Dwarkesh Podcast #72 - Mark Zuckerberg - Llama 3, Open Sourcing $10b Models, & Caesar Augustus

Published - April 18, 2024

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Mark, welcome to the podcast.

Mark Zuckerberg
Thanks for having me. Big fan of your podcast.

Dwarkesh Patel
Thank you, that's very nice of you to say. Let's start by talking about the releases that will go out when this interview goes out. Tell me about the models and Meta AI. What's new and exciting about them?

Mark Zuckerberg
I think the main thing that most people in the world are going to see is the new version of Meta AI. The most important thing that we're doing is the upgrade to the model. We're rolling out Llama-3. We're doing it both as open source for the dev community and it is now going to be powering Meta AI. There's a lot that I'm sure we'll get into around Llama-3, but I think the bottom line on this is that we think now that Meta AI is the most intelligent, freely-available AI assistant that people can use. We're also integrating Google and Bing for real-time knowledge.

We're going to make it a lot more prominent across our apps. At the top of Facebook and Messenger, you'll be able to just use the search box right there to ask any question. There's a bunch of new creation features that we added that I think are pretty cool and that I think people will enjoy. I think animations is a good one. You can basically take any image and just animate it.

One that people are going to find pretty wild is that it now generates high quality images so quickly that it actually generates it as you're typing and updates it in real time. So you're typing your query and it's honing in. It's like "show me a picture of a cow in a field with mountains in the background, eating macadamia nuts, drinking beer" and it's updating the image in real time. It's pretty wild. I think people are going to enjoy that. So I think that's what most people are going to see in the world. We're rolling that out, not everywhere, but we're starting in a handful of countries and we'll do more over the coming weeks and months. I think that's going to be a pretty big deal and I'm really excited to get that in people's hands. It's a big step forward for Meta AI.

But I think if you want to get under the hood a bit, the Llama-3 stuff is obviously the most technically interesting. We're training three versions: an 8 billion parameter model and a 70 billion, which we're releasing today, and a 405 billion dense model, which is still training. So we're not releasing that today, but I'm pretty excited about how the 8B and the 70B turned out. They're leading for their scale. We'll release a blog post with all the benchmarks so

people can check it out themselves. Obviously it's open source so people get a chance to play with it.

We have a roadmap of new releases coming that are going to bring multimodality, more multi-linguality, and bigger context windows as well. Hopefully, sometime later in the year we'll get to roll out the 405B. For where it is right now in training, it is already at around 85 MMLU and we expect that it's going to have leading benchmarks on a bunch of the benchmarks. I'm pretty excited about all of that. The 70 billion is great too. We're releasing that today. It's around 82 MMLU and has leading scores on math and reasoning. I think just getting this in people's hands is going to be pretty wild.

Dwarkesh Patel
Oh, interesting. That's the first I'm hearing of it as a benchmark. That's super impressive.

Mark Zuckerberg
The 8 billion is nearly as powerful as the biggest version of Llama-2 that we released. So the smallest Llama-3 is basically as powerful as the biggest Llama-2.

Dwarkesh Patel
Before we dig into these models, I want to go back in time. I'm assuming 2022 is when you started acquiring these H100s, or you can tell me when. The stock price is getting hammered. People are asking what's happening with all this capex. People aren't buying the metaverse. Presumably you're spending that capex to get these H100s. How did you know back then to get the H100s? How did you know that you'd need the GPUs?

Mark Zuckerberg
I think it was because we were working on Reels. We always want to have enough capacity to build something that we can't quite see on the horizon yet. We got into this position with Reels where we needed more GPUs to train the models. It was this big evolution for our services. Instead of just ranking content from people or pages you follow, we made this big push to start recommending what we call unconnected content, content from people or pages that you're not following.

The corpus of content candidates that we could potentially show you expanded from on the order of thousands to on the order of hundreds of millions. It needed a completely different infrastructure. We started working on doing that and we were constrained on the infrastructure in catching up to what TikTok was doing as quickly as we wanted to. I basically looked at that and I was like "hey, we have to make sure that we're never in this situation again. So let's order enough GPUs to do what we need to do on Reels and ranking content and feed. But let's also double that." Again, our normal principle is that there's going to be something on the horizon that we can't see yet.

Dwarkesh Patel
Did you know it would be AI?

Mark Zuckerberg
We thought it was going to be something that had to do with training large models. At the time I thought it was probably going to be something that had to do with content. It's just the pattern matching of running the company, there's always another thing. At that time I was so deep into trying to get the recommendations working for Reels and other content. That's just such a big unlock for Instagram and Facebook now, being able to show people content that's interesting to them from people that they're not even following.
But that ended up being a very good decision in retrospect. And it came from being behind. It wasn't like "oh, I was so far ahead." Actually, most of the times where we make some decision that ends up seeming good is because we messed something up before and just didn't want to repeat the mistake.

Dwarkesh Patel
This is a total detour, but I want to ask about this while we're on this. We'll get back to AI in a second. In 2006 you didn't sell for $1 billion but presumably there's some amount you would have sold for, right? Did you write down in your head like "I think the actual valuation of Facebook at the time is this and they're not actually getting the valuation right"? If they'd offered you $5 trillion, of course you would have sold. So how did you think about that choice?

Mark Zuckerberg
I think some of these things are just personal. I don't know that at the time I was sophisticated enough to do that analysis. I had all these people around me who were making all these arguments for a billion dollars like "here's the revenue that we need to make and here's how big we need to be. It's clearly so many years in the future." It was very far ahead of where we were at the time. I didn't really have the financial sophistication to really engage with that kind of debate.
Deep down I believed in what we were doing. I did some analysis like "what would I do if I weren't doing this? Well, I really like building things and I like helping people communicate. I like understanding what's going on with people and the dynamics between people. So I think if I sold this company, I'd just go build another company like this and I kind of like the one I have. So why?" I think a lot of the biggest bets that people make are often just based on conviction and values. It's actually usually very hard to do the analyses trying to connect the dots forward.

Dwarkesh Patel
You've had Facebook AI Research for a long time. Now it's become seemingly central to your company. At what point did making AGI, or however you consider that mission, become a key priority of what Meta is doing?

Mark Zuckerberg
It's been a big deal for a while. We started FAIR about 10 years ago. The idea was that, along the way to general intelligence or whatever you wanna call it, there are going to be all these different innovations and that's going to just improve everything that we do. So we didn't conceive of it as a product. It was more of a research group. Over the last 10 years it has created a lot of different things that have improved all of our products. It's advanced the field and allowed other people in the field to create things that have improved our products too. I think that that's been great.

There's obviously a big change in the last few years with ChatGPT and the diffusion models around image creation coming out. This is some pretty wild stuff that is pretty clearly going to affect how people interact with every app that's out there. At that point we started a second group, the gen AI group, with the goal of bringing that stuff into our products and building leading foundation models that would power all these different products.

When we started doing that the theory initially was that a lot of the stuff we're doing is pretty social. It's helping people interact with creators, helping people interact with businesses, helping businesses sell things or do customer support. There's also basic assistant functionality, whether it's for our apps or the smart glasses or VR. So it wasn't completely clear at first that you were going to need full AGI to be able to support those use cases. But in all these subtle ways, through working on them, I think it's actually become clear that you do. For example, when we were working on Llama-2, we didn't prioritize coding because people aren't going to ask Meta AI a lot of coding questions in WhatsApp.

Dwarkesh Patel
Now they will, right?

Mark Zuckerberg
I don't know. I'm not sure that WhatsApp, or Facebook or Instagram, is the UI where people are going to be doing a lot of coding questions. Maybe the website, meta.ai, that we're launching. But the thing that has been a somewhat surprising result over the last 18 months is that it turns out that coding is important for a lot of domains, not just coding. Even if people aren't asking coding questions, training the models on coding helps them become more rigorous in answering the question and helps them reason across a lot of different types of domains. That's one example where for Llama-3, we really focused on training it with a lot of coding because that's going to make it better on all these things even if people aren't asking primarily coding questions.

Reasoning is another example. Maybe you want to chat with a creator or you're a business and you're trying to interact with a customer. That interaction is not just like "okay, the person sends you a message and you just reply." It's a multi-step interaction where you're trying to think through "how do I accomplish the person's goals?" A lot of times when a

customer comes, they don't necessarily know exactly what they're looking for or how to ask their questions. So it's not really the job of the AI to just respond to the question.

You need to kind of think about it more holistically. It really becomes a reasoning problem. So if someone else solves reasoning, or makes good advances on reasoning, and we're sitting here with a basic chat bot, then our product is lame compared to what other people are building. At the end of the day, we basically realized we've got to solve general intelligence and we just upped the ante and the investment to make sure that we could do that.

Dwarkesh Patel
So the version of Llama that's going to solve all these use cases for users, is that the version that will be powerful enough to replace a programmer you might have in this building?

Mark Zuckerberg
I just think that all this stuff is going to be progressive over time.

Dwarkesh Patel
But in the end case: Llama-10.

Mark Zuckerberg
I think that there's a lot baked into that question. I'm not sure that we're replacing people as much as we're giving people tools to do more stuff.

Dwarkesh Patel
Is the programmer in this building 10x more productive after Llama-10?

Mark Zuckerberg
I would hope more. I don't believe that there's a single threshold of intelligence for humanity because people have different skills. I think that at some point AI is probably going to surpass people at most of those things, depending on how powerful the models are. But I think it's progressive and I don't think AGI is one thing. You're basically adding different capabilities. Multimodality is a key one that we're focused on now, initially with photos and images and text but eventually with videos. Because we're so focused on the metaverse, 3D type stuff is important too. One modality that I'm pretty focused on, that I haven't seen as many other people in the industry focus on, is emotional understanding. So much of the human brain is just dedicated to understanding people and understanding expressions and emotions. I think that's its own whole modality, right? You could say that maybe it's just video or image, but it's clearly a very specialized version of those two.

So there are all these different capabilities that you want to train the models to focus on, in addition to getting a lot better at reasoning and memory, which is its own whole thing. I

don't think in the future we're going to be primarily shoving things into a query context window to ask more complicated questions. There will be different stores of memory or different custom models that are more personalized to people. These are all just different capabilities. Obviously then there's making them big and small. We care about both. If you're running something like Meta AI, that's pretty server-based. We also want it running on smart glasses and there's not a lot of space in smart glasses. So you want to have something that's very efficient for that.

Dwarkesh Patel
If you're doing $10Bs worth of inference or even eventually $100Bs, if you're using intelligence in an industrial scale what is the use case? Is it simulations? Is it the AIs that will be in the metaverse? What will we be using the data centers for?

Mark Zuckerberg
Our bet is that it's going to basically change all of the products. I think that there's going to be a kind of Meta AI general assistant product. I think that that will shift from something that feels more like a chatbot, where you ask a question and it formulates an answer, to things where you're giving it more complicated tasks and then it goes away and does them. That's going to take a lot of inference and it's going to take a lot of compute in other ways too.

Then I think interacting with other agents for other people is going to be a big part of what we do, whether it's for businesses or creators. A big part of my theory on this is that there's not going to be just one singular AI that you interact with. Every business is going to want an AI that represents their interests. They're not going to want to primarily interact with you through an AI that is going to sell their competitors' products.

I think creators is going to be a big one. There are about 200 million creators on our platforms. They basically all have the pattern where they want to engage their community but they're limited by the hours in the day. Their community generally wants to engage them, but they don't know that they're limited by the hours in the day. If you could create something where that creator can basically own the AI, train it in the way they want, and engage their community, I think that's going to be super powerful. There's going to be a ton of engagement across all these things.

These are just the consumer use cases. My wife and I run our foundation, Chan Zuckerberg Initiative. We're doing a bunch of stuff on science and there's obviously a lot of AI work that is going to advance science and healthcare and all these things. So it will end up affecting basically every area of the products and the economy.

Dwarkesh Patel
You mentioned AI that can just go out and do something for you that's multi-step. Is that a bigger model? With Llama-4 for example, will there still be a version that's 70B but you'll just train it on the right data and that will be super powerful? What does the progression look like? Is it scaling? Is it just the same size but different banks like you were talking about?

Mark Zuckerberg
I don't know that we know the answer to that. I think one thing that seems to be a pattern is that you have the Llama model and then you build some kind of other application specific code around it. Some of it is the fine-tuning for the use case, but some of it is, for example, logic for how Meta AI should work with tools like Google or Bing to bring in real-time knowledge. That's not part of the base Llama model. For Llama-2, we had some of that and it was a little more hand-engineered. Part of our goal for Llama-3 was to bring more of that into the model itself. For Llama-3, as we start getting into more of these agent-like behaviors, I think some of that is going to be more hand-engineered. Our goal for Llama-4 will be to bring more of that into the model.

At each step along the way you have a sense of what's going to be possible on the horizon. You start messing with it and hacking around it. I think that helps you then hone your intuition for what you want to try to train into the next version of the model itself. That makes it more general because obviously for anything that you're hand-coding you can unlock some use cases, but it's just inherently brittle and non-general.

Dwarkesh Patel
When you say "into the model itself," you train it on the thing that you want in the model itself? What do you mean by "into the model itself"?

Mark Zuckerberg
For Llama- 2, the tool use was very specific, whereas Llama-3 has much better tool use. We don't have to hand code all the stuff to have it use Google and go do a search. It can just do that. Similarly for coding and running code and a bunch of stuff like that. Once you kind of get that capability, then you get a peek at what we can start doing next. We don't necessarily want to wait until Llama-4 is around to start building those capabilities, so we can start hacking around it. You do a bunch of hand coding and that makes the products better, if only for the interim. That helps show the way then of what we want to build into the next version of the model.

Dwarkesh Patel
What is the community fine tune of Llama-3 that you're most excited for? Maybe not the one that will be most useful to you, but the one you'll just enjoy playing with the most. They fine-tune it on antiquity and you'll just be talking to Virgil or something. What are you excited about?

Mark Zuckerberg
I think the nature of the stuff is that you get surprised. Any specific thing that I thought would be valuable, we'd probably be building. I think you'll get distilled versions. I think you'll get smaller versions. One thing is that I think 8B isn't quite small enough for a bunch of use cases. Over time I'd love to get a 1-2B parameter model, or even a 500M parameter model and see what you can do with that.

If with 8B parameters we're nearly as powerful as the largest Llama-2 model, then with a billion parameters you should be able to do something that's interesting, and faster. It'd be good for classification, or a lot of basic things that people do before understanding the intent of a user query and feeding it to the most powerful model to hone in on what the prompt should be. I think that's one thing that maybe the community can help fill in. We're also thinking about getting around to distilling some of these ourselves but right now the GPUs are pegged training the 405B.

Dwarkesh Patel
So you have all these GPUs. I think you said 350,000 by the end of the year.

Mark Zuckerberg
That's the whole fleet. We built two, I think 22,000 or 24,000 clusters that are the single clusters that we have for training the big models, obviously across a lot of the stuff that we do. A lot of our stuff goes towards training Reels models and Facebook News Feed and Instagram Feed. Inference is a huge thing for us because we serve a ton of people. Our ratio of inference compute required to training is probably much higher than most other companies that are doing this stuff just because of the sheer volume of the community that we're serving.

Dwarkesh Patel
In the material they shared with me before, it was really interesting that you trained it on more data than is compute optimal just for training. The inference is such a big deal for you guys, and also for the community, that it makes sense to just have this thing and have trillions of tokens in there.

Mark Zuckerberg
Although one of the interesting things about it, even with the 70B, is that we thought it would get more saturated. We trained it on around 15 trillion tokens. I guess our prediction going in was that it was going to asymptote more, but even by the end it was still learning. We probably could have fed it more tokens and it would have gotten somewhat better.

At some point you're running a company and you need to do these meta reasoning questions. Do I want to spend our GPUs on training the 70B model further? Do we want to get on with it so we can start testing hypotheses for Llama-4? We needed to make that call

and I think we got a reasonable balance for this version of the 70B. There'll be others in the future, the 70B multimodal one, that'll come over the next period. But that was fascinating that the architectures at this point can just take so much data.

Dwarkesh Patel
That's really interesting. What does this imply about future models? You mentioned that the Llama-3 8B is better than the Llama-2 70B.

Mark Zuckerberg
No, no, it's nearly as good. I don't want to overstate it. It's in a similar order of magnitude.

Dwarkesh Patel
Does that mean the Llama-4 70B will be as good as the Llama-3 405B? What does the future of this look like?

Mark Zuckerberg
This is one of the great questions, right? I think no one knows. One of the trickiest things in the world to plan around is an exponential curve. How long does it keep going for? I think it's likely enough that we'll keep going. I think it's worth investing the $10Bs or $100B+ in building the infrastructure and assuming that if it keeps going you're going to get some really amazing things that are going to make amazing products. I don't think anyone in the industry can really tell you that it will continue scaling at that rate for sure. In general in history, you hit bottlenecks at certain points. Now there's so much energy on this that maybe those bottlenecks get knocked over pretty quickly. I think that's an interesting question.

Dwarkesh Patel
What does the world look like where there aren't these bottlenecks? Suppose progress just continues at this pace, which seems plausible. Zooming out and forgetting about Llamas...

Mark Zuckerberg
Well, there are going to be different bottlenecks. Over the last few years, I think there was this issue of GPU production. Even companies that had the money to pay for the GPUs couldn't necessarily get as many as they wanted because there were all these supply constraints. Now I think that's sort of getting less. So you're seeing a bunch of companies thinking now about investing a lot of money in building out these things. I think that that will go on for some period of time. There is a capital question. At what point does it stop being worth it to put the capital in?

I actually think before we hit that, you're going to run into energy constraints. I don't think anyone's built a gigawatt single training cluster yet. You run into these things that just end up being slower in the world. Getting energy permitted is a very heavily regulated

government function. You're going from software, which is somewhat regulated and I'd argue it's more regulated than a lot of people in the tech community feel. Obviously it's different if you're starting a small company, maybe you feel that less. We interact with different governments and regulators and we have lots of rules that we need to follow and make sure we do a good job with around the world. But I think that there's no doubt about energy.

If you're talking about building large new power plants or large build-outs and then building transmission lines that cross other private or public land, that's just a heavily regulated thing. You're talking about many years of lead time. If we wanted to stand up some massive facility, powering that is a very long-term project. I think people do it but I don't think this is something that can be quite as magical as just getting to a level of AI, getting a bunch of capital and putting it in, and then all of a sudden the models are just going to... You do hit different bottlenecks along the way.

Dwarkesh Patel
Is there something, maybe an AI-related project or maybe not, that even a company like Meta doesn't have the resources for? Something where if your R&D budget or capex budget were 10x what it is now, then you could pursue it? Something that's in the back of your mind but with Meta today, you can't even issue stock or bonds for it? It's just like 10x bigger than your budget?

Mark Zuckerberg
I think energy is one piece. I think we would probably build out bigger clusters than we currently can if we could get the energy to do it.

Dwarkesh Patel
That's fundamentally money-bottlenecked in the limit? If you had $1 trillion...

Mark Zuckerberg
I think it's time. It depends on how far the exponential curves go. Right now a lot of data centers are on the order of 50 megawatts or 100MW, or a big one might be 150MW. Take a whole data center and fill it up with all the stuff that you need to do for training and you build the biggest cluster you can. I think a bunch of companies are running at stuff like that. But when you start getting into building a data center that's like 300MW or 500MW or 1 GW, no one has built a 1GW data center yet. I think it will happen. This is only a matter of time but it's not going to be next year. Some of these things will take some number of years to build out. Just to put this in perspective, I think a gigawatt would be the size of a meaningful nuclear power plant only going towards training a model.

Dwarkesh Patel
Didn't Amazon do this? They have a 950MW –

Mark Zuckerberg
I'm not exactly sure what they did. You'd have to ask them.

Dwarkesh Patel
But it doesn't have to be in the same place, right? If distributed training works, it can be distributed.

Mark Zuckerberg
Well, I think that is a big question, how that's going to work. It seems quite possible that in the future, more of what we call training for these big models is actually more along the lines of inference generating synthetic data to then go feed into the model. I don't know what that ratio is going to be but I consider the generation of synthetic data to be more inference than training today. Obviously if you're doing it in order to train a model, it's part of the broader training process. So that's an open question, the balance of that and how that plays out.

Dwarkesh Patel
Would that potentially also be the case with Llama-3, and maybe Llama-4 onwards? As in, you put this out and if somebody has a ton of compute, then they can just keep making these things arbitrarily smarter using the models that you've put out. Let's say there's some random country, like Kuwait or the UAE, that has a ton of compute and they can actually just use Llama-4 to make something much smarter.

Mark Zuckerberg
I do think there are going to be dynamics like that, but I also think there is a fundamental limitation on the model architecture. I think like a 70B model that we trained with a Llama-3 architecture can get better, it can keep going. As I was saying, we felt that if we kept on feeding it more data or rotated the high value tokens through again, then it would continue getting better. We've seen a bunch of different companies around the world basically take the Llama-2 70B model architecture and then build a new model. But it's still the case that when you make a generational improvement to something like the Llama-3 70B or the Llama-3 405B, there isn't anything like that open source today. I think that's a big step function. What people are going to be able to build on top of that I think can't go infinitely from there. There can be some optimization in that until you get to the next step function.

Dwarkesh Patel
Let's zoom out a little bit from specific models and even the multi-year lead times you would need to get energy approvals and so on. Big picture, what's happening with AI these next couple of decades? Does it feel like another technology like the metaverse or social, or does it feel like a fundamentally different thing in the course of human history?

Mark Zuckerberg
I think it's going to be pretty fundamental. I think it's going to be more like the creation of computing in the first place. You'll get all these new apps in the same way as when you got the web or you got mobile phones. People basically rethought all these experiences as a lot of things that weren't possible before became possible. So I think that will happen, but I think it's a much lower-level innovation. My sense is that it's going to be more like people going from not having computers to having computers.

It's very hard to reason about exactly how this goes. In the cosmic scale obviously it'll happen quickly, over a couple of decades or something. There is some set of people who are afraid of it really spinning out and going from being somewhat intelligent to extremely intelligent overnight. I just think that there's all these physical constraints that make that unlikely to happen. I just don't really see that playing out. I think we'll have time to acclimate a bit. But it will really change the way that we work and give people all these creative tools to do different things. I think it's going to really enable people to do the things that they want a lot more.

Dwarkesh Patel
So maybe not overnight, but is it your view that on a cosmic scale we can think of these milestones in this way? Humans evolved, and then AI happened, and then they went out into the galaxy. Maybe it takes many decades, maybe it takes a century, but is that the grand scheme of what's happening right now in history?

Mark Zuckerberg
Sorry, in what sense?

Dwarkesh Patel
In the sense that there were other technologies, like computers and even fire, but the development of AI itself is as significant as humans evolving in the first place.

Mark Zuckerberg
I think that's tricky. The history of humanity has been people basically thinking that certain aspects of humanity are really unique in different ways and then coming to grips with the fact that that's not true, but that humanity is actually still super special. We thought that the earth was the center of the universe and it's not, but humans are still pretty awesome and pretty unique, right?

I think another bias that people tend to have is thinking that intelligence is somehow fundamentally connected to life. It's not actually clear that it is. I don't know that we have a clear enough definition of consciousness or life to fully interrogate this. There's all this science fiction about creating intelligence where it starts to take on all these human-like behaviors and things like that. The current incarnation of all this stuff feels like it's going in

a direction where intelligence can be pretty separated from consciousness, agency, and things like that, which I think just makes it a super valuable tool.

Obviously it's very difficult to predict what direction this stuff goes in over time, which is why I don't think anyone should be dogmatic about how they plan to develop it or what they plan to do. You want to look at it with each release. We're obviously very pro open source, but I haven't committed to releasing every single thing that we do. I'm basically very inclined to think that open sourcing is going to be good for the community and also good for us because we'll benefit from the innovations. If at some point however there's some qualitative change in what the thing is capable of, and we feel like it's not responsible to open source it, then we won't. It's all very difficult to predict.

Dwarkesh Patel
What is a kind of specific qualitative change where you'd be training Llama-5 or Llama-4, and if you see it, it'd make you think "you know what, I'm not sure about open sourcing it"?

Mark Zuckerberg
It's a little hard to answer that in the abstract because there are negative behaviors that any product can exhibit where as long as you can mitigate it, it's okay. There's bad things about social media that we work to mitigate. There's bad things about Llama-2 where we spend a lot of time trying to make sure that it's not like helping people commit violent acts or things like that. That doesn't mean that it's a kind of autonomous or intelligent agent. It just means that it's learned a lot about the world and it can answer a set of questions that we think would be unhelpful for it to answer. I think the question isn't really what behaviors would it show, it's what things would we not be able to mitigate after it shows that.

I think that there's so many ways in which something can be good or bad that it's hard to actually enumerate them all up front. Look at what we've had to deal with in social media and the different types of harms. We've basically gotten to like 18 or 19 categories of harmful things that people do and we've basically built AI systems to identify what those things are and to make sure that doesn't happen on our network as much as possible. Over time I think you'll be able to break this down into more of a taxonomy too. I think this is a thing that we spend time researching as well, because we want to make sure that we understand that.

Dwarkesh Patel
It seems to me that it would be a good idea. I would be disappointed in a future where AI systems aren't broadly deployed and everybody doesn't have access to them. At the same time, I want to better understand the mitigations. If the mitigation is the fine-tuning, the whole thing about open weights is that you can then remove the fine-tuning, which is often superficial on top of these capabilities. If it's like talking on Slack with a biology researcher... I think models are very far from this. Right now, they're like Google search. But if I can show them my Petri dish and they can explain why my smallpox sample didn't grow

and what to change, how do you mitigate that? Because somebody can just fine-tune that in there, right?

Mark Zuckerberg
That's true. I think a lot of people will basically use the off-the-shelf model and some people who have basically bad faith are going to try to strip out all the bad stuff. So I do think that's an issue. On the flip side, one of the reasons why I'm philosophically so pro open source is that I do think that a concentration of AI in the future has the potential to be as dangerous as it being widespread. I think a lot of people think about the questions of "if we can do this stuff, is it bad for it to be out in the wild and just widely available?" I think another version of this is that it's probably also pretty bad for one institution to have an AI that is way more powerful than everyone else's AI.

There's one security analogy that I think of. There are so many security holes in so many different things. If you could travel back in time a year or two years, let's say you just have one or two years more knowledge of the security holes. You can pretty much hack into any system. That's not AI. So it's not that far-fetched to believe that a very intelligent AI probably would be able to identify some holes and basically be like a human who could go back in time a year or two and compromise all these systems.

So how have we dealt with that as a society? One big part is open source software that makes it so that when improvements are made to the software, it doesn't just get stuck in one company's products but can be broadly deployed to a lot of different systems, whether they're banks or hospitals or government stuff. As the software gets hardened, which happens because more people can see it and more people can bang on it, there are standards on how this stuff works. The world can get upgraded together pretty quickly.

I think that a world where AI is very widely deployed, in a way where it's gotten hardened progressively over time, is one where all the different systems will be in check in a way. That seems fundamentally more healthy to me than one where this is more concentrated. So there are risks on all sides, but I think that's a risk that I don't hear people talking about quite as much. There's the risk of the AI system doing something bad. But I stay up at night worrying more about an untrustworthy actor having the super strong AI, whether it's an adversarial government or an untrustworthy company or whatever. I think that that's potentially a much bigger risk.

Dwarkesh Patel
As in, they could overthrow our government because they have a weapon that nobody else has?

Mark Zuckerberg
Or just cause a lot of mayhem. I think the intuition is that this stuff ends up being pretty important and valuable for both economic and security reasons and other things. If someone whom you don't trust or an adversary gets something more powerful, then I think that that could be an issue. Probably the best way to mitigate that is to have good open source AI that becomes the standard and in a lot of ways can become the leader. It just ensures that it's a much more even and balanced playing field.

Dwarkesh Patel
That seems plausible to me. If that works out, that would be the future I prefer. I want to understand mechanistically how the fact that there are open source AI systems in the world prevents somebody causing mayhem with their AI system? With the specific example of somebody coming with a bioweapon, is it just that we'll do a bunch of R&D in the rest of the world to figure out vaccines really fast? What's happening?

Mark Zuckerberg
If you take the security one that I was talking about, I think someone with a weaker AI trying to hack into a system that is protected by a stronger AI will succeed less. In terms of software security –

Dwarkesh Patel
How do we know everything in the world is like that? What if bioweapons aren't like that?

Mark Zuckerberg
I mean, I don't know that everything in the world is like that. Bioweapons are one of the areas where the people who are most worried about this stuff are focused and I think it makes a lot of sense. There are certain mitigations. You can try to not train certain knowledge into the model. There are different things but at some level if you get a sufficiently bad actor, and you don't have other AI that can balance them and understand what the threats are, then that could be a risk. That's one of the things that we need to watch out for.

Dwarkesh Patel
Is there something you could see in the deployment of these systems where you're training Llama-4 and it lied to you because it thought you weren't noticing or something and you're like "whoa what's going on here?" This is probably not likely with a Llama-4 type system, but is there something you can imagine like that where you'd be really concerned about deceptiveness and billions of copies of this being out in the wild?

Mark Zuckerberg
I mean right now we see a lot of hallucinations. It's more so that. I think it's an interesting question, how you would tell the difference between hallucination and deception. There are

a lot of risks and things to think about. I try, in running our company at least, to balance these longer-term theoretical risks with what I actually think are quite real risks that exist today. So when you talk about deception, the form of that that I worry about most is people using this to generate misinformation and then pump that through our networks or others. The way that we've combated this type of harmful content is by building AI systems that are smarter than the adversarial ones.

This informs part of my theory on this. If you look at the different types of harm that people do or try to do through social networks, there are ones that are not very adversarial. For example, hate speech is not super adversarial in the sense that people aren't getting better at being racist. That's one where I think the AIs are generally getting way more sophisticated faster than people are at those issues. And we have issues both ways. People do bad things, whether they're trying to incite violence or something, but we also have a lot of false positives where we basically censor stuff that we shouldn't. I think that understandably makes a lot of people annoyed. So I think having an AI that gets increasingly precise on that is going to be good over time.

But let me give you another example: nation states trying to interfere in elections. That's an example where they absolutely have cutting edge technology and absolutely get better each year. So we block some technique, they learn what we did and come at us with a different technique. It's not like a person trying to say mean things, They have a goal. They're sophisticated. They have a lot of technology. In those cases, I still think about the ability to have our AI systems grow in sophistication at a faster rate than theirs do. It's an arms race but I think we're at least winning that arms race currently. This is a lot of the stuff that I spend time thinking about.

Yes, whether it's Llama-4 or Llama-6, we need to think about what behaviors we're observing and it's not just us. Part of the reason why you make this open source is that there are a lot of other people who study this too. So we want to see what other people are observing, what we're observing, what we can mitigate, and then we'll make our assessment on whether we can make it open source. For the foreseeable future I'm optimistic we will be able to. In the near term, I don't want to take our eye off the ball in terms of what are actual bad things that people are trying to use the models for today. Even if they're not existential, there are pretty bad day-to-day harms that we're familiar with in running our services. That's actually a lot of what we have to spend our time on as well.

Dwarkesh Patel
I found the synthetic data thing really curious. With current models it makes sense why there might be an asymptote with just doing the synthetic data again and again. But let's say they get smarter and you use the kinds of techniques—you talk about in the paper or the blog posts that are coming out on the day this will be released—where it goes to the thought chain that is the most correct. Why do you think this wouldn't lead to a loop where it gets

smarter, makes better output, gets smarter and so forth. Of course it wouldn't be overnight, but over many months or years of training potentially with a smarter model.

Mark Zuckerberg
I think it could, within the parameters of whatever the model architecture is. It's just that with today's 8B parameter models, I don't think you're going to get to be as good as the state-of-the-art multi-hundred billion parameter models that are incorporating new research into the architecture itself.

Dwarkesh Patel
But those will be open source as well, right?

Mark Zuckerberg
Well yeah, subject to all the questions that we just talked about but yes. We would hope that that'll be the case. But I think that at each point, when you're building software there's a ton of stuff that you can do with software but then at some level you're constrained by the chips that it's running on. So there are always going to be different physical constraints. How big the models are is going to be constrained by how much energy you can get and use for inference. I'm simultaneously very optimistic that this stuff will continue to improve quickly and also a little more measured than I think some people are about it. I don't think the runaway case is a particularly likely one.

Dwarkesh Patel
I think it makes sense to keep your options open. There's so much we don't know. There's a case in which it's really important to keep the balance of power so nobody becomes a totalitarian dictator. There's a case in which you don't want to open source the architecture because China can use it to catch up to America's AIs and there is an intelligence explosion and they win that. A lot of things seem possible. Keeping your options open considering all of them seems reasonable.

Mark Zuckerberg
Yeah.

Dwarkesh Patel
Let's talk about some other things. Metaverse. What time period in human history would you be most interested in going into? 100,000 BCE to now, you just want to see what it was like?

Mark Zuckerberg
It has to be the past?

Dwarkesh Patel
Oh yeah, it has to be the past.

Mark Zuckerberg
I'm really interested in American history and classical history. I'm really interested in the history of science too. I actually think seeing and trying to understand more about how some of the big advances came about would be interesting. All we have are somewhat limited writings about some of that stuff. I'm not sure the metaverse is going to let you do that because it's going to be hard to go back in time for things that we don't have records of. I'm actually not sure that going back in time is going to be that important of a thing. I think it's going to be cool for like history classes and stuff, but that's probably not the use case that I'm most excited about for the metaverse overall.

The main thing is just the ability to feel present with people, no matter where you are. I think that's going to be killer. In the AI conversation that we were having, so much of it is about physical constraints that underlie all of this. I think one lesson of technology is that you want to move things from the physical constraint realm into software as much as possible because software is so much easier to build and evolve. You can democratize it more because not everyone is going to have a data center but a lot of people can write code and take open source code and modify it. The metaverse version of this is enabling realistic digital presence. That's going to be an absolutely huge difference so people don't feel like they have to be physically together for as many things. Now I think that there can be things that are better about being physically together. These things aren't binary. It's not going to be like "okay, now you don't need to do that anymore." But overall, I think it's just going to be really powerful for socializing, for feeling connected with people, for working, for parts of industry, for medicine, for so many things.

Dwarkesh Patel
I want to go back to something you said at the beginning of the conversation. You didn't sell the company for a billion dollars. And with the metaverse, you knew you were going to do this even though the market was hammering you for it. I'm curious. What is the source of that edge? You said "oh, values, I have this intuition," but everybody says that. If you had to say something that's specific to you, how would you express what that is? Why were you so convinced about the metaverse?

Mark Zuckerberg
I think that those are different questions. What are the things that power me? We've talked about a bunch of the themes. I just really like building things. I specifically like building things around how people communicate and understanding how people express themselves and how people work. When I was in college I studied computer science and psychology. I think a lot of other people in the industry studied computer science. So, it's always been the intersection of those two things for me.

It's also sort of this really deep drive. I don't know how to explain it but I just feel constitutionally that I'm doing something wrong if I'm not building something new. Even when we were putting together the business case for investing a $100 billion in AI or some huge amount in the metaverse, we have plans that I think made it pretty clear that if our stuff works, it'll be a good investment. But you can't know for certain from the outset. There are all these arguments that people have, with advisors or different folks. It's like, "how are you confident enough to do this?" Well the day I stop trying to build new things, I'm just done. I'm going to go build new things somewhere else. I'm fundamentally incapable of running something, or in my own life, and not trying to build new things that I think are interesting. That's not even a question for me, whether we're going to take a swing at building the next thing. I'm just incapable of not doing that. I don't know.

I'm kind of like this in all the different aspects of my life. Our family built this ranch in Kauai and I worked on designing all these buildings. We started raising cattle and I'm like "alright, I want to make the best cattle in the world so how do we architect this so that way we can figure this out and build all the stuff up that we need to try to do that." I don't know, that's me. What was the other part of the question?

Dwarkesh Patel
I'm not sure but I'm actually curious about something else. So a 19-year-old Mark reads a bunch of antiquity and classics in high school and college. What important lesson did you learn from it? Not just interesting things you found, but there aren't that many tokens you consume by the time you're 19. A bunch of them were about the classics. Clearly that was important in some way.

Mark Zuckerberg
There aren't that many tokens you consume... That's a good question. Here's one of the things I thought was really fascinating. Augustus became emperor and he was trying to establish peace. There was no real conception of peace at the time. The people's understanding of peace was peace as the temporary time between when your enemies inevitably attack you. So you get a short rest. He had this view of changing the economy from being something mercenary and militaristic to this actually positive-sum thing. It was a very novel idea at the time.

That's something that's really fundamental: the bounds on what people can conceive of at the time as rational ways to work. This applies to both the metaverse and the AI stuff. A lot of investors, and other people, can't wrap their head around why we would open source this. It's like "I don't understand, it's open source. That must just be the temporary time between which you're making things proprietary, right?" I think it's this very profound thing in tech that it actually creates a lot of winners.

I don't want to strain the analogy too much but I do think that a lot of the time, there are models for building things that people often can't even wrap their head around. They can't understand how that would be a valuable thing for people to do or how it would be a reasonable state of the world. I think there are more reasonable things than people think.

Dwarkesh Patel
That's super fascinating. Can I give you what I was thinking in terms of what you might have gotten from it? This is probably totally off, but I think it's just how young some of these people are, who have very important roles in the empire. For example, Caesar Augustus, by the time he's 19, is already one of the most important people in Roman politics. He's leading battles and forming the Second Triumvirate. I wonder if the 19-year-old you was thinking "I can do this because Caesar Augustus did this."

Mark Zuckerberg
That's an interesting example, both from a lot of history and American history too. One of my favorite quotes is this Picasso quote that all children are artists and the challenge is to remain an artist as you grow up. When you're younger, it's just easier to have wild ideas. There are all these analogies to the innovator's dilemma that exist in your life as well as for your company or whatever you've built. You're earlier on in your trajectory so it's easier to pivot and take in new ideas without disrupting other commitments to different things. I think that's an interesting part of running a company. How do you stay dynamic?

Dwarkesh Patel
Let's go back to the investors and open source. The $10B model, suppose it's totally safe. You've done these evaluations and unlike in this case the evaluators can also fine-tune the model, which hopefully will be the case in future models. Would you open source the $10 billion model?

Mark Zuckerberg
As long as it's helping us then yeah.

Dwarkesh Patel
But would it? $10 billion of R&D and now it's open source.

Mark Zuckerberg
That's a question which we'll have to evaluate as time goes on too. We have a long history of open sourcing software. We don't tend to open source our product. We don't take the code for Instagram and make it open source. We take a lot of the low-level infrastructure and we make that open source. Probably the biggest one in our history was our Open Compute Project where we took the designs for all of our servers, network switches, and data centers, and made it open source and it ended up being super helpful. Although a lot of people can design servers the industry now standardized on our design, which meant that

the supply chains basically all got built out around our design. So volumes went up, it got cheaper for everyone, and it saved us billions of dollars which was awesome.

So there's multiple ways where open source could be helpful for us. One is if people figure out how to run the models more cheaply. We're going to be spending tens, or a hundred billion dollars or more over time on all this stuff. So if we can do that 10% more efficiently, we're saving billions or tens of billions of dollars. That's probably worth a lot by itself. Especially if there are other competitive models out there, it's not like our thing is giving away some kind of crazy advantage.

Dwarkesh Patel
So is your view that the training will be commodified?

Mark Zuckerberg
I think there's a bunch of ways that this could play out and that's one. So "commodity" implies that it's going to get very cheap because there are lots of options. The other direction that this could go in is qualitative improvements. You mentioned fine-tuning. Right now it's pretty limited what you can do with fine-tuning major other models out there. There are some options but generally not for the biggest models. There's being able to do that, different app specific things or use case specific things or building them into specific tool chains. I think that will not only enable more efficient development, but it could enable qualitatively different things.

Here's one analogy on this. One thing that I think generally sucks about the mobile ecosystem is that you have these two gatekeeper companies, Apple and Google, that can tell you what you're allowed to build. There's the economic version of that which is like when we build something and they just take a bunch of your money. But then there's the qualitative version, which is actually what upsets me more. There's a bunch of times when we've launched or wanted to launch features and Apple's just like "nope, you're not launching that." That sucks, right? So the question is, are we set up for a world like that with AI? You're going to get a handful of companies that run these closed models that are going to be in control of the APIs and therefore able to tell you what you can build?

For us I can say it is worth it to go build a model ourselves to make sure that we're not in that position. I don't want any of those other companies telling us what we can build. From an open source perspective, I think a lot of developers don't want those companies telling them what they can build either. So the question is, what is the ecosystem that gets built out around that? What are interesting new things? How much does that improve our products? I think there are lots of cases where if this ends up being like our databases or caching systems or architecture, we'll get valuable contributions from the community that will make our stuff better. Our app specific work that we do will then still be so

differentiated that it won't really matter. We'll be able to do what we do. We'll benefit and all the systems, ours and the communities', will be better because it's open source.

There is one world where maybe that's not the case. Maybe the model ends up being more of the product itself. I think it's a trickier economic calculation then, whether you open source that. You are commoditizing yourself then a lot. But from what I can see so far, it doesn't seem like we're in that zone.

Dwarkesh Patel
Do you expect to earn significant revenue from licensing your model to the cloud providers? So they have to pay you a fee to actually serve the model.

Mark Zuckerberg
We want to have an arrangement like that but I don't know how significant it'll be. This is basically our license for Llama. In a lot of ways it's a very permissive open source license, except that we have a limit for the largest companies using it. This is why we put that limit in. We're not trying to prevent them from using it. We just want them to come talk to us if they're going to just basically take what we built and resell it and make money off of it. If you're like Microsoft Azure or Amazon, if you're going to be reselling the model then we should have some revenue share on that. So just come talk to us before you go do that. That's how that's played out.

So for Llama-2, we just have deals with basically all these major cloud companies and Llama-2 is available as a hosted service on all those clouds. I assume that as we release bigger and bigger models, that will become a bigger thing. It's not the main thing that we're doing, but I think if those companies are going to be selling our models it just makes sense that we should share the upside of that somehow.

Dwarkesh Patel
Regarding other open source dangers, I think you have genuine legitimate points about the balance of power stuff and potentially the harms you can get rid of because we have better alignment techniques or something. I wish there were some sort of framework that Meta had. Other labs have this where they say "if we see this concrete thing, then that's a no go on the open source or even potentially on deployment." Just writing it down so the company is ready for it and people have expectations around it and so forth.

Mark Zuckerberg
That's a fair point on the existential risk side. Right now we focus more on the types of risks that we see today, which are more of these content risks. We don't want the model to be doing things that are helping people commit violence or fraud or just harming people in different ways. While it is maybe more intellectually interesting to talk about the existential risks, I actually think the real harms that need more energy in being mitigated are things

where someone takes a model and does something to hurt a person. In practice for the current models, and I would guess the next generation and maybe even the generation after that, those are the types of more mundane harms that we see today, people committing fraud against each other or things like that. I just don't want to shortchange that. I think we have a responsibility to make sure we do a good job on that.

Dwarkesh Patel
Meta's a big company. You can handle both.
As far as open source goes, I'm actually curious if you think the impact of open source, from PyTorch, React, Open Compute and other things, has been bigger for the world than even the social media aspects of Meta. I've talked to people who use these services and they think that it's plausible because a big part of the internet runs on these things.

Mark Zuckerberg
It's an interesting question. I mean almost half the world uses our consumer products so it's hard to beat that. But I think open source is really powerful as a new way of building things. I mean, it's possible. It may be one of these things like Bell Labs, where they were working on the transistor because they wanted to enable long-distance calling. They did and it ended up being really profitable for them that they were able to enable long-distance calling. 5 to 10 years out from that, if you asked them what was the most useful thing that they invented it's like "okay, we enabled long distance calling and now all these people are long-distance calling." But if you asked a hundred years later maybe it's a different answer.

I think that's true of a lot of the things that we're building: Reality Labs, some of the AI stuff, some of the open source stuff. The specific products evolve, and to some degree come and go, but the advances for humanity persist and that's a cool part of what we all get to do.

Dwarkesh Patel
By when will the Llama models be trained on your own custom silicon?

Mark Zuckerberg
Soon, not Llama-4. The approach that we took is we first built custom silicon that could handle inference for our ranking and recommendation type stuff, so Reels, News Feed ads, etc. That was consuming a lot of GPUs. When we were able to move that to our own silicon, we're now able to use the more expensive NVIDIA GPUs only for training. At some point we will hopefully have silicon ourselves that we can be using for at first training some of the simpler things, then eventually training these really large models. In the meantime, I'd say the program is going quite well and we're just rolling it out methodically and we have a long-term roadmap for it.

Dwarkesh Patel
Final question. This is totally out of left field. If you were made CEO of Google+ could you have made it work?

Mark Zuckerberg
Google+? Oof. I don't know. That's a very difficult counterfactual.

Dwarkesh Patel
Okay, then the real final question will be: when Gemini was launched, was there any chance that somebody in the office uttered: “Carthago delenda est”.

Mark Zuckerberg
No, I think we're tamer now. It's a good question. The problem is there was no CEO of Google+. It was just a division within a company. You asked before about what are the scarcest commodities but you asked about it in terms of dollars. I actually think for most companies, of this scale at least, it's focus. When you're a startup maybe you're more constrained on capital. You’re just working on one idea and you might not have all the resources. You cross some threshold at some point with the nature of what you're doing. You're building multiple things. You're creating more value across them but you become more constrained on what you can direct to go well.

There are always the cases where something random awesome happens in the organization and I don't even know about it. Those are great. But I think in general, the organization's capacity is largely limited by what the CEO and the management team are able to oversee and manage. That's been a big focus for us. As Ben Horowitz says “keep the main thing, the main thing” and try to stay focused on your key priorities.

Dwarkesh Patel
Awesome, that was excellent, Mark. Thanks so much. That was a lot of fun.

Mark Zuckerberg
Yeah, really fun. Thanks for having me.

Dwarkesh Patel
Absolutely.