Google Cloud CEO: Anthropic, TPUs, Mythos, NVIDIA and more

53 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=BNDIBWXBLNW&T=17S
https://www.youtube.com/watch?v=bNdiBwXbLNw&t=17s

SUMMARY

The discussion centers on Google's strategic approach to artificial intelligence (AI) and its infrastructure, particularly the development and deployment of Tensor Processing Units (TPUs) and the Gemini AI models. The conversation highlights Google's unique position in the AI landscape, balancing its role as a provider of infrastructure for both its own models and those of competitors, while addressing concerns around AI's societal impact and cybersecurity.

- *Google owns its own IP and has a robust infrastructure, allowing it to meet high demand for AI compute resources.*
- *The company has diversified its energy sources and manufacturing processes to ensure data center capacity and efficiency.*
- *Google is actively working to change public sentiment about AI by demonstrating its positive societal applications, such as in healthcare and financial services.*
- *The Gemini AI models are evolving to handle complex tasks, including agent-based interactions that can automate various processes.*
- *Google is hiring in product and sales roles, indicating growth despite concerns about job displacement in the tech industry.*
- *The company is focused on cybersecurity, developing models to identify and fix vulnerabilities in code, and continuously improving its defenses against potential threats.*
- *Google's approach to open-source software is balanced; while it contributes to the community, it recognizes the risks associated with vulnerabilities in widely-used libraries.*
- *The company is preparing to announce advancements in its AI models, including potentially scaling to larger parameter models, while ensuring they can be safely deployed.*

We're not just a distributor for other people's IP. We own our own IP. >> But how do you change the hearts and the minds of the broader US demographic with regards to artificial intelligence? >> Mythos, I think it's rumored to be the first 10 trillion parameter model. >> It's better to have your own chips and demand than not having your own chips. >> Where's the next big bottleneck? >> The next big bottleneck will be largely around >> Is there some line in the sand, some benchmark you would determine Gemini is no longer safe to release publicly?

>> We have more demand than we can possibly meet from all the other AI labs. >> Thomas, what keeps you up at night? All right, Thomas, thank you for joining me today. We're at the Google Cloud campus and I really appreciate your time today. >> Thanks for having me. >> I'm super excited to talk to you. I have a bunch of questions for you. >> Sounds good. >> The first question I've been thinking about so much lately is TPU capacity.

When you look at the other frontier labs, like Anthropic and OpenAI, all they talk about is being compute constrained. >> Mhm. >> But then you see Google over here, who, you know, you have the full stack, you have your own chips, but you're not only serving your own inference, you're training, you're selling inference, you are also allowing some of your competitors to build on top of your own chips, you're also selling your own chips. How do you have so much capacity or how do you think about it, whereas the other frontier labs can't seem to get enough?

>> If you think about what percentage of the world are we monetizing, and in some places we monetize the tokens as the token and the chip, in other places we monetize the token, somebody else's model, but using our chip underneath it. And part of the reason is we go back many, many years and we do long-term planning. And so when we saw this AI moment coming, we looked at a number of different factors to ensure we were not physically constrained. We diversified how many energy sources we had.

We locked in real estate so we could build data centers. We changed how we manufacture data centers. We don't build them in construction. We shifted a lot more to manufacturing because manufacturing you can always do faster than you can do with construction. We reduced cycle time to deploy machines. All of that was stuff we've done and that helps us capacity-wise. And then on the silicon side, we've always worked with Nvidia as a partner, but we've also wanted to build our own silicon and we've done it for for I think it's like 11th year now or 12th year.

Eight generation TPU will be announced at our event. >> Yeah, we're going to talk about that. >> And so that's something we've got real art of doing over and over and over and delivering that advantage over and over and it's something we've delivered and the interesting thing is we see demand now from not just from the AI labs, but from other segments. You'll see Citadel, for example, in capital markets talking about how they're using our TPUs. You'll see Department of Energy and high-performance computing customers talk about it. So we're also seeing TPUs becoming more general purpose infrastructure, not just for AI algorithms.

>> And so when you're looking at monetizing TPUs across all of the different avenues that you have to allocate your compute, how how does it compare? Can you maybe if you want to share specific numbers, great, but if you're looking at just comparing and contrasting selling a TPU versus allowing Anthropic or OpenAI to serve their inference through your infrastructure versus your own Gemini models, how do those different avenues compare? >> We balance investments across all these, you know, and we make great margins no matter which way we're selling it because we own our own IP. We're not just a distributor for other people's IP. I think that's helped us and you've seen us improve both top line and operating margin. We've also moved TPUs.

Now, as for example, when you look at capital markets, one of the things we find that's very interesting is that algorithmic trading was done using numerical computation, which was largely done on traditional compute. That's been constrained by the Moore's law. The incremental improvement you see generation to generation is getting slower. And so many of the top firms have seen the huge improvements they can get by shifting to inference. Instead of doing computation using numerical techniques, if you shift to inference, you can ride the improvements you're seeing in inference time. And as they've come in, they want our machines in their venues,

where it's closer to where the exchanges are, for example.

>> Right. >> So, we've started taking TPUs and making available in other people's data center for some of our key customers and that's a slightly different business model. And we have, you know, in in macro, I would say diversification improves product because you see requirements from many places. Diversification in monetization also helps us grow. I mean, when we deal with supply chain vendors, for example, because we are using these chips not just for our own needs, but also offering them in market, they say, "Well, Google's demand is a sum total of a much larger pool and as a result, we get favorable terms."

>> I want to stick on this point for a moment longer. If compute demand is infinite, I mean, even just for the R&D side, why not why not just hoard the compute? Why not just keep it? And to put a finer point on it, if AGI is is really the goal that all of the AI labs are heading to towards and whoever hits it kind of first and is able to scale it out wins, it seems like keeping the capacity, keeping it for yourself, keeping it for your own models is actually quite beneficial. What am I missing?

>> You have to make money to fund all of this. >> Well, Google makes a ton of money. >> But you have to keep generating cash flow to do it. >> Mhm. >> And this is one more lever for us to generate sufficient cash flow. And the amount we allocate to other people is balanced always against our own needs and our own capital requirements. And you know, no matter which lab you're in, venture capital cannot fund you indefinitely.

>> Yeah. >> And as compute costs grow, if you're running a loss leader business where you're losing money, and you're not making enough money from inference and other techniques to cover the cost of training, as that gap was gets wider, the number of sources you can go to get smaller. >> I've been talking about how Google is in such a unique position. They have the cash cows. They have the chips. They have the models. >> That's right.

>> Do does does your Gemini team ever come to you and say like we don't have enough? Or I I I know I'm really stuck on this point. It's just so wild for me to hear where where these other companies are just not able to keep up. >> There's always demand for these things and it'll I think for the next 10 years there will always be more demand than supply. And that's a good place to be in if you have your own chip. If you don't, you're reselling other people's stuff. And in a capacity constrained environment, your unit economics get more expensive.

And in our case, because we control the chip, >> >> the unit economics remain attractive. So, that is going to be an advantage for us because we own the silicon. >> So, if you look at the whole pie of your TPUs, of your compute infrastructure, can you talk a little bit about what the split is between training, inference, selling TPUs, serving inference for other labs? >> Broad brush, I think we don't talk about the details, so I'm not going to go through every element.

>> Sure. >> But broad brush, if you look in macro, cloud is about half of Alphabet's capital, and it's growing because it's growing much faster, as you know. So, that's like a split, and then on our side, a significant part of our growth is coming from Gemini and our models. And so, you can use that as a rough approximation. >> Okay. You you mentioned data centers and building out data centers. Can you explain what the difference is

between construction and manufacturing when you're talking about data centers?

>> Yeah, it's just what is the unit at which you're deploying capacity. So, you could take, for example, a rack of machines, assemble it in a data center. You could take an entire row of machines and deploy it in a data center. The higher the grain at which you can deploy, the more you can pre-construct it, pre-test it in a central location, which means you're much faster in deployment. >> When you're planning deployment of a new data center, I I you're probably more aware than anybody there's a pretty negative sentiment about data centers in the US specifically. I think it's 20% favorability.

H- h- how do you think about that? And and h- how do you think the broader AI industry can start to change the opinion, change the sentiment around artificial intelligence, and specifically deploying data centers, which gives the US a strategic advantage. It's I'm I tend to be very optimistic about AI in general. So, how how do you think about that? >> What people are really concerned about on data centers is a couple of things. Number one, you know, will energy costs in my state or my county go up?

>> Right. >> Second, will there be sufficient employment in the local community in which the data center operates? And so, there's a couple of things that we're doing. First one is we're investing in behind-the-meter technology where we're not taking energy off the grid and we cross connect to the grid if the state wants it so that in case there's short supply on the grid our energy can fuel the grid. We're investing in alternate forms of energy because we think that the traditional mode of generate and distribute is not necessarily the only way that energy supply will come into the market. And so one of the things we're we're looking at is can you reduce the unit cost of energy with new forms of energy delivery created by the demand for AI but then can serve the broader market. Third, we take a lot we pay a lot of attention to making sure the unit of energy we're consuming what's called PUE that we have the best in the industry meaning if you need a 100 megawatts of compute how little incremental megawatts do you need from the energy source so you're not wasting energy? And we're by far the most efficient of that in the world.

And there's a thousand things that go into that on thermodynamic exchange, how we do heating, all of that. Lastly, we're investing in the communities in which we're in. And to avoid the communities feeling like Google's deploying in one giant location, we distribute it in many places so that no individual state feels like we're becoming a big burden on their resources. And we've had a great track record. I travel to many many of our data centers and when you go to the local economy and see the kids in the school systems and you see the employees who operate our data centers who are super important to us, how much economic development we bring to those rural communities where you know, we think that's part of our responsibility.

>> That's fantastic. What about the broader sentiment in not the local community cuz when you go in, you create jobs, you're investing, you're not using the electricity and driving up the prices directly. that's all wonderful, but how do you actually change the the hearts and the minds of the broader US demographic with regards to artificial intelligence? >> That's going to be a process. You know, and I think it's finding places where you can apply the technology in a way that's good for society rather than just you know, causing issues where people are worried about job displacement. I'll give you just a few examples of things.

>> You'll see in our you know, keynote the company called Signal, we do not talk about it. They're a health insurer based in Germany. They're the largest health insurer in Germany. They today deploy a lot of agents built on Gemini Enterprise to to help their teams do work. Now, really interesting thing was there was a lot of anxiety when we started the work with them that it would mean job displacement. They have not let go anybody.

And in fact, what they found is the accuracy with which and the speed with which they can answer questions from people about am I eligible for this treatment or not? It's cut down from in some cases 23 minutes to research and answer now to less than a few seconds. And so, that's improved efficiency, it's improved quality of their customer care, and they've not let go a single job.

we work for example with the American Society for Clinical Oncology. They're the largest 51,000 members, every oncologist in the United States. They wanted an application where AI was helping a doctor sitting down to deal with a patient to understand standard of care guidelines, which is this person's come in they have breast cancer, what's the guideline? It turns out they're also diabetic. I can't prescribe chemo if they're diabetic of this kind. There's a lot. These rules are incredibly complicated. In many cases, they overlap.

And they wanted some help on providing answers. The answers have to be 100% correct. You can have hallucination. And we've helped them. And it helps doctors take care of patients. And the feedback from their membership has been incredibly rewarding to see. So, there are lots of examples. And we always say the most important thing, Citigroup, for example, is building a wealth advisor. So, today, if you think of the average citizen, they get if you're a high-net-worth person, you can go to a private bank and you have wealth management professional advise you.

If you're an average person who doesn't have those financial resources, you may not get high-quality advice. Citigroup is building a wealth advisor. They're going to be showing it at the event, which allows you to use our reasoning and task management capabilities in Gemini to advise people and also to help them take investment actions if they want. So, these are examples of things that society's going to find beneficial. It takes time to get the balance between the AI's going to cause massive job displacement to also hear from this side of things. And that's part of the journey we're on as a as a society.

>> I I do agree. I think job displacement in particular is is something the general population in the US is extremely worried about. Let me ask you directly for your organization, Google Cloud, now that you're seeing your engineers and and other parts of your organization be more productive because of artificial intelligence, more automation, are you hiring? Are you letting go folks? Are you stable? Where where are you on that front?
>> We're adding people for products and sales. We're hiring a lot of people in our go-to-market organization. We're hiring a lot of forward deployed engineers. And in places where we're building new product, we're adding capability. Like an example I will tell you here's an example of what people don't see. A long time ago we said as models became more sophisticated in understanding code and second as models learn how to use computers to do tasks there's lots of things they can do amazingly well.

But one of the issues with understanding code is they can also find vulnerabilities in code. >> Mhm. >> And so enormous anxiety about cybersecurity vulnerabilities from some of the new models. >> We're going to talk about that. >> made a decision a long time ago to do three things. Number one helping improve Gemini as a way to detect issues in code and we've a lot of customers using it. Second, helping build a model that can repair code.

Because if you're finding vulnerabilities very quickly, people may not be able to keep up. So can a model assist you in fixing it? And we have new capability coming for that. When we acquired this company Wiz, you'll see us showing new capability with Wiz and it's really about continuous detection. >> Mhm. >> You know, people call it continuous red teaming. So we're going to show three different types of agents. An agent that continually attacks you to make sure your vulnerabilities are being fixed and you don't get caught off guard which you couldn't do before.

An agent that prioritizes the issues that are being discovered so that you can get okay, these are the ones I really need to fix. And then a third one that helps you fix >> I'm glad to hear you're you're still hiring. >> Yeah. more >> and hiring. I I There are companies out there. I think Block is the the big one. You know, Jack Dorsey put out this blog post. Block laid off half of their organization, blamed AI or pointed at AI as the reason. What do you think the difference is between like how Google sees this productivity increase and is still increasing employment versus Block who said, "No, we're actually going to transform the company. We need We need half as many people and we're going to do things better." What Where's the discrepancy there?

>> Every company has demand for its products and services and each CEO makes their own decisions. We're seeing plenty of demand and so we're investing. >> Let's talk about Nvidia for a moment. Jensen just did a podcast with Tarkash and he talked about how Nvidia and their architecture is the cheapest on a per token basis overall total cost of ownership. That's because of CUDA and NVLink networking tooling delivering better tokenomics. Can Do you Do you agree with that assessment? Do you think Google is the best overall total cost of ownership?

and if not, how does Google catch up? >> We have a lot of customers who say we are the best total cost of ownership. >> Yeah, I guess that was the answer, right? >> I mean, the reality is if you're an AI lab, you choose the best platform. It's not just our own teams that use it. We have more demand than we can possibly meet from all the other AI labs. And so, I would just tell you that they would not be asking for TPU if you were much more expensive.

>> Is a big factor of what makes TPU special the speed? I've noticed the Gemini family of models is very fast and as a speed maxi myself, I very much appreciate the speed. And typically when you're looking at ASICs, they're specialized, they tend to be a lot faster than the generalized GPU. Like is that a selling point for a lot of AI labs or for your own customers or are they still saying quality all day, quality, quality?

>> It's a combination, I would say, three core elements because I think it's not the chip, it's the system. A TPU system, for example, 8T has 9,600 chips. 8I is, I think, 1,152. All on a single optical torus network. So, there's incredibly high bandwidth, super predictable latency across all the the the chips in a pod. And that gives you, for

example, when you look at the speed with which we're able to take stuff out of the memory for processing and to put stuff back in memory, it's extraordinarily efficient.

just to give an example, 8T, the training chip, can fit 2 petabytes of memory in a single system. 2 petabytes is like 100 times the size of all the Library of Congress digitized. >> Yeah. >> and because it's the super low latency network, your throughput from memory into the chips themselves are extremely fast. Third, as as if you look above that layer, from a programming stack point of view, you have a lot of tools that Google has built and given to the industry that's used for compiler optimization.

For example, JAX, we've done great work with PyTorch, you know, XLA, Pathways. These are all technology that Google's built. And so, do you you put all of that together. And even if you look at inference via LLM, there's a number of technologies that are very super optimized. It's that whole stack that makes that TPU system so efficient and so powerful. And you see that measured through what we call goodput. Goodput is how much effective throughput are you seeing? We also made some decisions a long time ago, like 3-4 years ago, for instance, we saw that energy, to your point earlier on energy, was going to be short supply. So, we focused on optimizing the dollars per watt or tokens per watt, and I think that's another element that you see a lot of people wanting.

>> So, you you've talked a little bit about planning and the TPU remind me you said 11 years? >> Yes. >> 11 years ago. It's kind of wild to see that a decision made so long ago, long in in the tech world, has bared so much fruit in the last few years. how how much change, how much variance in your planning happens based on what you're seeing in the market today? do the decisions that you made years and years ago apply and they're just steadfast, or are you having to change things constantly?

>> I would say the the the history we have across the different layers of the stack has compounded over time. When we did TensorFlow, we realized you needed a large-scale distributed programming model for training, and we built JAX, for example. That was something that was compounded on our history of learning from what people were trying with TensorFlow and needing a new distributed you know, training model, right? So, some of these accumulate over time because we learn from what we're doing from prior, and we're we're making new improvements.

We're also incredibly attuned to the market, listening to customers, making decisions like people asked us, "Why did you build 8i, you know, the inference chip?" It's because it's we've seen that as you eventually, no matter how rich you are, you cannot fund training without making money on inference. And so, you have to at least cover the cost of your training from a break-even point of view over time. You can't just always depend on venture capitalist to fund you.

and so, we said there's going to be a big demand for inference. We knew what the factors we needed to optimize for inference. And you know, frankly, the demand for the inference the 8i has been way, way more than we expected. >> So, let's talk about the eighth-gen chip. So, this is the first time where you have split out two two different chips, family, but two different chips, one for inference, one for pre-training. first, just confirm

Ironwood was more built for inference.

>> Ironwood was a mixed chip that's used for training and inference. >> Yeah, I think there was >> people run to for example, inference, there's a lot of diurnality to it. During for like chat, during the daytime, people wake up and they ask a bunch of questions. At night, even some people still sleep. And so, at that time, a lot of people were using a spot for inference. Like post-training, a lot of people were doing on spot instances at night.

So, it's a it's a general-purpose chip. with eight, T is mostly for training. Some people are considering using it for inference. And I is primarily for inference. Although, people with smaller models also use it for training. >> Based on the fact that you decided to split the chips, what does that say about where the workloads are heading? What do you see today? And then what what do you think we're going to see over the next, let's say, 5 years? Where are the major workloads going to be?

>> So, we You see that in the work we're doing with Gemini as much as in the silicon. So, if you look at Gemini, we've seen sort of three phases, if you will, with the models. The first phase was where people were asking the model a set of questions, and it was answering, and you may iterate on it in a multi-turn, but it was primarily kind of a a search chatbot-like experience.

So, our Gemini Enterprise product does provide the ability to do search and answer questions. It also added deep research to do deep analysis. >> Mhm. >> Then the second phase came along where people used to use diffusion models primarily to create content, like images, audio, video, and then with 25 nano banana, we added media in was always true, but media out became part of the main model.

And so, we saw people from creative, for example, WPP, a variety of CPG firms using Gemini Enterprise, our enterprise AI platform, to create content. And and there's all kinds of content creation now going on with it. >> Mhm. >> And then, the models became really good at dealing with the abstractions of the world. And when I say abstraction of the world, if you go to a company, the model has to be hooked into a variety of different systems. It has to talk to your CRM system to ask and answer questions about customers.

It may have to look at your supply chain and planning systems. And as the models became really good at dealing with those, and the ultimate abstraction is abstracting the rest of the world as a computer. Because if you can talk to a computer, the computer can talk to everything. Because all these forms of software are just abstractions for how a computer can talk to it. >> Do you do you think that is the ultimate abstraction, a model being able to control computers, computer use, browser use.

>> But understanding the information that comes from those systems as well. It's not just I can talk to a computer, but I need to be able to respond to the information that computer gives me. Do you see what I mean? And so that's then led to this notion of agent. An agent is you know, a module, let's call it that, that you can delegate tasks to. The agent describes itself with a set of skills and it knows how to operate a set of tools, and then it can operate those including a computer and do tasks in your behalf.

Now, for us, that allows people whether it's Xfinity using us for their you know, to schedule and manage all their customer care, Walmart using us for a variety of things in their organization from planning to scheduling, Bosch in manufacturing using us. Merck has talked about how they're using us for research and you know, patient from the drug discovery out to delivering to patients the whole cycle being automated.

And that's the next phase of evolution. And so part of it is we're sort of co-designing as the skills of the model advance, we're able to broaden the set of things that can be done. >> Tie it back to how that informs the decision to split up the two chips between inference and training. >> So, if you look back when you asked search a quest the first phase, there were a lot more input tokens than output tokens. When you asked the model a set of questions, cuz you would ask it a very complicated long described question and it would say this is the answer.

Then, when you came to content generation, you would give it a simple prompt, create a video that shows my dog wearing a Superman cape and driving a car. And then it would take a while to generate the output tokens. So, that generated a very different mix of tokens of the type of tokens. Multimodality was one big thing, and then the volume of output tokens grew. Then you come along to agents.

It informed chip design in three or four different ways. It informed us on how long you need to maintain stuff in memory. So, for example, how what kind of KV cache would you want? Because now you're delegating something that could run for 6, 7, 12 hours. You don't want to be shuffling things in and out as tokens because they get expensive. So, that's one example of something. Second example. You want this talk this system to operate a computer.

By the way, that computer is a traditional classical compute machine. >> >> So, when people asked us, "How did that inform your chip effort?" We not only work with Intel. We all have our own arm chips. And so, we built that because we saw a general purpose compute usage coming from these tools. When you run for inference an agent that does many, many different steps, there are things about how you want to hold and pin objects in memory in the model so that the model runs things super efficiently because that can really optimize the cost of inference.

There are many things we've done internally on how the chip can hold things in memory. and then because people wanted even a practical example, people want inference in many locations cuz they want to manage latency, unlike training where you can put it in a few big locations. So, practical example is 8i can be run in non-water-cooled mode so that you can put it in many more locations cuz air cooling is still the primary thing in most data centers. So, there's a lot of thought that goes into these decisions. I'm just giving you three simple examples to illustrate.

>> Yeah, I I think that the aging piece is really interesting because it really changes the way that those tokens are actually used in practice. obviously Nvidia talks a lot about extreme co-design. Google seems like they have extreme co-design on every layer. >> Yes. >> First talk about like with agentic usage especially if you're doing a lot of read and writes to a hard drive or you you know there there's like a lot of pieces that you need to optimize for. What's the latest thing that you've optimized for in the TPU stack and then where do you think the next big

bottleneck is based on agentic usage growth?

>> So we look at the whole system all the time. couple examples. We're announcing two new storage solutions next week. one is our managed Lustre solution. We've improved it to run 10 terabytes per second throughput. It is really designed for large-scale training. >> Mhm. >> So you can cross-connect it to a giant cluster. And because you have large data sets you can read them from the the large-scale Lustre cluster now into a large training fleet for super efficient scale. So that's one.

Second thing we introduced is a a new ultra-low latency inference storage system a rapid storage. and the idea is you can centrally keep the information you want to inference on in our cloud storage. But you can mount it close to where think of it as a forward proxy like thing wherever your inference chips are running. And so from your inference processor down to the storage system rapid storage to fetch for inference It's incredibly fast. it's 15 terabits per second.

So you get ultra low latency. You want to optimize all this stuff on a common network backbone. So they're introducing a new form of networking called Virgel, which gives you ultra low connectivity speed across a giant cluster. so there are many many other parts to the stack we're also co-designing because of agents coming in. And the idea is to give people the most efficient cost structure to run agents with the best performance and quality. >> Where's the next big bottleneck?

>> The next big bottleneck will be largely around when consumers use virtual machines. You know, they Let's say I'm a consumer at home. I build an agent. And the agent is going to schedule travel for me, just hypothetically, if you're going on vacation. And you ask it to do a bunch of tasks like go look up eight travel sites, which are exposed as tools. You know, this common thing now people call MCPs or APIs.

Let's go find all the travel sites. Let's say it's booking a trip to Europe or to Southeast Asia. Run that, calculate for me the total cost, and tell me my budget. Consumers cannot afford to have VMs running forever. It's extremely expensive, as you know. So people want to activate deactivate VMs whenever a task gets done. And because they these these tools need local storage, they these virtual machines can be over subscribed, but you can also have local disk in from which you read and write super efficiently. And so that's going to be a bottleneck because it's going to directly affect how widespread you can make this technology available.

Because companies can pay for stuff. Obviously, the cheaper and more efficient they can use more stuff. But if you want to bring this to consumers, you know, for them it gets expensive very quickly. And if you want to reach everybody, you're going to have to engineer the cost structure of these things. And having again that ability to go across the layers from the agent down to Gemini, down to the storage system and the compute systems, that allows us to co-design.

>> Thank you for sharing that. I want to talk about Anthropic a little bit. Anthropic is one of Google's customers. >> Yes. >> They are a unique company in a lot of ways. Claude is one of Google's biggest rivals at the same time. Yet, you are essentially their backbone for a lot of the training, a lot of the inference. How do you

think about that decision? And I know we touched on it earlier, but I want want to go into more detail. How do you think about powering Anthropic's models? And then they are also competing with Google. Is that the AWS playbook where it's power everybody and we're just not going to play favorites, or is it something different?

>> Google's a platform company, you know, so when you're a platform company, different parts of business compete with different players in the market. Some parts of business may supply them and some part of the business may compete with them. And so we're determined to be best in class in the models. And we're very proud of what we've done, not just with Gemini the model, but also the whole tool chain that we're bringing around Gemini with our enterprise portfolio of tools.

At the same time, you know, there are customers who want, for example, our TPUs, and so Anthropic is an example of them. And it's just part of being a platform company. It's the same way that people ask us, how well do you optimize your model with Apple? Apple, for example, has signed a contract with us for the model, as you know. And so, people go, "Isn't that competing with your Android platform and ecosystem?" Yes, but that's part of being a platform company.

>> Yeah. I think I'm I'm kind of stuck on the Anthropic piece because I mean, they are competing on the enterprise level where Apple is not. Yeah, I'm I'm just thinking you're you know, you're you're powering them, and then at a certain point we may get and although there's plenty of TPU capacity to go around right now, as you said, but at a certain point there might have to be a difficult decision. How do you make that decision of well, can we give the capacity to an Anthropic, or do we keep it for Gemini, do we keep it for our own research? How do you make that decision?

>> We have an executive team with Sundar, and we discuss these, and as any mature company, we make those decisions. There's difficult calls every day. For instance, we have demand not just from Anthropic. So, what percentage for even if you said there's X amount for Gemini, and there's Y amount for the rest of the world, what amount do you give Anthropic versus the hundreds of other labs and other customers ask us for it? That's all complicated decisions that anybody has to make. What I'll tell you is this, it's better to have your own chips and demand than not having your own chips.

>> Yeah, well said. Mythos, you you kind of hinted at a little bit. I think it's rumored to be the first 10 trillion parameter model. Is Google playing in the 10 trillion parameter model space yet? Are you Are you close to it? Where are you in that in that life cycle? >> You'll see new stuff from us on Gemini with announcements coming both at next and soon after. I think on the the capability of the model, we're very proud of where Gemini is.

I mean, it's been state-of-the-art for a long time. We have a new version of Gemini coming very, very soon. And from all the benchmarks we've seen, we've been very confident on that as well. >> So, hypothetically, if you think about a 10 trillion parameter model based on what you oversee on the TPU side, is that even a feasible size to serve in the current state of the the the world? >> We've had a capability to do disaggregated serving, which allows us to scale very large dense models super well, and that's been in place for a long time. And so, we're able to we would not design a model that we couldn't serve.

And so, we're very confident if TPUs can serve the largest models in the world, and most importantly, our serving stack we use for disaggregated serving, by definition, is the most efficient on TPU in of all the model providers in the industry. So, we're very confident we can serve the largest models, particularly the largest Gemini models. >> Does this mean that we're not seeing any slowdown on the scaling pre-training side? You're you're not feeling it at all, cuz there was for a while in the industry people talking about pre-training is slowing down. Now, let's focus on RL, let's focus on the thinking time.

You're not seeing that at all. >> We're not seeing that from the point of view of chip design or system design or lack of capacity or any of that. >> And then what about the underlying data? Are you seeing more an an effective use of synthetic data? >> We are seeing I mean, I'll give you two or three examples of things we are seeing. Historically, a lot of the data that was fed into models was unstructured data, like text, audio, video files, etc. Those continue to grow.

But, the reality of those is like there are many elements in an enterprise context which make them actually really simple to deal with. When you ask a question to an agent and you have the agent respond and tell you, "Tell me the citation of where did you derive this answer from? It's easy if it's in a document because you can just show a link to that document. Now, just imagine you ask the model a question, tell me how much inventory we'll need to meet demand for this product.

That is going to translate to a query on a system like an SAP system or some kind of supply chain system. That's dynamically going against set of tables. So, first being accurate on decom- decomposing that query into which table is it getting it from and showing the response like where is the citation? Like, how did you get this How do I know the answer that you gave me is correct? Is a much more complicated problem. And so, because of the work we do in enterprise, we're able to feed Gemini a lot more cycles into our you know, trajectory optimization harness with structured data. Complex things like complex fields. Have you ever seen a You know, when you talk about computer use in a browser use? If you ever see an enterprise application with a thousand fields, drop-down lists, etc.

There's no consumer app that would ever have that complexity. >> Yeah. >> So, being in this space also allows us to teach our Gemini system some of those things and put it into the harness. >> I I I let's let's continue on harnesses and agentic coding in general. I I've been doing a lot of coding myself. There was kind of a viral tweet that went around about somebody who had a friend at Google who basically said Google isn't on the frontier of agentic coding internally.

What's your take on that? How How has Google adopted agentic coding? And especially again, I have to bring up Anthropic. The rate at which they're shipping is incredible. How is Google adopting the frontier of agentic coding today? >> We have a lot of engineers using Jetski, which is our internal coding harness. And that feedback has been going directly to DeepMind in the you know, in that reinforcement loop, and it's improving quality of Gemini for coding every day.

And we have a lot of people in my organization using it. >> One thing I've noticed, I'm more productive than

ever. >> Mhm. >> I'm shipping so fast. I'm having so much fun doing it. I'm not reviewing every line of code. >> Mhm. >> Actually, I'm reviewing very few lines of code. >> Mhm. >> But Google can't do that. I have little toy projects. Google, you have high-stakes projects and services and products that you're serving. How how can you both be on the frontier of agentic coding and be produ- producing so many lines of code, but also making sure that you're maintaining that quality. Also, make sure that you are actually reviewing every single line of code that gets deployed.

>> So, when we talk about software engineering productivity, we look at it slightly differently than is reported externally. So, if you work in a in a company that builds products like Google does, the reality is that there are two or three examples of things that you find that are really important. Like a senior engineer writes much more compact code than a junior engineer. So, we don't count how many lines of code as a measure because that's generally, you know, a a weaker engineer writes a lot more code to do the same task that a senior engineer does.

>> It's kind of a cliché, right? Over the years, that's don't don't count lines of code. But I think now more than ever, it's just the shipping speed overall. >> so it's how much functions do we add? That's important. >> Yeah. >> The second thing, we've always had a tradition at Google that when you go to check in code, you need peer review. Typically, the peer review is done by senior managers, right? So, and they become the bottleneck. So, we've introduced and people are using Gemini and we, for example, recently in Cloud introduced it to scan for security vulnerabilities in code.

So, it's not just that the tool is being used to generate code, we're also using it to inspect code. And that helps us get when the senior engineers come in for the review, a bunch of free work has been done. The third one is, for the long-term, in any real software company, the bulk of the time of the engineers where they find doing less productive work is debugging issues. So, we built a version of Gemini and one of the things we're going to show next week is, you know, what's the most complex computer in the world?

The most complex computer in the world is a cloud. It makes a PC look like a toy. And so, we've taken all of our cloud and exposed it as tools to the model. And so, now we're using Gemini to troubleshoot incidents happening. And so, that's also helped us improve the speed with which people can function and in turn improve the quality of the model itself. So, there's a number of dimensions that through which we look at the issue.

>> But as productivity increases and you're shipping more features more quickly, I know lines of code is not the measurement, but it's certainly an an output of this increased velocity. >> Yes. >> There comes a a point where you just cannot review every single line of code. And then, I think kind of if you think about abstracting and going beyond that, there's a point at which humans are understanding the actual code less and less over time, especially as you mentioned, if you're using AI to review the code to debug. So, if you're having AI create code, AI review code, are we losing the core understanding of of code and the functionality being deployed.

>> That's a risk that we have to manage as an industry. People talk about I'm going to give you a prompt and the prompt's going to generate a block code. >> Mhm. >> And you don't need to understand the code because you understand the prompt. In reality, for a complex system, the prompt will not explain all the potential behavior

of the system. >> Possibility. >> Right? >> Yeah. >> And so, for example, how do you deal with exceptions?

and so, that's something that I think every time you find there's one area, like if some time ago, people said you won't need all these software engineers. And then along comes the model and finds a lot of security vulnerabilities. And just when we need a ton of software engineers to work with models. Like we're introducing a version of our model that can actually fix bugs. A fix security vulnerability, specifically.

>> Yeah. >> But you still need a human to use the tool and and focus on it. Sometimes the industry over-rotates and so, you say you don't need anybody just when you need it. >> Yeah. >> And so, we we take a much longer term view of things. And so, we're constantly looking at for instance, do you need a supervisor model to look at code in a different way to actually review the code? And that's why when I said we still do peer review of the code. And we're helping our senior engineers use the tool to do the reviews.

And then the question comes, will the tool be self-aware enough? If it generated the code, will it find an issue with code that it generated? Because it's not self-aware of certain patterns. That's something we're looking at approaches to solve. And so, our goal has always been to make sure we have the best model is to apply it at scale. And in my team alone we have thousands of people using it every single day. I mean, if you walk just over there to the campus, you can see people like six different windows open, one in which they're coding, one in which they're compiling, one in which they're you know, deploying and testing and another one where they've got a background job running to run code review. I mean, the the whole there's a lot of people using the JetSki tool harness and it's part of just evolving how work is getting done.

>> You touched on cybersecurity. Let's finish on that. Anthropic decided the Mythos model was too advanced in the cybersecurity capabilities to release publicly, at least not yet. For Google, how do you think about that? What was your reaction? And then also, is there some line in the sand, some benchmark that you think or you would determine Gemini is no longer safe to release publicly? >> We're working through that on what would that line be. But our you know, our issue has been So, if Mythos finds a set of issues, what percentage of those issues could be found with an open-source model?

And the reason I mentioned open-source model is, no matter how much you defend and you can say, "Well, I'll make sure closed-source models don't fall in adversary's hands." Open-source models for sure are going to fall into adversary's hands. >> And they're just getting better. >> And they're getting better. So, sooner or later, some part of this, it may not be all of the patterns, but some part of it can be detected. So, what should you do in response? And we are unique because we're a hyperscaler, we're a model provider, and we also have a cybersecurity organization.

Both our Mandiant team and Wiz. So, we've done three practical things. If people are going to find issues using a model, you need to have a model help fix issues because they're going to find them way faster than humans can fix. So, you need a model to help fix and so we're looking at something there. Second, if they're going to find issues with models, they're going to cause use the model and computer use to launch a large-scale attack.

And so to defend that, using like I'll red team my system once a month is not going to be sufficient. So, introducing agents that can do continuous red teaming and agents that can actually help fix. Like, for example, it's one thing to fix the code. It's the second thing to find all the places the old code was running and to remove it and then deploy the new code that's been patched and updated, right? So, that's the second piece.

And the third piece is there's so much code out there. What do I start with? >> Yeah. >> So, again, that's another thing where we built tools to help people identify and prioritize what. >> Is it Is this an argument for or against open-source software? Not models, but software. If you're open source, your your your your all your code is out there. It is ripe for models to go look at it, find vulnerabilities, and exploit them.

Close source, you don't have that problem. But on the other hand, open source is going to get hardened much more quickly. What is your take on that? Is that an argument for or against? >> No, we we as Google use a ton of open source and we contribute a ton of open source. We're going to help the open-source community using our tools to actually go fix these things. I'm just pointing out the reality of where things are is that adversaries are going to use the model and the first place they're going to try and scan is popular open-source libraries.

>> Yeah. Yeah. >> Because that gives them the maximum surface area to try and attack. And so it's those are all elements where we think it's important to go and address and fix and we're in process with the rest of the industry. >> Thomas, last question for you. What keeps you up at night? >> We, you know, we're balancing so many things. Making sure we have to one part of your discussion, do we have right term long term plans for capital infrastructure, for data centers, networks, enough of those lovely TPUs to go around.

Second, are we constantly pushing the domain problems, the important problems? Three years ago when we said we should solve the problem of as AI gets better, cyber is going to be definitely an area that's affected. And when we made the offer to buy Waze, people asked like, why would you guys be doing that? When we look at our Gemini enterprise platform, just to give you an example, between January and now, our token count has jumped from 10 billion a minute to 16 billion a minute.

And the number of enterprise users of Gemini enterprise has jumped by 40% sequentially. So, we're always looking at how we're solving the right problems for for customers and users, and that's always the focus for us. And as long as we keep pushing aggressively and solving those problems and staying ahead of the market when the the technology is evolving so quickly that when something happens, you got to have solutions before that occurs to the most part, and our teams have done an amazing job and we're super proud of what they've done and looking forward to the event.

>> Thomas, thank you so much. Really appreciate it.