider.

Like there's just not a lot of value captured there.

In the cloud today, for the most part, obviously, some people do this specifically.

But when you're asking for a server, you don't really care whether it's Intel or AMD

or some kind of thing that's virtualized running ARM underneath it.

And so especially now, if you have specific use cases that require that type of

a certain instruction set that only Intel provides for performance, you'll ask for it.

But I think those clouds have all those options available.

All right, everybody.

Stephen Estes is a principal at CLA.

Clifton Larson Allen is a professional service provider that specializes in CPA,

tax consulting and wealth advisory.

Welcome to the program, Stephen.

Thank you for having me.

So at what point should we start to seek out professional accounting and tax services?

I think there's a couple of tipping points, right?

One of them is really as soon as funding and equity-based comp come into play.

So whether it's a 500 case seed or pre-seed or a $10 million series A,

I mean, as soon as you're raising and you're looking to hire talent and give equity to those

people, you should have quality advisors for both legal and tax.

And like I said, another marker really is that foreign activity, right?

Having a foreign subsidiary or foreign founders, you can really find yourselves in hot water

real quick if you don't know what you're doing in that regard because all companies

have to play by the same rules in the international sandbox as whether it's a startup or Coca-Cola.

Get started right now at claconnect.com slash tech.

Let them know your boy, Jake, how I sent you.

claconnect.com slash tech to get started right now.

Let's go back to Mishra for a second.

I love this.

It's fascinating as well.

Yeah, I just ran this live while we were sitting here.

So, you know, folks that were watching, you could see that's the time frame.

Yeah, but...

Okay, so suffice to say, you run it on your own servers.

You could go faster and it's gonna just get faster every week.

And then, you know, these open source projects can sometimes have such a diversity of talent in them.

And because they don't have a command and control structure,

they're gonna have more interesting insights, right?

That's the nature of open source projects.

You might have these four or five, you know, developers in South America and then these

12 in, you know, Japan and then these six in Ukraine.

And they all have some other use case and they contribute to the model
in a way that's different than if Microsoft or Google or OpenAI are building software.
They don't need to ask permission to work on some part of the model or the open source project,
correct?
Yep.
Yeah, I mean, not at all.
It's permissionless.
Yeah.
Now, there might be some permissions that occur when you do want to, you know,
actually commit those changes.
Exactly.
Yeah.
So to go now, usually how these projects are governed is, you know, they, you know,
like Bill was mentioning, they have organizations that are built to, you know, kind of decide like,
you know, where the direction of the project is going.
And so those then directions, then basically the developers on the open source project
use those high level, like think about that as roadmap and vision say,
oh, we want to build this next and everyone aims towards it.
But nothing stops you from taking it if you don't agree with that and doing something and then,
you know, submitting it for commitment into the project and either they can take that
or you can fork it.
Many projects have been forked, right?
And so that's happened over the years as well.
That's always your option if the core project doesn't want to go off on your side quest,
you can just make a copy of any, you fork it.
And you then start a new project and we've seen that happen over and over and over again.
Okay, let's get back to Mistral.
M-I-S-T-R-A-L.
Mistral, yeah.
R-A-L, yeah, correct.
Okay, now you mentioned how many parameters.
Let's explain that for civilians listening.
Yeah.
Yeah, so the way I kind of put this into simple thinking is more parameters equals,
like parameters are like neurons in a brain.
And so if you look at a small organism, it doesn't, you know, has a very small brain,
doesn't have a lot of neurons.
I think it's generally considered that humans have somewhere between 40 and 80 billion neurons.
And so when we think about it, the more neurons exist, the more data that this has access to that,
not from like, well, it started from what it was trained on, but that data is then processed
and held in kind of these weird fragments that we've talked about before.
And so the more of that that's there, the more it's able to basically reason,
think, provide you really incredible results like we've seen.

And so these folks at Mistral have done an incredible job of creating a smaller model that performs like the larger ones, that's probably due to some of the technological approaches they've taken and also perhaps even the training data.

And they've shown that in knowledge and reasoning and across these different tasks, which, you know, math, code, which is really incredible.

I thought this was very exciting.

So the exciting thing here is, you know, people are taking different approaches.

They're making, you know, these language models more efficient and faster, smaller, better, cheaper, all of those things.

And how many language models is the open source community grinding on right now?

And when I say grinding, like making daily progress on, like, you know, of the, you know, it could be some projects that are abandoned, et cetera.

Yeah.

So like listing all the language model projects on GitHub or wherever, or a hugging phase, that's not productive.

But let's say major ones that are getting daily updates to them, right?

Because daily updates would be a sign that this thing is cooking.

So how many are there?

There's probably under 10 that are like, and so what we really have to do is maybe take a step back because the space is playing out in a really unique way.

The folks that are, there's a less set of people that are going after really, really large models like OpenAI that are general intelligence models, right?

Where more of the energy is going is smaller run in, you know, kind of more confined, you know, compute requirements.

And that area has a lot of models, like too many to count.

And what many folks have realized is that it's a better approach to go after a smaller model that's, you know, tuned or trained for specific tasks, then to try to compete with general purpose models.

Yeah.

And so that means that when we talk about this competition and regulatory capture, if you're OpenAI, if you're anthropic and you want to lock all this down, if you're Reid Hoffman or Musafa or Sam Altman, you know, Greg, whatever, you're saying, hey, slow it down, trust us, we're going to protect everybody.

You then have, they're under attack that, hey, I'm going to make a verticalized one that's just for audio, just for video, just for code, just for, you know, a specific language or a specific culture, whatever the vertical is, you're going to see, you know, these large language models be attacked by a thousand cuts by verticals and by thousands of people contributing to an open source model. Is the goal to do smaller, more nimble models like and have the best model with the least parameters? Is that like an attack vector here that people are trying to make these things smaller and more efficient and cheaper to run?

Yeah. I think that's the ultimate goal because what we've already seen in the last year and, you know, with the rise of NVIDIA stock was these really, really large models require, you know, compute and energy, like last people talk about the energy, but compute and energy,

when you have like a cluster of 10,000 or 100,000 GPUs, the amount of energy it's using is really substantial. I wonder what one of those costs to run a day, like one of those GPUs, you know, an H100 per day, if it's actually doing work, it's doing jobs.
Maybe the guys can look at this rather, but honestly, I think it was like the amount of electricity, more than like a house, it was like something really substantial that we can probably research in the background here. Well, then a house might cost $1,000 a month or something,
you know, so if these things cost $1,000 a month to run, and you've got 1,000 of them, that's a million dollars a month, it's 10 million a month in energy costs for 10,000 of them, which I think is what some of these big clusters are now doing. That's not nothing, you know, it's like $120 million a year and electricity costs on top of these things costing $40K each or whatever they cost. Yeah. The energy cost might be 25% of the yearly cost, so over four years, it might be the same. Yeah, it's crazy. Yeah, so I'm just like, I'm looking at it here while we're doing it, but it basically is saying like an H100 card runs at 700 watts, right?
And so think about that 24-7, that turns into like a really, really big night. That basically turns into, I would say, my guess is close to a megawatt a month. Yeah, so we have to then figure out what a megawatt a month cost. And here's from Reuters. Running ChatGPT is very expensive
for the company. Each query costs roughly $0.04 according to an analysis from Bernstein. If ChatGPT queries grow to a tenth the scale of Google Surgery require roughly $48 billion worth of GPUs initially and about $16 billion worth of chips a year to keep operational, which is just crazy. But I think this is going to plummet, right? It's going to go down 50%, 90% a year. And yeah, I get the sense this is going to be, if it's $0.04 a query, it's going to be $0.04 and then $0.04 and then nobody's going to even think about it. Kind of like storage became so de minimis. But the energy cost is pretty crazy. There's been reports that everybody's making chips. So OpenAI supposedly is looking into making chips. We heard last week that part of, I think, the anthropic deal was that Amazon was making their own chips and they wanted Anthropic to use it. Obviously Apple makes the M1, the M2, and whatever's in the A, whatever they're up to, A14, 15, 16, whatever's in the iPhone. So we now have everybody's going to make their own chips. I heard Google's going to make their own chips too. So now we have...
Well, Google's been doing it for a long time, right? They have the TPU, right? Which is, yeah, they've been doing that for a very long time. There seems to be some intense ramp up of this, because not only can't you get Nvidia chips, there's a line out the door for them, people maybe don't feel great about the pricing of them. So now everybody makes their own chips. So that is something I didn't anticipate. And so here's the headline from Friday, Microsoft to debut AI chip next month that could cut Nvidia GPU costs. And that's from the information.
So I started my career in actually making chips, but for networking equipment. And this happened in networking at the late 90s. Everyone needed their own chips to basically interface with the immense growth of the internet on the... Sort of the core side of the internet with optical, with fiber optics, right? And so everyone was making chips there. And there was a time and place somewhere between, say, 97 and 2001 where there would have been hundreds of chip

companies making front-end chips to interface with optical transceivers. That ultimately all consolidated down to maybe like three companies. And it happened for the same reason, like no one could get these chips, right? And then no one could develop them in ways that each particular vendor required. And I think we're just kind of seeing that movie again here. And we'll see an explosion, which will bring the cost down. And then ultimately, I think it'll consolidate as well. But those are the waves that we see in tech anyways, right? Where we see like expansion and consolidation and expansion again. And so a massive amount of investment and then it becomes commoditized or so cheap that people maybe some number of players bow out and let...

Or roll up and... Yeah, exactly. That's another way. I mean, we saw that. I predicted this in GPUs because we saw it with fiber. People were building out so much fiber that they overbuilt. And then I think Google and some other providers bought up a lot of that fiber, pennies on the dollar that have been overbuilt. So we might be in the overbelting, overbel phase. What other demos do we have for this week?

Yeah, so I just had like Mistroll, which we can run through that one. Actually, like we just... I had the paper there so I can pull that one up. And so this is their playground on hugging face as well, where we spend some time. All right, let's... Give me a drop in here, Jay Cal, and I'll send you this link too if you want to try one. Give me a question here.

Oh, question. Well, you know, how about what are the best restaurants in Napa?

That's it. I mean, this is like some live information. Who knows where they got this information? But it's screaming fast, even on hugging face. I mean, instant answer. So that's the first thing I'm noticing is that this is... And it, of course, got it exactly right. Bouchon, French Laundry, Solbar, I mean, Adhoc. Yeah, these are all great ones. Botag I know a bunch of these.

And if you asked it what...

Put those in a table and add the average cost of dinner. I wonder if it has anything there.

I don't think so. That would be like more live data. But it did pretty good here, you know, depending on what... I don't think so. Yeah, I don't think it'll have that. But like, yeah, so... Yeah. Yeah. But, you know, that's a tough one, I think, even for chat GPT. But what I really liked about this one, and I'm glad you did this question, was like... And this is what we're going to get as we put these out there. It first of all kind of says, hey, in order to provide the best, but here are my assumptions, right? And it talks about, hey, you can define best. And here's like a combination of food quality, good service and that. And then target audience and research, how it got there. I really, really am a fan of like, like you said, the speed. You could run this on your laptop if you want to, right? And I think that's really incredible here. Yeah.

So you could give this language model to an astronaut, you know, going to Mars. And if it lost contact with the internet, it would still be able to give those answers, which is just for people to think about this. If you were stranded on a desert island, having one of these language models fully trained on a laptop, you would be able to have like an increasingly impressive conversation that could wind up saving your life. If you asked it how to start a fire on a desert island, what would it say? Yeah. I mean, actually ask it, you know, I'm trapped. If you were trapped on a desert island, how would you start a fire? Now, if you didn't go to survival school and you had this, all of a sudden, you've got this bot with you on some foreign location. I wonder if it will actually even help you build that fire. You could, you find a piece of glass is one technique,

rubbing two sticks together is another obvious technique, trying to find flint or stones or metal that could spark something. Gather material, collect dry leaves, small branches, of course, that's kindling, find a suitable location, okay, create a fire pit, prepare the tinder, okay, this is all the same stuff, kindling, prepare fuel, create a fire layer structure, okay, ignite the fire, there are several methods to start a fire, such as using flint and steel, magnifying glass or a battery, I forgot the battery one. So that's interesting. So tell it to explain to me the techniques in number 10. See if it understands that we're asking it about its existing answer, because steps one through nine are preparing a fire, putting the wood together, but not actually sparking the fire. What we care about is actually starting it. I wonder if it will be able to teach us how to start a fire. Oh yeah, here we go, flint and steel. This method involves striking flint against steel rod, okay, magnifying glass on a sunny day focus, sunlight onto the tinder using magnifying glass, battery and steel wool, that works, yeah, lighter matches, yeah, if you have them. It's a pretty good answer. And what else do we have in the demos? Everybody loves the demos here. Yeah, so last week, we kind of, right at the end, we got into WhatsApp and the model and the AI. And I think one of the things I wanted to correct, I wanted to pull WhatsApp back up because it is available in the group chat. So I was getting a second here. Yeah, so you can do a group chat and include an agent. So because that was one of the reasons you didn't give it even a higher grade last week. Yeah, and I loved it last week. I was super impressed. And I just thought, you know, if my wife and I had a chat going and we had slash Gordon Ramsay, and we could, you know, ask Gordon, hey, here's what they, you know, here's what we have, what should we make?

Hey, we got, you know, some salmon and we've got some pasta, you know, we've got butter and cheese.

And it was like, okay, yeah, make some farfalli with salmon in it. Here's a recipe for you.

So this is our AI, you know, Twist WhatsApp group that I created. Let me get this over to the side. And in here, we have the meta agent, right? And so, and if you can see here, you can do sort of an at and it allows you to add meta AI and say, I am thinking about booking a trip to Napa.

What are the top five hotels? Okay, here we go. Let's see. And what's great about this is like I could then ask a follow up to it, right? It should respond back in a second.

Maybe they've gone slow as well now. Here we go. Well, here we go. Yeah,

Arbaj, Meadowood, Four Seasons. Yeah, I mean, it's not great. It's not terrible. Yeah, not great, not bad. What's a fun activity for the three of us to do while in Lake Tahoe over Christmas?

I mean, it's pretty obvious what you do in Lake Tahoe over Christmas, right?

I consider ice skating, taking a sleigh ride, enjoying a festive atmosphere. Yeah.

Yeah, we could have also gone skiing, but we're snowmobiling. Not a great answer.

Not a lot of snow at Christmas sometimes, though. Maybe it knows.

How about more answers? Additionally, you can explore emerald waters of clear kayak toward not in the winter. You can do that. So, you can see the meta AI is like a very rudimentary AI, I think. I'm not sure which model that is. Do you think that's Lama? It must be, right? Is there a language model? Yeah, it's Lama, right? Yeah, it's definitely that one. Yeah.

Yeah, I think that leaves a lot to be desired right now. This is where I think verticalized AI is going to be much better. When you ask about travel, you really need a travel AI. That's just that, right? Yeah. So, I think that that's where these models are going to wind up,

is that they'll be very verticalized and fine-tuned ones by vertical. And if you ask about food,
it should really be narrowing you down and not just giving you the generic one. It should be
just like when you search now for recipes on Google, it doesn't just give you 10 blue links.
It really is thoughtful and they have a lot more information on it. Listen,
it's been another amazing episode. Thanks for Bill Gurley for tuning in. It's pretty clear.
The community must fight for more open source anytime anybody says, trust us.
We will be the sole source. You should be wondering if you should trust them.
And just listen, you need to fight for open source and you don't want one private company.
It's not a dig to any... It's not a dig to Sam. It's not a dig to Anthropic.
You don't want any one party controlling a platform technology. You want it to be open
source. You want closed source solutions, private company solutions, open source solutions. You want
a range of opportunities. And remember the last time we had some founder say,
trust me, I'll get us some regulation. That was SBF. And he was going to be the one who
got us federal regulation for crypto. And he's on trial at this very moment for a bunch of
Meshugana. I mean, are you following that? Sunny, you were down the crypto rabbit hole.
Yeah. And very deeply and following it. And it's really sad to see. I think one of the...
I think one of the more... There's a lot of shocking things, not to say, but one of the
more shocking things I saw last week is they were running some kind of insurance service.
And there was a counter that was saying, oh, this much money is insured by this.
And that counter was basically like a random number generator that was picking a number
between 7,500 and 10,000. And I mean, talk about explicit fraud. There you go. Exactly.
I mean, it's deranged. Your point is, to say that this was a real business and
that they were just in over their heads when they were doing free meditated things like this.
And this is where all the chats eventually get dumped, all the emails, all the slacks,
everything, everybody winds up flipping. No matter what the case is, it's very rare that a group
of people will circle the wagons. Even the mob and the mafia had a hard time maintaining that.
And this is like their entire lives, their families, their traditions were around this.
Maybe the cartels on the margins can, but not a bunch of dopey kids who've been working together
for 18 months, they were committing fraud after fraud. In this case, they were using a random
number generator with some parameters on it to dupe the public when they were getting insurance,
if I'm understanding it correctly. And listen. And it takes a certain amount of evil to do that,
right? Because there are lots of cases people have gone sideways, like Theranos, Elizabeth Holmes,
and at the core, I think she was trying to make a testing machine.
That's the case of it maybe got ahead of you, right? You got ahead of your skis,
you were a bulls*** artist. You thought, I can fake it till I make it. This wasn't fake it till
you make it. This was, let's orchestrate a huge crime. And listen, if you're in the mob and you
flip, you get whacked. With these kids, if they flip, there's nobody there to whack them. It's
not like SBF's parents are like some criminal mastermind cartel. They're a bunch of dopey
Stanford professors who didn't raise their kid correctly. And they're not going to whack
the other members. They're all going to flip on each other. It's all going to come out. I
told everybody I will bet dollars to donuts that he gets over 30 years. I'm saying over 30 years.
One thing in defense. We take the over-the-under 30-year sentence.

I would take the over. Yeah. If I said 50, we should take the over-the-under. I think under. So 40 would be the line we have to think it through. What would you take at 40? Pardon? We take the over-the-under at 40 years. I think the under. I think it'll come right around that 30. Okay. There we go. Wow. Look at that. I set a pretty good line. So the line is probably 36 years. He gets up before retirement kind of thing. I think they're going to burn you made off him. Bernie Madoff got multiple life sentences. So I set the 30 year, but I actually think there's a chance he gets life. I think this could be a life kind of situation. Well, it's very rare that you catch somebody red-handed doing a multi-billion dollar crime. There's a small number of multi-billion dollar crimes in the history of humanity. It's hard to pull off. And so if you're going to put some kids in a local city who were dealing drugs 20 years, 30 years, and they were dealing a half million dollars in drugs, a million dollars in drugs. Let's just call it a million dollars in drugs. A million dollar drug cartel in Chicago. And this kid was stealing over a billion. I mean, it doesn't feel proportional, right? If he were to get anything but a magnitude more, because it's a thousand X that crime. This is serious crime, folks. You're spot on there. I just don't know how to do that relative calculation. I mean, there's people in jail today for marijuana use. It's allowed now.

Luckily, there is consensus. Trump, Biden, Obama before everybody is a line that we need to reverse those for nonviolent felons and really think about those. All right. Listen, another great episode. And listen, he's guilty. I've made my decision already. The evidence is there. The jury will make their decision, but I hope he gets a huge sentence. And that's it. All right. Sandeep Madra, they have it, folks. Definitive intelligence. If your company is looking for AI and analysis of big data and you need help, Sunny's your guy. Just email Sunny at definitive.io. And we'll see you all next time. Bye-bye.