Jensen Huang – Will Nvidia's moat persist?

103 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=HRBQ66XQTCO
https://www.youtube.com/watch?v=Hrbq66XqtCo

We've seen the valuations of a bunch of
software companies crash because people are expecting AI to commoditize software.
There's a potentially naive way of thinking about things, which is: look,
Nvidia sends a GDS2 file to TSMC. TSMC builds the logic dies, it builds the
switches, then it packages them with the HBM that SK Hynix, Micron, and Samsung make.
Then it sends it to an ODM in Taiwan where they assemble the racks.
Nvidia is fundamentally making software that other people are manufacturing, and if software
gets commoditized, does Nvidia get commoditized?

In the end, something has to
transform electrons to tokens. The transformation of electrons to tokens and making those tokens more valuable over
time is hard to completely commoditize. The transformation from electrons to
tokens is such an incredible journey.

Making that token is like making one molecule
more valuable than another molecule, making one token more valuable than another.
The amount of artistry, engineering, science, and invention that goes
into making that token valuable, obviously we're watching it happen in real time.
The transformation, the manufacturing, all of the science that goes in there is far from deeply
understood and the journey is far from over.

I doubt that it will happen.
We're going to make it more efficient, of course. The way that you framed the question
is my mental model of our company. The input is electrons, the output is
tokens. In the middle is Nvidia. Our job is to do as much as necessary and as little
as possible to enable that transformation to be done at incredible capabilities.
What I mean by "as little as possible," whatever I don't need to do, I partner with
somebody and make it part of my ecosystem.

If you look at Nvidia today, we probably
have the largest ecosystem of partners, both in the supply chain upstream and
downstream, all of the computer companies, application developers, and model makers.
AI is a five-layer cake, if you will. We have ecosystems across the entire five layers.

We try to do as little as possible, but the part that we have to do,
as it turns out, is insanely hard.

I don't think that gets commoditized.
In fact, I also don't think the enterprise software companies, the tools makers... Most software companies today are tool makers. Some of them are not. Some of them are workflow codification systems. But for a lot of companies, they're tool makers.
For example, Excel is a tool, PowerPoint is a tool, Cadence makes tools, Synopsys makes tools.
I actually see the opposite of what people see.

I think the number of agents is going
to grow exponentially, and the number of tool users is going to grow exponentially.
It's very likely that the number of instances of all these tools is going to skyrocket.
It's very likely that the number of instances of Synopsys Design Compiler is going to skyrocket, along with the number of agents using the floor planners, our layout tools, and our design rule checkers.

Today we're limited by the number of engineers.
Tomorrow, those engineers are going to be supported by a bunch of agents.
We're going to be exploring the design space like you've never seen before, and we're going to use the tools that we use today. I think tool use is going to cause the software companies to skyrocket. The reason why it hasn't happened yet is because the agents aren't good enough at using their tools yet.
Either these companies are going to build the agents themselves, or agents are going to get good enough to be able to use those tools.

I think it's going to be a combination of both.
I think in your latest filings, you had almost a $100 billion in purchase commitments with foundries, memory, and packaging. SemiAnalysis has reported that you will have $250 billion of these kinds of purchase commitments. One interpretation is that Nvidia's moat is really that you've locked up many years of these scarce components.
Somebody else might have an accelerator, but can they actually get the memory to build it?
Can they actually get the logic to build it?

Is this really Nvidia's big
moat for the next few years? It's one of the things that we can do that
is hard for someone else to do. We've made enormous commitments upstream. Some of it is explicit, these commitments that you mentioned. Some of it is implicit. For example, a lot of the investments that are upstream are made by our supply chain because I said to the CEOs, "Let me tell you how big this industry is going to be, let me explain to you why, let me reason through it with you, and let me show you what I see."

As a result of that process of informing,
inspiring, and aligning with CEOs of all different industries upstream, they're
willing to make the investments. Why are they willing to make the
investments for me and not someone else? The reason for that is because they know
that I have the capacity to buy their supply and sell it through my downstream.
The fact is that Nvidia's downstream supply chain and our downstream demand is so large,
they're willing to make the investment upstream.

If you look at GTC, people are marveled
by the scale of it and the people that go. It's a full 360 degrees, the entire
universe of AI all in one place. They're all in one place because
they need to see each other. I bring them together so that the
downstream can see the upstream, the upstream can see the downstream, and
all of them can see the advances in AI. Very importantly, they can all meet the AI
natives, all the AI startups being built, and all the amazing things happening so they can
see firsthand all the things that I tell them.

I spend a lot of my time informing, directly
or indirectly, our supply chain, partners, and ecosystem about the opportunity in front of us.
Some people always say, "Jensen, in most keynotes, it's one announcement after another."
With our keynotes, there's always a part of it that's a little torturous in the sense
that it almost comes across like education.

In fact, that's exactly on my mind.
I need to make sure the entire supply chain, upstream and downstream, the ecosystem,
understands what is coming at us, why it's coming, when it's coming, how big it's going to be,
and is able to reason about it systematically, just like I reason about it.
Regarding the moat as you describe it, we're able to build for a future.
If our next several years are a trillion dollars in scale, we have the supply chain to do it.
Without our reach, the velocity of our business... Just as there's cash flow, there's
supply chain flow, there's churns.

Nobody is going to build a supply chain for an
architecture if the business churns are low. Our ability to sustain the scale is only
because our downstream demand is so great. And they see it, they hear about
it, they see it all coming. That allows us to do the things we're
able to do at the scale we do them. I do want to understand more concretely
whether the upstream can keep up. For many years now, you guys have
been 2x-ing revenue year over year.

You've been more than tripling the amount of flops

you're providing to the world year over year. And 2x-ing at this scale now is really incredible.
Exactly. But then you look at logic. You're the biggest customer on TSMC's N3
node, and you're one of the biggest on N2. AI as a whole this year is
going to be sixty percent of N3. It's going to be 86% next year,
according to SemiAnalysis. How do you double if you're the majority?
And how do you do that year over year?

Are we in a regime now where the growth rate
in AI compute has to slow because of upstream? Do you see a way to get around this?
How do we build 2x more fabs year over year, ultimately?
At some level, the instantaneous demand is greater than the
supply upstream and downstream in the world. At any instant, we could be limited by the
number of plumbers, which actually happens.

The plumbers are invited to next year's GTC.
By the way, great idea. But that's a good condition. You want an industry where
the instantaneous demand is greater than the total supply of the industry.
The opposite is obviously less good. If we're too far apart, if one particular
component is too far away, the industry swarms it. For example, notice people aren't
talking very much about CoWoS anymore.

The reason for that is because for two years
we swarmed the living daylights out of it. We doubled, doubled, doubled on several doubles.
Now I think we're in fairly good shape. TSMC now knows that CoWoS supply has to
keep up with the rest of the logic demand and the memory demand.
They're scaling CoWoS and future packaging technologies at
the same level as they scale logic. This is terrific, because for a long time,
CoWoS and HBM memory were rather specialty.

But they're not specialties anymore. People now
realize they're mainstream computing technology. Of course, we're now much more able to
influence a larger scope of our supply chain. At the beginning of the AI
revolution, all the things that I say now, I was saying five years ago.
Some people believed in it and invested in it, for example, Sanjay and the Micron team.
I still remember the meeting really well where I was clear about exactly what was
going to happen, why it was going to happen, and the predictions of today.
They really doubled down on it.

We partnered with them across LPDDR and HBM
memories, and they really invested in it. It obviously has been tremendous for the company. Some people came
a little bit

later, but now they're all here. Each one of these bottlenecks
gets a great deal of attention. Now we're prefetching the
bottlenecks years in advance. For example, the investments that
we've done with Lumentum, Coherent, and the silicon photonics ecosystem over the last
several years really reshaped the supply chain.

We built up an entire supply chain around TSMC.
We partnered with them on COUPE, invented a whole bunch of technology, and licensed those patents
to the supply chain to keep it nice and open. We're preparing the supply chain through the
invention of new technologies, new workflows, new testing equipment like double-sided
probing, investing in companies, and helping them scale up their capacity.
You can see that we're trying to shape the ecosystem so that the supply
chain is ready to support the scale.

It seems like some bottlenecks
are easier than others. Scaling up CoWoS versus scaling up—
I went to the hardest one, by the way. Which is?
Plumbers. Plumbers and electricians. This is one of the concerns that I have about the doomers
describing the end of work and killing of jobs. If we discourage people from
being software engineers, we're going to run out of software engineers.
The same prediction happened ten years ago.

Some of the doomers were telling people,
"Whatever you do, don't be a radiologist." You might hear some of those videos still
on the web saying radiology is going to be the first career to go and the world is
not going to need any more radiologists. Guess what we're short of? Radiologists.
Going back to this point about how some things you can scale, and other things... How do you actually
manufacture 2x the amount of logic a year?

Ultimately, memory and logic
are bottlenecked by EUV. How do you get to 2x as many
EUV machines year over year? None of that is impossible to scale quickly.
All of that is easy to do within two or three years.
You just need a demand signal. Once you can build one, you can build ten, and
once you can build ten, you can build a million. These things are not hard to replicate.
How far down the supply chain do you go?

Do you go to ASML and say, "Hey,
if I look out three years from now, for Nvidia to be generating two trillion a year
in revenue, we need way more EUV machines"? Some of them I have to directly, some
of them indirectly, and some of them... If I can convince TSMC, ASML will be convinced.

We have to think about the critical pinch points. But if TSMC is convinced, you'll have plenty of EUV machines in a few years.

My point is that none of the bottlenecks last
longer than a couple of years, two, three years, none of them.
Meanwhile, we're improving computing efficiency by 10x 20x, and
in the case of Hopper to Blackwell, 30x to 50x. We're coming up with new algorithms
because CUDA is so flexible. We're developing all kinds of new
techniques so that we drive efficiency in addition to increasing capacity.
None of those things worry me.

It's the stuff that's downstream from us.
Energy policies that prevent energy from... You can't create an industry without energy.
You can't create a whole new manufacturing industry without energy.
We want to reindustrialize the United States. We want to bring back chip manufacturing,
computer manufacturing, and packaging. We want to build new things like EVs and robots.
We want to build AI factories. You can't build any of these things without
energy, and those things take a long time.

More chip capacity, that's a 2-3 year problem.
More CoWoS capacity, 2-3 year problem. Interesting. I feel like I have guests
tell me the exact opposite thing sometimes. In this case, I just don't have the
technical knowledge to adjudicate. The beautiful thing is
you're talking to the expert. True. I want to ask about your competitors.
If you look at the TPU, arguably two out of the top three models in the world,
Claude and Gemini, were trained on TPU.

What does that mean for Nvidia going forward?
We build a very different thing. What Nvidia built is accelerated
computing, not a tensor processing unit. Accelerated computing is used for all
kinds of things: molecular dynamics, quantum chromodynamics, data processing, data
frames, structured data, and unstructured data. It's also used for fluid
dynamics and particle physics.

In addition, we use it for AI.
Accelerated computing is much more diverse. Although AI is the conversation today and
is obviously very important and impactful, computing is much broader than that.
Nvidia has reinvented the way computing is done, moving from general-purpose
computing to accelerated computing. Our market reach is far greater than
any TPU or ASIC can possibly have.

If you look at our position, we're the only
company that accelerates applications of all kinds. We have a gigantic ecosystem. So all kinds of frameworks and algorithms run on Nvidia. Because our computers are designed to be operated by other people, anyone who's an operator can buy our systems. With most of these home-built systems, you have to be your own operator because they were never designed to be flexible enough for others to operate. Because anybody can operate our systems, we're in every cloud, including Google, Amazon, Azure, and OCI.

If you want to operate it to rent, you
better have a large ecosystem of customers in many industries to be the offtakers. If you want to operate it for yourself, we obviously have the ability to help you operate it yourself, like we did for Elon with xAI. And because we can enable operators in any company and any industry, you could use it to build a supercomputer for scientific research and drug discovery at Lilly.

We can help them operate their own
supercomputer and use it for the entire diversity of drug discovery and biological sciences that we accelerate. There are just a whole bunch of applications that we can address that you can't do with TPUs. Nvidia built CUDA to be a fantastic tensor processing unit as well, but it also handles every life cycle of data processing, computing, AI, and so on.

Our market opportunity is just a lot
larger, and our reach is a lot greater. Because we support every application in the world now, you can build Nvidia systems anywhere and know that there will be customers for it. It's a very different thing. This is going to be a long question. You have spectacular revenue, and you're not making $60 billion
a quarter from pharma and quantum. You're making it because AI is an unprecedented technology that is growing unprecedentedly fast.

The question then is what
is best for AI specifically. I'm not in the details, but I talk to my
AI researcher friends and they say, "Look, when I use a TPU, it's this big systolic array that's perfect for doing matrix multiplies, whereas a GPU is very flexible. It's great when you have lots of branching or irregular memory access." But what is AI? It's just these very predictable matrix multiplies again and again and again. You don't have to give up any die area for warp schedulers or switches between threads and memory banks.

And the TPU is really optimized for the bulk

of this growth in revenue and use case for compute that is coming online right now. I wonder how you react to that. Matrix multiplies are an important part of AI, but they're not the only part. If you want to come up with a new attention mechanism, disaggregate in a different way, or invent a whole new type of architecture altogether—like a hybrid SSM—you want an architecture that's generally programmable. If you want to create a model that fuses diffusion and autoregressive techniques, you want an architecture that's just generally programmable.

We run everything you can imagine. That's the advantage. It allows for the invention of new algorithms a lot more easily, because it's a programmable system. The ability to invent new algorithms is really what makes AI advance so quickly. TPUs, like anything else, are impacted by Moore's Law, which we know is increasing by about 25% per year. The only way to really get 10x or 100x leaps is to fundamentally change the algorithm and how it's computed every single year. That's Nvidia's fundamental advantage. The only reason we were able to make Blackwell to Hopper 50x... When I first announced Blackwell was going to be 35x more energy efficient than Hopper, nobody believed it.

Then Dylan wrote an article saying I sandbagged, and it's actually fifty times. You can't reasonably do that with just Moore's Law. The way we solve that problem is with new models, like MoEs, that are parallelized, disaggregated, and distributed across a computing system. Without the ability to really get down and come up with new kernels with CUDA, it's really hard to do.

It's the combination of the programmability of our architecture and the fact that Nvidia is an extreme co-design company. We can even offload some of the computation into the fabric itself, like NVLink, or into the network with Spectrum-X. We could affect change across the processors, the system, the fabric, the libraries, and the algorithm simultaneously. Without CUDA to do that, I wouldn't even know where to start. My sponsor Crusoe was among the first clouds to offer NVIDIA's Blackwell and Blackwell Ultra platforms.

And they just announced their NVIDIA Vera Rubin deployment scheduled for later this year. But access to state-of-the-art hardware is only part of the story. For example, most inference engines already do KV caching for a single user's forward passes. But Crusoe does it across users and GPUs. So if a thousand agents are running on the same system prompt, Crusoe only has to compute the KV cache once for it to become available to every single GPU in the cluster. This is especially important as systems get more agentic and require much longer prefixes

in order to use tools and access files.

In a recent benchmark, Crusoe was able to
deliver up to 10x faster time-to-first token and up to 5x better throughput than vLLM.
This is just one among many reasons that you should run your inference workload with Crusoe.
And if you need GPUs for training, you don't need to switch clouds.
Crusoe's got you covered there too. Go to crusoe.ai/dwarkesh to learn more.
This gets at an interesting question about Nvidia's clientele. 60% of your revenue is
coming from these big five hyperscalers.

In a different era with different customers—let's
say professors running experiments—they need CUDA. They can't use another accelerator. They
just needed to run PyTorch with CUDA and have everything optimized.
But these hyperscalers have the resources to write their own kernels.
In fact, they have to in order to get that last 5% of performance they need
for their specific architecture. Anthropic and Google are mostly running their
own accelerators or running TPUs and Trainium. But even OpenAI, using GPUs, has Triton
because they need their own kernels.

Down to CUDA C++, instead of using cuBLAS
and NCCL, they've got their own stack which compiles to other accelerators as well.
If most of your customers can and do make replacements for CUDA, to what extent
is CUDA really the thing that is going to make frontier AI happen on Nvidia?
CUDA is a rich ecosystem. If you want to build on any computer first, building
on CUDA first is incredibly smart.

Because the ecosystem is so
rich, we support every framework. If you want to create custom kernels... For
example, we contribute enormously to Triton. So the back end of Triton has
huge amounts of Nvidia technology. We're delighted to help every
framework become as great as it can be. There are lots and lots of frameworks.
There's Triton, vLLM, SGLang, and more. Now there's a whole bunch of new
reinforcement learning frameworks coming out, like verl and NeMo RL.
With post-training and reinforcement learning, that entire area is just exploding.
So if you want to build on an architecture, building on CUDA makes the most sense
because you know the ecosystem is great.

You know that if something happens,
it's more likely in your code and not in the mountain of code underneath.
Don't forget the amount of code you're dealing with when building these systems.
When something doesn't work, was it you or was it the computer?

You would like it to always be you and to be able to trust the computer.
Obviously, we still have lots of bugs ourselves, but our system is so well wrung out that you can at least build on top of the foundation.

That's number one: the richness,
programmability, and capability of the ecosystem. The second thing is, if you're a developer building anything at all, the single most important thing you want is an install base.
You want the software you write to run on a whole bunch of other computers.
You're not building software just for yourself. You're building it for your fleet or everybody else's fleet because you're a framework builder. Nvidia's CUDA ecosystem is ultimately its great treasure.

We have several hundred million GPUs out there
now. Every cloud has it. It goes back to the A10, A100, H100, H200, the L series, the P series.
There's a whole bunch of them. They're in all kinds of sizes and shapes.
If you're a robotics company, you want that CUDA stack to actually run in the robot itself.
We're literally everywhere. The install base means that once you develop the software or the model, it's going to be useful everywhere. That is just incredibly valuable. Lastly, the fact that we're in every single cloud makes us genuinely unique.

If you're an AI company or
developer, you're not exactly sure which cloud service provider you're going to partner with or where you'd like to run it. We run everywhere, including on-prem for you if you like. The combination of the richness of the ecosystem,
the expansiveness of the install base, and the versatility of where we are makes CUDA invaluable.
That makes a lot of sense.

I guess the thing I'm curious
about is whether those advantages matter a lot to your main customers.
There's many people for whom they might matter. The kind of person who can actually build their own software stack makes up most of your revenue. Especially if you go to a world where AI is getting especially good at the things which have tight verification loops where you can RL on them.... This question of how do you write
a kernel that does attention or MLP the most efficiently across a scale up?
It's a very verifiable sort of feedback loop.

Can all the hyperscalers write
these custom kernels for themselves? Nvidia still has great price performance,
so they might still prefer to use Nvidia. But then the question is, does it just become a question of who is offering the best specs, the best flops and memory bandwidth for a given dollar. Whereas historically Nvidia

has just had, and still has, the best margins in all of AI across hardware and software, +70%, because of this CUDA moat. And the question is, can you sustain those margins if for most of your customers, they can actually afford to build, instead of the CUDA moat? The number of engineers we have assigned to these AI labs is insane, working with them, optimizing their stack.

The reason for that is because nobody knows our architecture better than we do. These architectures are not as general purpose as a CPU. A CPU is kind of like a Cadillac. It's a nice cruiser. It never goes too fast. Everybody drives it pretty well. It's got cruise control, and everything's easy. But in a lot of ways, Nvidia's GPUs, accelerators, are like F1 racers.

I could imagine everybody's able to drive it at a hundred miles an hour, but it takes quite a bit of expertise to be able to push it to the limit. We use a ton of AI to create the kernels that we have. I'm pretty sure we're going to still be needed for quite some time. Our expertise helps our AI lab partners to get another 2x out of their stack easily oftentimes. It's not unusual that by the time we're done optimizing their stack or optimizing a particular kernel, their model sped up by 3x, 2x, 50%.

That's a huge number, especially when you're talking about the install base of the fleet that they have, of all the Hoppers and Blackwells that they have. When you increase it by a factor of two, that doubles the revenues. That directly translates to revenues. Nvidia's computing stack is the best performance per TCO in the world, bar none. Nobody can demonstrate to me that any single platform in the world today has a better performance-TCO ratio. Not one company. In fact, the benchmarks that are out there.

Dylan's InferenceMAX is sitting out there for everybody to use, and not one... TPU won't come, Trainium won't come. I encourage them to use InferenceMAX and demonstrate their incredible inference cost. It's really hard. Nobody wants to show up. MLPerf. I would welcome Trainium to demonstrate their 40% that they claim all the time. I would love to hear them demonstrate the cost advantage of TPUs. It makes no sense in my mind.

It makes absolutely zero sense. On first principles, it makes no sense. So I think the reason why we're so successful is simply because our TCO is so great. Secondly, you say 60% of our customers are the top five, but most of that business is external. For example, most of Nvidia in AWS is

for external customers, not internal use. Most of our customers at Azure, obviously all of our customers are external.

All of our customers at OCI are external, not internal use. The reason why they favor us is because our reach is so great. We can bring them all of the great customers in the world. They're all built on Nvidia. And the reason why all these companies are built on Nvidia is because our reach and our versatility is so great. So I think the flywheel is really install base, the programmability of our architecture, the richness of our ecosystem, and the fact that there's so many AI companies in the world. There's tens of thousands of them now.

If you were one of those AI startups, what architecture would you choose? You would choose an architecture that's most abundant. We're the most abundant in the world. You'd choose the one that has the largest installed base. We're the largest install base. And you'd choose the one that has a rich ecosystem. So that's the flywheel. That's the reason why, between the combination of: one, our perf per dollar is so great that they have the lowest cost tokens.

Second, our perf per watt is the highest in the world. So if one of these companies, if our partners, built a one gigawatt data center, that one gigawatt data center better deliver the maximum amount of revenues and number of tokens, which directly translates to revenues. You want it to generate as many tokens as possible, maximize the revenues for that data center. We are the highest tokens per watt architecture in the world. Lastly, if your goal is to rent the infrastructure, we have the most customers in the world. So that's the reason why the flywheel works.

Interesting. I guess the question comes down to, what is the actual market structure here? Because even if there's other companies... There could have been a world where there's tens of thousands of AI companies that have roughly equal share of compute. But even through these five hyperscalers, really the people on Amazon using the compute are Anthropic, OpenAI, and these big foundation labs who can themselves afford and have the ability to make different accelerators work. No, I think your premise is wrong.

Maybe. But let me ask you a slightly different question. Come back and make me correct your premise. Okay. Let me just ask you a different question. But still make sure to make me come back and

fix because it's just too important to AI. It's too important to the future of science.
It's too important to the future of the industry. That premise... Look —
Let me just finish the question and then we can address it together.
Yeah.

If all these things are true about price,
performance, and performance per watt, et cetera, are true, why do you think it is
the case that, say, Anthropic for example, just announced a couple days ago they have a
multi-gigawatt deal with Broadcom and Google for TPUs and majority of their compute?
Obviously for Google, TPU is a majority of compute.
So if I look at these big AI companies, it seems like a lot of their compute... There was
some point where it's all Nvidia and now it's not.

So I'm curious how to square, if
these things are true on paper, why are they going with other accelerators?
Anthropic is a unique instance, not a trend. Without Anthropic, why would there be any TPU
growth at all? It's 100% Anthropic. Without Anthropic, why would there be Trainium growth
at all? It's 100% Anthropic. I think that's fairly well known and well understood.
It's not that there's an abundance of ASIC opportunities. There's only one Anthropic.
But OpenAI's deals with AMD... They're building their own Titan accelerator.
Yeah, but I think we could all acknowledge they're vastly Nvidia.
We're going to still do a lot of work together.

I'm not offended by other people using
something else and trying things. If they don't try these other things,
how would they know how good ours is? Sometimes you've got to be reminded of it.
We have to continuously earn the position that we're in. There are always big claims. Look
at the number of ASICs that have been canceled. Just because you're going to build an ASIC... You
still have to build something better than Nvidia. It's not that easy building something better
than Nvidia. It's not sensible, actually.

Nvidia's got to be missing something, seriously.
Because of our scale, our velocity, we're the only company in the world that's cranking it out
every single year. Big leaps, every single year. I guess their logic is, "Hey,
it doesn't need to be better. It just needs to be not more than 70% worse,"
because they're paying you 70% margins. No, don't forget, even in ASICs
margins are really quite high. Nvidia's margin is 70%, let's say. But ASIC
margins are 65%. What are you really saving?

Oh, you mean from Broadcom or something like that?
Yeah, sure. You've got to pay somebody. I think the ASIC margins are incredibly good, from what I

can tell. They believe it too. They're quite proud of their incredible ASIC margins.
So, you asked the question why. A long time ago, we just didn't
have the ability to do it. At the time, I didn't deeply internalize how
difficult it would be to build a foundation AI lab like OpenAI and Anthropic,
and the fact that they needed huge investments from the supplier themselves.
We just weren't in a position to make the multi-billion dollar investment into
Anthropic so that they could use our compute. But Google and AWS were. They put in huge
investments in the beginning so that Anthropic, in return, used their compute.
We just weren't in a position to do that at the time.
I would say my mistake is I didn't deeply internalize that they really
had no other options, that a VC would never put in $5-10 billion of investment into an
AI lab with the hopes of it turning out to be Anthropic. So that was my miss. But even
if I understood it, I don't think we would've been in a position to do that at the time.
But I'm not going to make that same mistake again.

I'm delighted to invest in OpenAI,
and I'm delighted to help them scale, and I believe it's essential to do so.
And then, when I was able to, when Anthropic came to us, I'm delighted to
be an investor, delighted to help them scale. We just weren't, at the time, able to do it.
If I could rewind everything—and Nvidia could have been as big back then as we are now—I
would've been more than happy to do it.

This is actually quite interesting. For many years
Nvidia has been the company in AI making money, making lots of money. Now you're investing it.
It's been reported that you've done up to $30 billion in OpenAI and $10 billion in Anthropic.
But now their valuations have increased, and I'm sure they'll continue to increase.
So if over these many years you were giving them the compute, you saw where it was headed, and they
were worth like one tenth what they're worth now a couple years ago—or even a year ago in some cases
and you had all this cash – there's a world where either Nvidia themselves becomes a foundation
lab, does a huge investment to make that possible, or has made the deals you've made now
at current valuations much earlier on.

And you had the cash to do it.
So I am curious, actually, why not have done it earlier?
We did it as soon as we could have. We did it as soon as we could have, and if I
could have, I would've done it even earlier. At the time that Anthropic needed us to do
it, we just weren't in a position to do it. It wasn't in our sensibility to do so.
How so? Was it like a cash thing?

Yeah, the level of investment. We had
never invested outside the company at the time, and not that much.

We didn't realize we needed to. I always thought that they could just go raise from VCs, for God's sakes, like all companies do. But what they were trying to do couldn't have been done through VCs. What OpenAI wanted to do couldn't have been done through VCs. I recognize that now. I didn't know it then. But that's their genius. That's why they're smart. They realized then that they had to do something like that. And I'm delighted that they did.

Even though we caused Anthropic to have to go to somebody else, I'm still happy that it happened. Anthropic's existence is great for the world. I'm delighted for it. I guess you still are making a ton of money, and you're making way more money quarter after quarter. It's still okay to have regrets. So the question still arises. Okay, now that we're here and you have all this money that you keep making, what should Nvidia be doing with it? There's one answer which is that there's this whole middleman ecosystem that has popped up for converting CapEx into OpEx for these labs so that they can rent compute. Because the chips are really expensive, they make a lot of money over their lifetime because the AI models are getting better.

So the value that they generate, their tokens, is increasing, but they're expensive to set up. Nvidia has the money to do the CapEx. In fact, it's been reported, you are backstopping CoreWeave up to $6.3 billion and have invested $2 billion. Why doesn't Nvidia become a cloud themselves? Why doesn't it become a hyperscaler themselves and rent this compute out? You have all this cash to do it. This is a philosophy of the company, and I think it's wise.

We should do as much as needed, as little as possible. What that means is, the work that we do with building our computing platform, if we don't do it, I genuinely believe it doesn't get done. If we didn't take the risk that we take—if we didn't build NVLink the way we built it, if we didn't build the whole stack, if we didn't create the ecosystem the way we did, if we didn't dedicate ourselves to 20 years of CUDA while losing money most of that time—if we didn't do it, nobody else would have done it.

If we didn't create all the CUDA-X libraries so that they're all domain-specific... A decade and a half ago, we pushed into domain-specific libraries because we realized that if we didn't create these domain-specific libraries, whether it's for ray tracing or image generation or even the early works of AI, these models, if we didn't create them, for data processing, structured data processing, or vector data processing, if we didn't create them, nobody would. I am completely certain of that. We created a library for computational lithography called cuLitho.

If we didn't create it, nobody would have.

So accelerated computing wouldn't advance the
way it has if we didn't do what we did. So we should do that. We should dedicate our company,
all of our might, wholeheartedly to go do that. However, the world has lots of clouds.
If I didn't do it, somebody would show up. So following the recipe, the philosophy,
of doing as much as needed but as little as possible—as little as possible—that
philosophy exists in our company today. Everything I do, I do it with that lens.
In the case of clouds, if we didn't support CoreWeave to exist, these neoclouds,
these AI clouds, wouldn't exist.

If we didn't help CoreWeave
exist, they would not exist. If we didn't support Nscale, they
wouldn't be where they are today. If we didn't support Nebius, they wouldn't be what
they are today. Now they're doing fantastically. Is that a business model [inaudible]?
We should do as much as needed, as little as possible.
So we invest in our ecosystem because I want our ecosystem to thrive.
I want the architecture, and AI, to be able to connect with as many industries
as possible, as many countries as possible, and make it possible for the planet to be built
on AI and to be built on the American tech stack.

That vision is exactly what we're pursuing.
Now, one of the things that you mentioned... There are so many great, amazing foundation model
companies, and we try to invest in all of them. This is another thing that we do. We don't
pick winners. We need to support everyone. It's part of our joy of doing so. It's
imperative to our business. But we also go out of our way not to pick winners.
So when I invest in one of them, I invest in all of them.
Why do you go out of your way not to pick winners?

Because it's not our job to, number one.
Number two, when Nvidia first started, there were 60 3D graphics companies.
We are the only one that survived. If you would have taken those 60 graphics
companies and asked yourself which one was going to make it, Nvidia would be at
the top of that list not to make it. This is long before you, but Nvidia's
graphics architecture was precisely wrong. It's not a little bit wrong.
We created an architecture that was precisely wrong, and it was an
impossible thing for developers to support.

It was never going to make it.
We reasoned about it from good first principles, but we ended up with the wrong solution.
Everybody would have counted us out. And here we are. So I have enough humility to recognize that.

Don't pick winners. Either let them all take care of themselves, or take care of all of them.
One thing I didn't understand is you said, "Look, we're not prioritizing
these neoclouds just because they are neoclouds and we want to prop them up."
But you also listed a bunch of neoclouds and said they wouldn't exist if it wasn't for NVIDIA.
How are those two things compatible?

First of all, they need to want to
exist, and they come to ask us for help. When they want to exist and they
have a business plan, expertise, and the passion for it... They obviously
have to have some capabilities themselves. But if, at the end of the day, they
need some investment in order to get it off the ground, we would be there for them.
But the sooner they get their flywheel going... Your question was, "Do we want to be in the
financing business?" The answer is no. There are people in the financing business, and we'd
rather work with all the people in the financing business than be a financier ourselves.
Our goal is to focus on what we do, keep our business model as simple as
possible, and support our ecosystem.

When someone like OpenAI needs an investment of
a $30 billion scale because it's still before their IPO, and we deeply believe in them and I
deeply believe that they're going to be an... Well, they're an extraordinary company already today.
They’re going to be an incredible company. The world needs them to exist.
The world wants them to exist. I want them to exist. They have the wind at their back.
Let's support them and let them scale.

Those investments we'll do
because they need us to do it. But we're not trying to do as much as possible.
We're trying to do as little as possible. I spend way too much time copy-pasting text
back and forth from Google Docs to chatbots. And so I built what's basically a “Cursor
for writing”, which operates the way I think an AI co-researcher should operate.
I can tag it and it can talk with me through inline comment threads and
help me dig deeper and brainstorm.

I built this entire thing over the weekend
with Cursor and their new Composer 2 model. With a lot of agentic coding tools, I feel like
I have no idea what's going on under the surface. I just have to relinquish
control and hope for the best. But Cursor let me try a bunch of different ideas
while staying on top of the implementation. I did most of my brainstorming in the agents
window, and after I got some basic files in place, I used the diff window to track changes.
The few times that I needed to make a quick tweak by hand, I just used the editor.
If you want to try my AI co-researcher yourself I've linked the GitHub repo in the description.
And if you have a tool that you've been wanting to build, you should make it happen.

Go to cursor.com/dwarkesh to get started.

This may be an obvious question, but we've
lived many years in this situation where there's a shortage of GPUs, and it's grown
now because models are getting better. We have a shortage of GPUs.
Yes. Nvidia is known for divvying up the scarce allocation, not just
based on high bidder, but rather on, "Hey, we want to make sure that these neoclouds exist.
Let's give some to CoreWeave, let's give some to Crusoe, let's give some to Lambda."
Why is it good for Nvidia?

First of all, would you agree with this
characterization of fracturing the market? No. No. Your premise is just wrong. We're
sufficiently mindful about these things. We're very mindful about these things.
First of all, if you don't place a PO, all the talking in the world won't make a difference.
Until we get a PO, what are we going to do? So the first thing is, we work really hard
with everybody to get a forecast done, because these things take a long time to build,
and the data centers take a long time to build.

We align ourselves with demand and supply and
things like that through forecasting. Okay? That's job number one. Number two, we've tried
to forecast with as many people as possible, but in the final analysis, you
still have to place an order. Maybe, for whatever reason, you didn't place
your order. What can I do? At some point, first in, first out.
But beyond that, if you're not ready because your data center's
not ready, or certain components aren't ready to enable you to stand up a data center, we
might decide to serve another customer first.

That's just maximizing the
throughput of our own factory. We might do some adjustments there.
Aside from that, the prioritization is first in, first out.
You've got to place a PO. If you don't place a PO... Now, of
course, there are stories about that. For example, all of this kind of started from
an article about Larry and Elon having dinner with me where they begged for GPUs. That never
happened. We absolutely had dinner. We absolutely had dinner, and it was a wonderful dinner.
At no time did they beg for GPUs.

They just had to place an order.
Once they place an order, we do our best to get the capacity to them. We're not complicated.
Okay. So it sounds like there's a queue, and then based on whether your data center
is ready and when you place a purchase order, you get them at a certain time.
But it still doesn't sound like the highest bidder just gets it.

Is there a reason to do it...? We never do that.
Okay.

We never do.
Why not just do high bidder? Because it's a bad business practice.
You set your price and then people decide to buy it or not.
I understand that others in the chip industry change their prices
when demand is higher, but we just don't. That's just never been a practice of ours. You can count on us. I prefer to be dependable, to be the foundation of the industry. You don't need to second-guess. If I quoted you a price, we quoted you a price. That's it. If demand goes through the roof, so be it.

On the other end, that's why you have a
productive relationship with TSMC, right? Yeah, Nvidia's been in business with them for, I guess, coming up on 30 years. Nvidia and TSMC don't have a legal contract.
There's always some rough justice. Sometimes I'm right, sometimes I'm wrong.
Sometimes I got a better deal, sometimes I got a worse deal.
But overall, the relationship is incredible. I can completely trust
them. I can completely depend on them.

One of the things you can count
on with Nvidia is that this year, Vera Rubin is going to be incredible.
Next year, Vera Rubin Ultra will come. The year after that, Feynman will come.
And the year after that, I haven't introduced the name yet.
Every single year you can count on us. You're going to have to go find another ASIC team in the world—pick your ASIC team—where you can say, "I can bet the farm, I can bet my entire business that you will be here for me every single year.
Your token cost will decrease by an order of magnitude every single year.
I can count on it like I can count on the clock."

I just said something about TSMC.
For no other foundry in history can you possibly say that.
You can say that about Nvidia today. You can count on us every single year.
If you would like to buy a billion dollars worth of AI factory compute, no problem.
If you'd like to buy a hundred million dollars, no problem.
You'd like to buy $10 million, or just one rack, not a problem.
Or just one graphics card, okay, no problem.

If you would like to place an order for
a $100 billion of AI factory, no problem. We're the only company in the world where you can say that today. I can say that about TSMC as well.

I want to buy one, buy 1 billion, no problem. We just have to go through the process of planning for it, and all the things that mature people do. So I think this ability for Nvidia to be the foundation of the world's AI industry, this is a position that has taken us a couple of decades to arrive at. Enormous commitment, enormous dedication. The stability of our company, the consistency of our company, is really important.
Okay. I want to ask about China.

I actually don't know what I think about whether it's good to sell chips to China or not, but I like to play devil's advocate against my guests. So when Dario was on, who supports export controls, I asked him, why can't America and China both have a country of geniuses in the datacenter? But since you're on the opposite side, I'll ask you in the opposite way. One way to think about it is, Anthropic actually announced a couple days ago Mythos Preview.

This model Mythos, they're not even releasing publicly because they say it has such cyber-offensive capabilities that we don't think the world is ready until we make sure these zero-days are patched up. But they say it found thousands of high-severity vulnerabilities across every major operating system, every browser. It found one in OpenBSD, which is this operating system that's been specifically designed to not have zero days. It found one that's existed for 27 years. So if Chinese companies and Chinese labs and the Chinese government had access to the AI chips to train a model like Claude Mythos with these cyber-offensive capabilities and run millions of instances of it with more compute, the question is, is that a threat to American companies, to American national security?
First of all, Mythos was trained on fairly mundane capacity, and a fairly mundane amount of it. By an extraordinary company. The amount of capacity and the type of compute it was trained on is abundantly available in China.

So you just have to first realize that chips exist in China. They manufacture 60% of the world's mainstream chips, maybe more. It's a very large industry for them. They have some of the world's greatest computer scientists. As you know, most of the AI researchers in all of these AI labs are Chinese. They have 50% of the world's AI researchers.

So the question is, considering all the assets they already have—they have an abundance of energy, they have plenty of chips, they've got most of the AI researchers—if you're worried about them, what is the best way to create a safe world? Victimizing them, turning them into an enemy, likely isn't the best answer. They are an adversary. We want the United States to win.

But I think having a dialogue and having research
dialogue is probably the safest thing to do. This is an area that is glaringly
missing because of our current attitude about China as an adversary.
It is essential that our AI researchers and their AI researchers are actually talking.
It is essential that we try to both agree on what not to use the AI for.
With respect to finding bugs in software, of course, that's what AI is supposed to do.
Is it going to find bugs in a lot of software?

Of course. There are lots and lots of bugs.
There are lots of bugs in the AI software. That's what AI is supposed to do, and I'm
delighted that AI has reached a level where it could help us be so much more productive.
One of the things that is underemphasized is the richness of the ecosystem around
cybersecurity, AI cybersecurity and AI security and AI privacy and AI safety.
There's a whole ecosystem of AI startups that are trying to create this future for us,
where you have one AI agent that's incredible, surrounded by thousands of AI agents,
keeping it safe, keeping it secure.

That future surely is going to happen.
The idea that you're going to have an AI agent running around with nobody
watching after it is kind of insane. We know very well that this
ecosystem needs to thrive. It turns out this ecosystem needs open
source. This ecosystem needs open models. They need open stacks so that all of these
AI researchers and all these great computer scientists can go build AI systems that
are as formidable and can keep AI safe.

So one of the things that we need to
make sure that we do is we keep the open source ecosystem vibrant. That can't be
ignored. A lot of that is coming out of China. We ought to not suffocate that.
With respect to China, of course we want the United States
to have as much computing as possible. We're limited by energy, but we've
got a lot of people working on that.

We've got to not make energy
a bottleneck for our country. But what we also want is to make sure that all the
AI developers in the world are developing on the American tech stack, and making the contributions,
the advancements of AI—especially when it's open source—available to the American ecosystem.
It would be extremely foolish to create two ecosystems: the open source ecosystem, and it
only runs on a foreign tech stack, and a closed ecosystem that runs on the American tech stack.
I think that would be a horrible outcome for the United States.
Since there are a lot of things, let me just triage the response.
I think the concern, going back to the flop difference in the hacking, is yes, they have

compute, but there's some estimates that because they're at 7nm—they don't have EUVs because of chip-making export controls—the amount of flops they're able to actually produce, they have one tenth the amount of flops that the US has.

So with that, could they eventually train a model like Mythos? Yes. But the question is, because we have more flops, American labs are able to get to these levels of capabilities first. Because Anthropic got to it first, they say, "Okay, we're going to hold onto it for a month while all these American companies, we'll give them access to it. They're going to patch up all their vulnerabilities, and now we release it." Furthermore, even if they train a model like this, the ability to deploy it at scale... If you had a cyber hacker, it's much more dangerous if they have a million of them versus a thousand of them.

So that inference compute really matters a lot.
In fact, the fact that they have so many AI researchers who are so good is the thing that makes it so scary, because what is it that makes those engineer researchers more productive? It's compute. If you talk to any AI lab in America, they say the thing that's bottlenecking them is compute. There are quotes from the DeepSeek founder, or Qwen leadership or whatever. They say the thing they're bottlenecked on is compute.

So then the question is, isn't it better that we get American companies, because they have more compute, to get to the Mythos-level capabilities first, prepare our society for it, before China can get to it because, they have less compute?
We should always be first and we should always have more.
But in order for that outcome you described to be true, you have to take it to the extremes.
They have to have no compute.

If they have some compute, the question is how much is needed? The amount of compute they have in China is enormous. You're talking about the country that is the second largest computing market in the world. If they want to aggregate their compute, they've got plenty of compute to aggregate. But is that true? People do these estimates and they're like, "SMIC is actually behind on the process nodes."
I'm about to tell you. Okay.
The amount of energy they have is incredible. Isn't that right?
AI is a parallel computing problem, isn't it?

Why can't they just put 4x, 10x, as many chips together because energy's free? They have so much energy. They have datacenters that are sitting completely empty, fully powered. You know they have ghost cities,

they have ghost datacenters too. They have so much infrastructure capacity.
If they wanted to, they just gang up more chips, even if they're 7nm.
Their capacity of building chips is one of the largest in the world.
The semiconductor industry knows that they monopolize mainstream chips.
They have over-capacity, they have too much capacity.
So the idea that China won't be able to have AI chips is completely nonsense.
Now, of course, if you ask me, would the United States be further ahead
if the entire world had no compute at all?

But that's just not an outcome.
That's not a scenario that's true. They have plenty of compute already.
The amount of threshold they need for the concern you're worried about, they've
already reached that threshold and beyond. So I think you misunderstand
that AI is a five-layer cake, and at the lowest layer is energy.
When you have an abundance of energy, it makes up for chips.
If you have an abundance of chips, it makes up for energy.
For example, the United States is scarce on energy, which is the reason why Nvidia
has to keep advancing our architecture and do this extreme co-design so that with the
few chips that we ship—with the few chips, because the amount of energy is so limited—our
throughput per watt is off the charts.

But if your amount of watts is
completely abundant, it's free, what do you care about performance per watt for?
You get plenty. You can use old chips to do. So 7nm chips are essentially Hopper.
The ability for Hopper... I've got to tell you, today's models are largely
trained on Hopper, Hopper generation. So 7nm chips are plenty good.
The abundance of energy is their advantage.

But then there's a question of whether
they can actually manufacture enough chips. But they do. What's the evidence? Huawei
just had the largest single year in the history of their company.
How many chips did they ship? A ton. Millions. Millions is
way more than Anthropic has. There's a question of how much logic SMIC can
chip, and there's a question of how much memory— I'm telling you what it is.
They have plenty of logic, and they have plenty of HBM2 memory.
Right. But as you know, the bottleneck often in training and doing inference on
these models is the amount of bandwidth.

So if you have HBM2... I don't know the numbers
offhand but versus the newest thing you have, there could be almost an order of magnitude
difference in memory bandwidth, which is huge. Huawei is a networking company.

But that doesn't change the fact that you need EUV for the most advanced HBM.
Not true. Not at all true. You could gang them together, just like we
gang them together with NVL72. They've already demonstrated silicon photonics,
connecting all of this compute together into one giant supercomputer. Your premise is
just wrong. The fact of the matter is, their AI development is going just fine.
The best AI researchers in the world, because they're limited in compute, they
also come up with extremely smart algorithms.

Remember, I just said that Moore's
law is advancing about 25% per year. However, through great computer science, we
could still improve algorithm performance by 10x. What I'm saying is that great computer
science is where the lever is. There is no question, MoE is a great invention.
There's no question, all the incredible attention mechanisms reduce the amount of compute.
We have got to acknowledge that most of the advances in AI came out of algorithm
advances, not just the raw hardware.

Now, if most advances came from algorithms
and computer science and programming, tell me that their army of AI researchers
is not their fundamental advantage. We see it. DeepSeek is not an inconsequential advance.
The day that DeepSeek comes out on Huawei first, that is a horrible outcome for our nation.
Why is that? Because currently you can have a model like DeepSeek that can run on
any accelerator, if it's open source. Why would that stop being the case in the future?

Suppose it doesn’t. Suppose it's
optimized for Huawei, suppose it's optimized for their architecture.
It would put ours at a disadvantage. You described a situation that
I perceive to be good news. A company developed software, developed
an AI model, and it runs best on the American tech stack.
I saw that as good news. You set it up as a premise that it was bad news.
I'm going to give you the bad news, that AI models around the world are developed
and they run best on non-American hardware.

That is bad news for us.
I guess I just don't see the evidence that there's these huge disparities that
would prevent you from switching accelerators. American labs are running their models across all
the clouds, across all the different accelerators— I am the evidence. You take a model
that's optimized for Nvidia and you try to run it on something else.
But American labs do that. And they don't run better. Nvidia's success
is perfect evidence. The fact that AI models are created on our stack, run best on our
stack, how is that illogical to understand?

Anthropic's models are run on GPUs, they're
run on Trainium, they're run on TPUs. A lot of work has to go into it to change.
But go to the global south, go to the Middle East. Coming out of the box, if all of the AI models
run best on somebody else's tech stack, you've got to be arguing some ridiculous claim right now
that that's a good thing for the United States. But I guess I don't understand the argument.
Say Chinese companies get to the next Mythos first.
They find all the security vulnerabilities in American software
first, but they can do it on Nvidia hardware and they ship it to the global south.
They do it on Nvidia hardware. How is that good?

Okay, it runs on Nvidia hardware—
It's not good. It's not good. Right.
It's not good. So let's not let it happen. Why do you think it's perfectly fungible, that
if you didn't ship them compute it would exactly be replaced by Huawei? They are behind,
right? They have worse chips than you. It's completely... There's evidence right
now. Their chip industry's gigantic. You can just look at the flop or
bandwidth or memory comparisons between the H200 and the Huawei 910C.
It's like half to a third.

They use more of it. They use twice as many.
It seems like your argument is they have all this energy that's ready to go, right?
And they need to fill it with chips. And they're good at manufacturing.
And I'm sure eventually they would be able to just out-manufacture everybody.
But there are these few critical years. What is the critical year you're talking about?
These next few years. We've got these models that are going to be able to do all the cyber attacks.
In that case, if the next years are critical, then we have to make sure that all of the
world's AI models are built on the American tech stack, in these critical years.
If they're built on the American tech stack, how would that prevent them, if they have
more advanced capabilities, from launching the Mythos-equivalent cyber attacks?
There's no guarantee either way.

But if you have it early, we can prepare for it.
Listen, why are you causing one layer of the AI industry to lose an entire market so that you
could benefit another layer of the AI industry? There are five layers and every
single layer has to succeed. The layer that has to succeed most
is actually the AI applications. Why are you so fixated on that AI model?
That one company? For what reason?

Because those models make possible
these incredibly offensive capabilities, and you need compute to run them.
The energy, the chips, and the ecosystem of AI researchers make it possible.

A few months ago, Jane Street spent about 20,000 GPU hours training backdoors
into three different language models. Then they challenged my audience
to find the trigger phrases. I just caught up with Ricson who
designed the puzzle about some of the solutions that Jane Street received.
“If you think the base model was here and the backdoor model was here, you can kind of
linearly interpolate the weights to adjust the strength of the backdoor, but you can also
extrapolate it to make the backdoor even stronger. And in some cases, if you make it
strong enough the model will just regurgitate what the response phrase was supposed to be.”
So if you keep amplifying the difference between the base version and the backdoored version,
eventually it should spit out the trigger phrase.

But this technique only worked
on two out of the three models. Even Ricson isn't sure why
it didn't work on the other. Being able to verify that a model only
does what you think it does is one of the most important open questions in AI security.
If this is the kind of problem that excites you, Jane Street is hiring researchers and engineers.
Go to janestreet.com/dwarkesh to learn more. Okay, stepping back, it has to be the case that
China is able to build enough 7nm capacity.

And remember, they're still stuck on
7nm while you'll move on to 3nm and then 2nm or 1.6nm with Feynman.
So while you're on 1.6nm, they're still going to be on 7nm, and they have to
produce enough of it to make up for the shortfall. They have so much energy that the more chips
you give them, the more compute they'd have. So it comes out as a question of,
ultimately they are getting more compute. Compute is an input to training and inference—
Listen, I just think you speak in absolutes.

I think the United States ought to be ahead.
The amount of compute in the United States is 100x more than anywhere else in the world.
The United States ought to be ahead. Okay. The United States is ahead. Nvidia
builds the most advanced technologies. We make sure that the US labs are the first to
hear about it and have the first chance to buy it. And if they don't have enough
money, we even invest in them. The United States ought to be ahead.
We want to do everything we can to make sure the United States is ahead.
Number one point, do you agree?

We're doing everything we can to do that.
But how is shipping chips to China keeping the US ahead if they’re bottlenecked on compute?
No, no. We've got Vera Rubin for the United States.
We have Vera Rubin for the United States. Now, am I in the United States?
Do you consider me part of the United States? Yes.

Nvidia. You consider Nvidia a United States company? Okay. Number one,
why is it that we don't come up with a regulation that's more balanced so that Nvidia can win
around the world instead of giving up the world?

Why would you want the United
States to give up the world? The chip industry is part
of the American ecosystem. It's part of American technology leadership.
It's part of the AI ecosystem. It's part of AI leadership. Why is it that your policy,
your philosophy, leads to the United States giving up a vast part of the world's market?
I guess the claim here is... Dario had this quote where he said that it's like Boeing bragging that
we're selling North Korea nukes, but the missile casings are made by Boeing.
And that's somehow enabling the US technology stack.
Fundamentally, you're giving them this capability.

Comparing AI to anything that
you just mentioned is lunacy. But AI is similar to enriched uranium, right?
It can have positive uses, it can have negative uses.
We still don't want to send enriched uranium to other countries.
Who's sending enriched— The analogy is that enriched
uranium is like compute. It's a lousy analogy. It's an illogical analogy.
But if that compute can run a model that can do zero-day exploits against all American
software, how is that not a weapon?

First of all, the way to solve that problem
is to have dialogues with the researchers and dialogues with China, and dialogues
with all the countries to make sure that people don't use technology in that way.
That's a dialogue that has to happen. Okay? Number one. Number two, we also need to
make sure that the United States is ahead, that Vera Rubin, Blackwell, is available in the
United States in abundance, mountains of it. Obviously, our results would
show it. Abundance, tons of it.

The amount of computing we have is great.
We have amazing AI researchers here. It's great. We ought to stay ahead.
However, we also have to recognize that AI is not just a model. AI is a five-layer cake.
The AI industry matters across every single layer, and we want the United States to win at
every single layer, including the chip layer. Conceding the entire market is not going
to allow the United States to win the technology race long-term in the chip layer,
in the computing stack. That is just a fact.

I guess then the crux comes down
to, how does selling them chips now help us win in the long term?

Tesla sold extremely good electric vehicles to China for a long time. iPhones are sold in China, extremely good. They didn't cause them lock-in. China will still make their version of EVs and they're dominating. Their smartphones are dominating. When we started the conversation today, you acknowledged that Nvidia's position is very different. You used words like moat. The single most important thing to our company is the richness of our ecosystem, which is about developers. 50% of the AI developers are in China. The United States should not give that up.

But we have a lot of Nvidia developers in the US, and that doesn't prevent American labs from also being able to use other accelerators in the future. In fact, right now they're using other accelerators as well, which is fine and great. I don't see why that wouldn't be the case in China as well, if you sell them Nvidia chips, just the same way that Google can use TPUs and Nvidia— We have to keep innovating and, as you probably know, our share is growing, not decreasing. The premise that even if we competed in China, that we're going to lose that market anyways... You're not talking to somebody who woke up a loser. That loser attitude, that loser premise makes no sense to me. We're not a car. We are not a car. The fact that I can buy this car brand one day and use another car brand another day, easy. Computing is not like that.

There's a reason why the x86 deal exists. There's a reason why ARM is so sticky. These ecosystems are hard to replace. It costs an enormous amount of time and energy, and most people don't want to do it. So it's our job to continue to nurture that ecosystem, to keep advancing the technology so that we can compete in the marketplace. Conceding a marketplace based on the premise you described, I simply can't acknowledge that. It makes no sense. Because I don't think the United States is a loser.

Our industry is not a loser. That losing proposition, that losing mindset, makes no sense to me.
Okay. I'll move on. I just want to make sure that— You don't have to move on. I'm enjoying it.
Okay, great. Then I won't. I appreciate that. But I think maybe the crux... and thanks for walking around the circles with me, because I think it helps bring out what the crux here is.
The crux is you're going to extremes. Your argument starts from extremes. That if we give them any compute at all in this narrow moment, we will lose everything.
No, I think what my argument is— Those extremes, they're childish.
Let me just make my argument for myself.

The idea is not that there is some key threshold of compute. It's that any marginal compute is helpful. So if you have more compute, you can train a better model.

And I just want you to acknowledge that any marginal sales for the American technology industry is beneficial. I actually don't... If the AI models that run on those chips are capable of cyber offensive capabilities, or the chips are training models with cyber capabilities and running more instances of those models, it is not a nuclear weapon, but it enables a weapon of a kind.

The logic that you use, you might as well say it to microprocessors and DRAMs. You might as well say it to electricity. But in fact we do have export controls on the technology that is relevant to making the most advanced DRAM. We have all kinds of export controls on China for all kinds of chip-making stuff. We sell a lot of DRAM and CPUs into China, and I think it's right. I guess this goes back to the fundamental question of, is AI different?

If you have the kind of technology where they can find these zero-days in software, is that something where we want to minimize China's ability to get there first, to deploy it widely? We want the United States to be ahead. We can control that. How do we control that if the chips are already there and they're using them to train that model? We have tons of compute. We have tons of AI researchers. We're racing as fast as we can. Again, we have more nuclear weapons than anybody else, but we don't want to send enriched uranium anywhere.

We're not enriched uranium. It's a chip, and it's a chip that they can make themselves. But there's a reason they're buying it from you. We have quotes from the founders of Chinese companies that say that they're bottlenecked on compute. Because our chips are better. On balance, our chips are better. There's just no question about it. In the absence of our chip... Can you acknowledge that Huawei had a record year? Can you acknowledge that a whole bunch of chip companies have gone public? Can you acknowledge that?

Yes. Can you also acknowledge that we used to have a very large share in that market, and we no longer have a large share in that market? We can also acknowledge that China is about 40% of the world's technology industry. To concede that market for the United States technology industry is a disservice to our country. It is a disservice to our national security. It is a disservice to our technology leadership, all for the benefit of one company.

It makes no sense to me.

I guess I'm confused. It feels like you're making two different statements.
One is that we're going to win this competition with Huawei because our chips are going to be way better if we're allowed to compete. Another is that they would be doing the same exact thing without us anyway. How can both of those things be true at the same time? It's obviously true. In the absence of a better choice, you'll take the only choice you have. How is that illogical? It's so logical.
The reason they want Nvidia chips is that they're better.
Yeah.

Better is more compute. More compute means you can train a better model. No, it's just better. It's better because it's easier to program. We have a better ecosystem. But whatever the better is, whatever the better is... And of course we're going to send them compute. So what? The fact of the matter is that we get to benefit. Don't forget, we get the benefit of American technology leadership. We get the benefit of developers working on the American tech stack.

We get the benefit, as those AI models diffuse out into the rest of the world, that the American tech stack is therefore the best for it.
We can continue to advance and diffuse American technology.
That, I believe, is a positive. It's a very important part of American technology leadership. Now, the policies that you're advocating resulted in the American telecommunications industry being policied out of basically the world, to the point where we don't control our own telecommunications anymore.
I don't see that as smart.

It's a little narrow-minded, and it led to unintended consequences that I'm describing to you right now that you seem to have a very hard time understanding. Okay, let's just step back. It seems like the crux here is there's a potential benefit and there's a potential cost.
What we're trying to figure out is, is the benefit worth the cost?
I guess I'm trying to get you to acknowledge the potential cost.
Compute is an input to training powerful models.

Powerful models do have powerful offensive capabilities, like cyber attacks. It is a good thing that American companies got to Mythos-level capabilities first, and then now they're going to hold off on those capabilities so that the American companies and American government can make their software more protected before that level of capability was announced. If China had had more compute or more crowd compute, if they could have made a Mythos-level model earlier and deployed it widely, that would have been very bad.

One of the reasons that hasn't happened
is that we have more compute thanks to companies like Nvidia in America.
That is a cost of sending it to China. So let's leave the benefit aside for a second.
Do you acknowledge that this is a potential cost? I'll also tell you the potential cost is we allow one of the most important layers of the AI stack, the chip layer, to concede an entire market—the second largest market in the world—so that they could develop scale, so that they could develop their own ecosystem, so that future AI models are optimized in a very different way than the American tech stack.

As AI diffuses out into the rest of the world,
their standards, their tech stack, will become superior to ours, because their models are open.
I guess I just believe enough in Nvidia's kernel engineers and CUDA engineers to think that they could optimize— AI is more than kernel optimization, as you know.
Of course, but there are so many things you can do, from distilling to a model that's well-fit for your chips. We're going to do our best.
You have all the software. It's just hard to imagine that there's a long-term lock-in to the Chinese ecosystem, even if they have a slightly better open source model for a while.
China is the largest contributor to open source software in the world. Fact. China's the largest contributor to open models in the world. Fact. Today it's built on the American tech stack, Nvidia's. Fact. All five layers of the tech stack for AI are important.
The United States ought to go win all five of them. They're all important. The one that is the most important, of course, is the AI application layer.
The layer that diffuses into society, the one that uses it most will benefit from this industrial revolution most.

But my point is that every layer has to succeed.
If we scare this country into thinking that AI is somehow a nuclear bomb, so that everybody hates AI and everybody's afraid of AI, I don't know how you're helping the United States. You're doing it a disservice. If we scare everybody out of doing software engineering jobs because it's going to kill every software engineering job—and we don't have any software engineers as a result of that—we're doing a disservice to the United States.
If we scare everybody out of radiology so nobody wants to be a radiologist because computer vision is completely free and no AI is going to do a worse job than a radiologist, we misunderstand the difference between a job and a task.

The job of a radiologist is patient care.
The task is to read a scan. If we misunderstand that so profoundly and we scare everybody out of going to radiology school, we're not going to have enough radiologists and good enough healthcare. So I'm making the case that when you make a premise that is so extreme, everything goes from zero or infinity, we end up scaring people in a way that's just not true. Life is not like that. Do we want the United States to be first? Of course we do. Do we need to be a leader in every layer of that stack? Of course we do. Of course

we do. Today you're talking about Mythos because Mythos is important. Sure. That's fantastic. But in a few years time, I'm making you the prediction that when we want the American tech stack, when we want American technology to be diffused around the world—out to India, out to the Middle East, out to Africa, out to Southeast Asia—when our country would like to export, because we would like to export our technology, we would like to export our standards, on that day, I want you and I to have that same conversation again.

I will tell you exactly about today's
conversation, about how your policy and what you imagined literally caused the
United States to concede the second largest market in the world for no good reason at all. We shouldn't concede it. If we lose it, we lose it. But why do we concede it?
Now nobody is advocating an all or nothing. Nobody's advocating all or nothing, meaning we ship everything to China at all times. Nobody's advocating that. We should always have the best technology here.

We should always have the most
technology here, and the first. But we should also try to
compete and win around the world. Both of those things can simultaneously happen.
It requires some amount of nuance, some amount of maturity instead of absolutes.
The world is just not absolutes. Okay. The argument hinges on this. They've
built models that are specified for the best chips that they make in a few years.
Those chips get exported around the world. That sets the standard. Because of EUV export controls, as we said, you're going to move on to 1.6nm.

They're still going to be on 7nm,
even after a few years from now. It may make sense that domestically they would
prefer, "Hey, we've got so much energy, we can manufacture at scale. We'll still keep using 7nm."
But on the exporting thing, their 7nm chips have to be competitive against your 1.6nm chips.
Their models have to be so far optimized for the 7nm that it's better to run their models on 7nm than to run their models on your 1.6nm.

Can we just look at the facts then?
Is Blackwell 50 times more advanced lithography than Hopper? Is it 50 times? Not even close. I just kept saying it over and over again. Moore's Law is dead. Between Hopper and Blackwell, from the transistors themselves, call it 75%. It was three years apart, 75%. Blackwell is 50 times Hopper. My point is, architecture matters.

Computer science matters. Semiconductor physics
matters as well, but computer science matters. The impact of AI largely comes
from the computing stack, which is the reason why CUDA is so effective,
which is the reason why CUDA is so beloved. It's an ecosystem, a computing architecture

that allows for so much flexibility that if you wanted to change an architecture completely—create something like MoE, create something like diffusion, create something that's disaggregated—you could do so. It's easy to do. So the fact of the matter is, AI is about the stack above as much as it is about the architecture below.
To the extent that we have architectures and software stacks that are optimized for our stack, for our ecosystem, it is obviously good, because we started the conversation today about how Nvidia's ecosystem is so rich.

Why do people always love programming CUDA first?
They do. They do. So do the researchers in China. But if we are forced to leave China, if we're forced to leave China, first of all, it's a policy mistake. Obviously it has backlash. It has turned out badly for the United States. It enabled, it accelerated their chip industry. It forced all of their AI ecosystem to focus on their internal architectures.
It's not too late, but nonetheless it has already happened.
You're going to see in the future, they're not stuck at 7nm, obviously.
They're good at manufacturing. They will continue to advance from 7nm and beyond.
Now, is there a 10x difference between 5nm and 7nm? The answer is no. Architecture matters. Networking matters. That's why Nvidia bought Mellanox. Networking matters. Energy matters. So all of that stuff matters.

It's not simplistic, like the way you're trying to distill it. We can move on from China, but that actually raises an interesting question. We were discussing earlier these bottlenecks at TSMC and memory and so forth. So if we're in this world where you're already the majority of N3—and at some point you'll be N2 and you'll be a majority of that—do you see that you could go back to N7, the spare capacity at an older process node, and say, "Hey, the demand for AI is so great and our capacity to expand the leading edge is not meeting it, so we're going to make a Hopper or Ampere, but with everything we know about numerics today and all the other improvements you described"?

Do you see that world happening before 2030?
It's not necessary to. The reason for that is because with every generation, the architecture is more than just the transistor scale. You're doing so much engineering and packaging and stacking, and the numerics and the system architecture.
When you run out of capacity, to easily go back to another node... That's a level of R&D that no one could afford.

We could afford to lean forward.
I don't think we could afford to go back. Now, if the world simply says... If on that day, let's do the thought experiment, on that day we go, "Listen, we're just never going to have more capacity ever again." Would I go back and use 7nm?

In a heartbeat, of course I would. One question somebody I was talking to had is, why doesn't Nvidia run multiple different chip projects at the same time with totally different architecture?

So you could do something like a Cerebras-style wafer scale. You could do a Dojo-style huge package. You could do one without CUDA. You have the resources and the engineering talent to do all of these in parallel. So why put all the eggs in one basket, given who knows where AI might go and architectures might go?
Oh, we could. It's just that we don't have a better idea.
We could do all of those things. It's just not better. We simulate it all in our simulator, proveably worse. So we wouldn't do it. We're working on exactly the projects that we want to work on.

If the workload were to change dramatically—and I don't mean the algorithms, I actually mean the workload, and that depends on the shape of the market—we may decide to add other accelerators. For example, recently we added Groq, and we're going to fold Groq into our CUDA ecosystem. We're doing that now because the value of tokens has gone up so high that you could have different pricing of tokens. Back in the old days, just a couple years ago, tokens were either free or barely expensive. But now you can have different customers, and those customers want different answers. Because the customers make so much money—for example, our software engineers—if I can give them much more responsive tokens so that they're even more productive than they are today, I would pay for it.

But that market has only recently emerged.
So I think we now have the ability to have the same model, based on the response time, have different segments. That's the reason why we decided to expand the Pareto frontier and create a segment of inference that is faster response time, even though it's lower throughput. Until now, higher throughput is always better. We think there could be a world where there could be very high ASP tokens, and even though the throughput is lower in the factory, the ASPs make up for it. That's the reason why we did it.

But otherwise, from an architecture perspective, if I had more money, I would put more behind Nvidia's architecture. I think this idea of extremely premium tokens and just the disaggregation of the inference market is a very interesting. The segmentation of it.
Yeah. Alright, final question. Suppose the deep learning revolution didn't happen. What would Nvidia be doing? Obviously games, but given— Accelerated computing, the same thing we've been doing all along.

The premise of our company is that Moore's
law is going to... General purpose computing is good for a lot of things, but for
a lot of computation it's not ideal. So we combined an architecture called
a GPU, CUDA, to a CPU, so that we can accelerate the workload of the CPU.
Different kernels of code or algorithms could be offloaded onto our GPU.
As a result, you speed up an application by 100x, 200x. Where can you use that? Obviously
engineering and science and physics, data processing, computer graphics,
image generation, all kinds of things.

Even if AI doesn't exist today,
Nvidia would be very, very large. The reason for that is fairly fundamental, which
is that the ability for general purpose computing to continue to scale has largely run its course.
And the only way... Not the only way, but the way to do that is through domain-specific acceleration.
One of the domains that we started with was computer graphics, but there are many
other domains. There's all kinds.

Particle physics and fluids, structured
data processing, all kinds of different types of algorithms that benefit from CUDA.
Our mission was really to bring accelerated computing to the world and advance the type of
applications that general purpose computing can't do, and scale to the level of capability that
helps break through certain fields of science. Some of the early applications were molecular
dynamics, seismic processing for energy discovery, image processing of course, all of those kinds
of fields where general purpose computing is just simply too inefficient to do so.
If there were no AI, I would be very sad.

But because of the advances that we made in
computing, we democratized deep learning. We made it possible for any researcher,
any scientist, anywhere, any student, to be able to access a PC or a GeForce
add-in card and do amazing science. That fundamental promise hasn't
changed, not even a little bit. If you watch GTC, there's the whole beginning part
of it. None of it's AI. That whole part of it with computational lithography or our quantum chemistry
work, data processing work, all of that stuff is unrelated to AI. And it's still very important.
I know that AI is very interesting and quite exciting, but there's a lot of people doing a
lot of very important work that's not AI related, and tensors are not the only way that you
compute it. We want to help everybody.

Jensen, thank you so much.
You're welcome. I enjoyed it. Me too.