An Interview with Daniel Gross and Nat Friedman About Apple and AI

Thursday, June 13, 2024

Good morning,

I am happy to welcome Daniel Gross and Nat Friedman back for a Stratechery Interview, the seventh in our ongoing series (we previously talked in October 2022, December 2022, March 2023, August 2023, December 2023, and February 2024).

This series is somewhat unique for Stratechery in that my interview subjects are generally not investors; this series, though, started in October 2022, when Friedman and Gross were launching a grant program, precisely because we shared a sense that there wasn't nearly enough activity or discussion around AI; one month later ChatGPT came out and the world has changed dramatically, including for Friedman and Gross, who are two of the leading investors in the space. To that end, I think it has been very valuable — and popular — to continue this series, but do keep in mind that Friedman and Gross may be invested in some of the companies we discuss.

This week marks the end of developer conference season, and Apple is poised to be a big winner. We discuss what makes Apple uniquely capable of pivoting, and the power that comes from building compelling products. Then, we explore the status of the other major players, including OpenAI, Microsoft, and Google; discuss whether or not we are in a bubble, and if so, why things are different than the dot com era; and finally, as is our wont, we get philosophical and discuss why thinking about AI in human terms may be a mistake.

As a reminder, all Stratechery content, including interviews, is available as a podcast; click the link at the top of this email to add Stratechery to your podcast player.

On to the Interview:

An Interview with Daniel Gross and Nat Friedman About Apple and AI

This interview is lightly edited for clarity.

Topics:
Apple | OpenAI | The Bubble Question | What's Next

Apple

Nat and Daniel, welcome back to Stratechery.

DG: Thank you for having us, Ben.

NF: Great to be back.

As always, it feels like it has been forever since you have been on and once again, I had to look it up, it's only been a few months. This seemed like a good time to catch up though, given we have now reached the end of developer conference season for the biggest players, so are you ready to jump in?

NF: Let's do it.

Let's start with the current belle of the ball, Apple. Apparently we have a new obvious winner from AI. In case you're keeping track, I think Google was the obvious winner, then OpenAI was the obvious winner, then Microsoft, then Google again, then everyone just decided screw it, just buy Nvidia — I think that one still holds actually — and now we are to Apple, which by the way does not seem to be using Nvidia. Here's a meta question: has anything changed in the broader environment where we can say with any sort of confidence, who is best placed and why, or is this just sort of the general meta, particularly in media and analysts like myself, running around like chickens with their heads cut off?

NF: I think one thing that really plays to Apple's favor is that there seems to be multiple players reaching the same level of capabilities. If OpenAI had clearly broken away, such that they were 10 times better or even 2 times better than everyone else in terms of model quality, that would put Apple in a more difficult position. Apple benefits from the idea that either they can catch up or they have their choice of multiple players that they can work with, and it looks like we have somewhere between three and five companies that are all in it to win it and most of whom are planning to offer their models via APIs.

You have Google, OpenAI, Anthropic, you have X, you have Meta and so if you're on the side of application building, generally this is great news because prices are going to keep dropping 90% per year, capabilities are going to keep improving. None of those players will have pricing power and you get to pick, or in Apple's case, you can pick for now and have time to catch up in your own first party capabilities. The fact that no one's broken away or shown a dominant lead, at least in this moment, between major model releases. We haven't seen ChatGPT-5 yet, we haven't seen Q* yet. Yeah, on current evidence, I think that's good for people who are great at products, focus on products and applications and have massive distribution.

Which is Apple.

NF: Which is Apple.

What do you think, Daniel? You, I think, have consistently in these interviews been pretty positive about Apple's potential end position. Do you feel validated? Do you feel even stronger or are you going to take the chance to zag here after leading the charge for a while?

DG: Yeah, you've invited us here a couple of times, and I do think I've been saying for quite some time that Apple has a pretty strong hand to play. Everything that Nat said, and I'd add on top of it, I think they're the only company in the world that can book out more of TSMC than Nvidia can. If you believe that you actually want a high volume of chips and you solve all your packaging problems and whatnot, and it's just a question of how much intelligence you can buy from the intelligence well that is in Taiwan, they have the ability to do it first. I think that's why Nvidia, defensively, pre-purchased a bunch of capacity, I believe that was in 2022. You can see those dynamics at play and you see that Nvidia understands that.

Yeah, I think Apple has always had the ingredients, and I've always said has the ingredients to be a dominant winner in the space, and now we're seeing that. To me the interesting thing that I got out of the keynote is, they're painting the picture today that high-end language models are a kind of search-like thing and they're going to have partnerships with different companies and you have two excellent posts on this topic, so I won't bore your listeners with that, but if that is how this goes, then I think they'll probably dominate.

If you end up in a dynamic where, for whatever reason the frontier capabilities translate into a disruptive innovation that allows you to capture customers, then maybe that's an issue, but to date that has just not been the story.

Yeah, I mean I was writing today, I wrote about Apple three times this week, but the latest one was I perceive there being two risk factors for Apple. One is what you just said, which is one of these models actually figures it out to such a great extent that Apple becomes the commodity hardware provider providing access to this model. They'll have a business there, but not nearly as a profitable one as they're setting up right now where the models are the commodity, that's risk factor number one.

Risk factor number two is, can they actually execute on what they showed? Can this on-device inference work as well as they claim? Will using their own silicon, and I think it's probably going to be relatively inefficient, but given their scale and the way that they can architect it, they can probably pull it off having this one-to-one connection to the cloud. If they can do it, that's great, but maybe they can't do it. They're doing a lot of new interesting stuff in that regard. Of those two risk factors, which do you think is the more important one?

DG: I don't fully understand and I never fully have understood why local models can't get really, really good, and I think that the reason often people don't like hearing that is there's not enough epistemic humility around how simple most of what we do is, from a caloric energy perspective, and why you couldn't have a local model that does a lot of that. A human, I think, at rest is consuming like 100 watts maybe and an iPhone is using, I don't know, 10 watts, but your MacBook is probably using 80 watts. Anyway, it's within achievable confines to create something that has whatever the human level ability is, it's synthesizing information on a local model.

What I don't really know how to think about is what that means for the broader AI market, because at least as of now we obviously don't fully believe that. We're building all of this complicated data center capacity and we're doing a lot of things in the cloud which is in cognitive dissonance with this idea that local models can get really good. The economy is built around the intelligence of the mean, not the median. Most of the labor is being done that is fairly simple tasks, and I've yet to see any kind of mathematical refutation that local models can't get really good. You still may want cloud models for a bunch of other reasons, and there's still a lot of very high-end, high-complexity work that you're going to want a cloud model for, chemistry, physics, biology, maybe even doing your tax return, but for basic stuff like knowing how to use your iPhone and summarizing web results, I basically don't understand why local models can't get really good.

The other thing I'd add in by the way that's going to happen for free is there's going to be a ton of work both on the node density side from TSMC, but also on the efficiency side from every single major AI lab, because even though they run their models in the cloud, or because they run their models in the cloud, they really care about their COGS. You have this process that's happened pretty durably year-over-year, where a new frontier model is launched, it's super expensive to run and then it's distilled, quantized or compressed so that the COGS of that company are more efficient. Now if you continue to do that, yeah, you do sort of wonder, wait a minute, "Why can't the consumer run this model?". There's a ton of economic pressure to make these models not just very smart, but very cheap to run. At the limit, I don't know if it's going to be like your Apple TV, sort of computer at home is doing the work, or literally it's happening in your hands, but it feels like local models can become pretty powerful.

NF: Yeah. Six months ago, Andrej Karpathy published this vision of the LLM OS, and the idea was that the language model, in a way it’s a new kind of computer, it’s a new kind of operating system, and what it’s going to have connected to it are peripherals and tools that it can use through function calling, and I thought that was a really fun vision.

TBD how it plays out, but I thought Apple’s announcement really supported his view. What Apple’s done in fact is they’ve got a little LLM kernel on the device that’s listening to your requests and figuring out what to do with them, and it can decide to try to handle them itself. It can decide to invoke parts of apps locally and it can decide to dispatch parts of the work or all of the work to either Apple’s models in their cloud, or now it can invoke — after getting your approval — ChatGPT.

What I think Apple’s done with this architecture is they’ve provided themselves both kind of a hedge on the quality of local models and a ramp that they can use. They can as local models improve, they can handle more requests locally as appropriate, and as their own models running on their chips in their cloud improve, they can use those, and then to the extent that it makes sense for them, they could use third party models. I think Daniel’s probably right that local will improve dramatically, it doesn’t have to be able to do everything for Apple’s strategy to work though, they can smoothly dispatch. They have a little router on the phone and I think that’s Apple’s dream is to have a 2B, a 3B model that runs on your phone, that’s primarily a kind of tool-use model. It basically does function calling.

The most important agent in AI is going to be the local agent that decides where to dispatch jobs. It doesn’t need to be big, it doesn’t need to be complex, but it is at the linchpin and it will control all the value.

NF: Yeah. I think in the short term there’s good reasons to use remote models even for only moderate things. The cost of having even a 3B model in memory on a phone in terms of energy is not trivial, the cost of loading it into memory is high. The amount of compute that you have to deploy now, but as Daniel says, Moore’s Law or Jensen’s Law will continue or TSMC’s Law and quantization and distilling and other techniques will continue to improve and so local models will just get better and Apple is, I think, not entirely, but somewhat indifferent to this over time. They can bet on this and it can be a smooth transition from remote to local as things get better.

Daniel, I’m actually curious — you used to be the head of Machine Learning at Apple. It has been a few years, so not to say you have necessarily an intimate view of what’s happened over the last year, but to me the fact that Apple Intelligence will only run on the iPhone 15 Pro strikes me as confirmation that Apple was indeed late to this, because I think if they could go back and do it over again, at least the iPhone 15 would’ve had 8GB of ram, at least bring in the whole last generation. Sure, they want to sell new hardware, but they also don’t want people to feel screwed either.

That speaks to me that yeah, this was a sort of, “All hands on deck, we have to figure this out”, situation and it’s mostly all positive. Positive number one is, I think it’s safe to assume that their chips are actually not yet properly designed to support this use case. We can expect not just that TSMC’s process improves, but that Apple design and support, when they customize Apple silicon for that local model, it’s going to get that much more powerful and that much more efficient. Then number two, my assumption is this Apple Private Cloud is just M2 Ultras and I think that probably governs the size of the model in those clouds, but what happens when they do actually design and I think it’s safe to say they will now design their own server chips. The whole Nuvia team is like, “We were telling you a long time ago!”

But the other third bullish sign is if all this is true, they seem to deliver a really compelling vision that is substantiated sufficiently — a lot of this is in beta is going to roll it over the year — in a year which kind of calls back to the old Apple, the Steve Jobs, “We’re going to do iMovie, oh shoot, it’s actually music, we need to ship iTunes and we’re

going to ship an iPod in six months". Is that a fair characterization? What's your perception of what went on internally over the last 18 months?

DG: Just to be very specific, and this will be relevant in a minute, when I was there, I was the DRI, a Directly Responsible Individual, of an annual recurring project called OS Intelligence, which encompassed all the different Machine Learning and AI efforts across the company, and I was a tent pole. A tent pole is the Apple's way of declaring this is one of six or seven things we're really going to care about in a given year.

When I saw Apple Intelligence on the keynote, obviously it's a funny acronym, but I was seeing the internal org reflected externally, that was a tent pole for the year, and even within tent poles, the company has a sense of internal hierarchy of tent poles. Really the question is, when you work at Apple, what is the classification code for a tent pole? Are you a P0, P1, P2, P3? You want to be P0 or P1. P0 would be something like new hardware support, which is the company's going to crater if this doesn't happen, the parts are coming in from China and you have to work with the new quad-band LTE or whatever.

If I had to guess, the organizational expression of what you're saying is it got moved from a P1, which is where I used to sit, to a P0. I don't actually know this to be true, but that's what it felt to me.

Well, it felt like the rest of WWDC was just like, "No one else worked on anything all year". It almost was useful to have all the other stuff in the first hour, to really emphasize that, "No, we haven't been working on anything other than Apple Intelligence".

DG: I think Apple is a pretty flexible company. The first year, or the first couple of months I worked at Apple, we worked on a tent pole, and we were punted close to the presentation, because what we had built was not good enough and it was an incredibly painful career moment for me. [Apple SVP of Engineering] Craig (Federighi) came down to my office to handhold me a little bit. I know that it was painful moment because I remember every second of that day, but they do those things on the positive and on the negative very flexibly throughout the year. If the stuff isn't good enough, they'll punt you for the next year, and if it seems important, they will reorganize the company pretty rapidly.

I think in part this is Steve's genius with the DRI model, which is not particularly tied to the org chart. For example, I was principally sitting in [Apple SVP of Services] Eddy Cue's org, but I was commanding people on completely different teams in Craig's org because I was a DRI. If I had to guess, yeah, they did one of those P0 sprints, and it sort of worked out, and there's a lot of things that are cobbled together, including, I don't know anything, but I'm assuming that your theories with the chips are correct.

I also think if you think of it from Apple's perspective, having the same kind or a similar processor in the cloud as on the client is somewhat sensible. Meaning, you could envision a protocol where the client is trying to stream tokens and because it has the same architecture and model as the cloud, literally down to the bits and down to the silicon, it can very flexibly fall back, generate a little bit more locally, generate a little bit more from the cloud and even as I talk to you now, some words are easier to come to than others. You could imagine, for example, the simplest way I think of it in my head is if you have a particular math problem you're working on, writing out the math problem is actually fairly low complexity compute, and then you have the equal sign, then suddenly you have these high complexity tokens. You could imagine them making this kind of fairly flexible protocol and it'd be helpful that you'll have very reliable performance standards and characteristics between the local and server model.

Obviously from Apple's perspective, from an economic perspective, it's highly aligned with their business model, and it also may help them be frontier at what they do, even if they don't have enough data center energy capacity and whatnot.

I do think just the one thing I'd add, just to close the loop on the earlier comment, because I think you could walk away from that keynote, project the future a little bit and wonder then what the heck are we all doing and why has Vertiv gone up 4X this year and Nvidia is where it is. Sometimes I really try to wonder what the equivalent is of looking at the Telco bubble and imagining social networking and I think that stuff is really hard to do. It is really easy for us to look at the Telco bubble and imagine pets.com or Webvan, it's quite hard to imagine social networking.

In AI, I think a similar thing you could imagine is that the easy leap is, "Oh, just do all the basic current economy as much as you can", and then the question is, "What's the harder to imagine social networking like thing?", and I feel like it might be a kind of rediscovery of hard science and hard physics. Human progress today is bottlenecked in those fields purely because the rate of intelligence. You could imagine, how does the AI market unfold? Actually a lot of the local economy or a lot of the economy can be done locally, but then your real miracles of new fluid dynamics, a new Ozempic, new kinds of energy, that is a massive industry, and maybe that gets done in the cloud and that's sort of what we missed, that was our horseless carriage. We were so focused on tool-use and intermittent browsing and we didn't realize, "Oh, there's brand new kinds of science that are super lucrative and profitable", the new Eli Lilly is the equivalent maybe of the new social networking. Anyway, that's just my thought at least today.

I completely agree. Actually, one of my takes has been, I think this is sort of drawing on the Internet era, is, I kind of suspect there's an inverse correlation between the profoundness of an innovation and the time it takes for it to actually make a difference. The Internet was, it took 20 years to your point, for the feed to come along, or 15 years, wherever it might be, which is the feed to my mind is the core Internet sort of innovation. That was something that could not be done before to have a dynamically created list of content that never ends, that is personalized to every single individual, that is fundamentally new. It took 15 years to get to it, and that's what unlocked the advertising model, it unlocked the entire economy of the Internet. That's what changed our politics, changed society. It took a really long time, even though in retrospect it's very obvious — I'm making your social media point in a different way. But it seems like, what were we doing in the intervening 15 years? We were putting articles on the web and slapping an ad next to it, just like we did in newspapers, and saying, "Wow, the Internet doesn't make any money". For AI, it feels like to me, we started this podcast 18 months ago talking about like, "Wow, no one's building products here," and it almost feels like that's actually still the case.

NF: Yes.

People are certainly building products, but no one has actually figured out what is the product that no one could imagine.

NF: Yeah, I think I was going to say this about Apple in particular. There's this big debate in the AI field about what are the rate-limiters on progress, and the scaling purists think we need more scale, more compute. There are people who believe we need algorithmic breakthroughs, and so we are AI researcher limited, and then there are folks who believe we're hitting a data wall, and what we're actually gated on is high-quality data and maybe labeled data, maybe raw data, maybe video can provide it.

But at least in terms of felt progress, I think it is UI and products. There's still a massive capability overhang where we are still learning how to make these models useful to people. It's really shocking to me how little has happened.

I guess, to your feed point, this is just something that takes time. There's a distributed search that occurs across the industry for things that work. When we made GitHub Copilot, it was primarily — maybe not primarily — but there was a very large part of the problem solving that was like, "What's the UI for this? What's the tolerable latency? What does it look like? How does it know what to do? And how do you make it so that when it makes errors, those are tolerable to you?", because it's going to make some errors.

Yep.

NF: So what I loved from the Apple presentation was, we really started to see in a concrete way of vision of the UI, and what they're doing in a way is, they've decomposed their apps into little bits and pieces of useful functionality with totally separated UIs. The vision is clearly that through one means or another, you can have a kind of conversation with your device where you say, "Where's the dinner tonight? How long does it take to get there?"

The demo of the woman wanting to know when her mom's flight landed and where they're going to dinner was one of the best tech demos in years, and I mean that completely sincerely in that it was so simple, and yet it was immediately understandable to everyone what a hard problem this is and how annoying it is to get that information.

NF: Apple in that demo broke through a lot of the kind of technology boundaries, and I think this is something Apple is often so amazing at. They clearly have a set of people who either aren't aware of or are able to ignore the way things currently work and think just in terms of what the user experience should be, almost at a superficial level and I say that as a good thing, not as a bad thing.

I was talking to Andrew, my co-host on Sharp Tech, he's a relevant novice in tech, and he was just so blown away. I'm like, "Look, this is the path of the Apple fanboy." This is why people love Apple, because of this consistent capability to approach problems in this way.

NF: Yeah. So the idea that they can say, "Forget apps, they're not these separate things you switch between, but they're going to provide these intents or these little functions or capabilities, they'll have bits of UI that appear in the conversation at the right time", and the magic of that demo is you can say, "How long does it take to get there?", and it knows what "there" means and it has the context. So I got really excited by that, it's amazing actually. It sounds so simple, it's obvious in retrospect, but I do think they pointed the way to the user experience here beyond what anyone else has done so far.

Well, there's a sort of analogy, like Daniel just explained, and I've written about, Apple's functional organization and the flexibility that affords them, where you can have a DRI come in and assemble a team on the fly to do something. That is kind what Apple demonstrated with their approach, which is, "Look, there is all these vertical silos on your phone, which are apps, which is your messaging app and your email app and what you're actually looking for is, Apple Intelligence is your personal DRI, it is going to operate across these different things and pull out bits and pieces", and that is compelling, it's very effective.

NF: Yeah, it's effective. Also little things, I mean, in the Mail app, the little summaries under each mail that have traditionally been the first eight words of the mail are now LLM summaries. It's like, "Oh, shoot, why didn't I think of that? That's so obvious."

The notifications thing was just, "Of course this is the way notifications should work".

DG: There's an aspect of really good product design that is a really good joke or a really good observational humor joke in that it's extremely obvious in hindsight. But this is something Nat has really made the case for, and I think he's correct, it's extremely hard to do a priori. Maybe a magic trick is a better metaphor where just conceiving it and thinking of it is really hard and really thankless work because if it's really good, no one will notice.

NF: You see the thing that nobody else sees that was sort of all around us. So yeah, I think that I like the LLM OS approach, I like the UI and the product vision. When you ask the question, "What's the big risk for them?", I really think it's execution, making this work. We've seen with AI, the demos are easy. We learned that in the first month of working on Copilot back in 2020, but making it work reliably and so the failures feel okay and the successes give you the rewards when they happen is really tough.

One of the things we've seen in LLMs is that it's really data-driven. I think one of the big takeaways that I've taken from the last couple of years is that if you want a capability in your model, there have to be a lot of good examples of that in your training data. When I built natbot three years ago, that was one of the first things I learned. It was a little bot that would use GPT-3 at the time to browse the web and take actions in the web. It kind of worked, it was a neat demo, and then I got early access to GPT-4 and plugged it in, thinking it would improve a lot because GPT-4 was evidently so much better, and it didn't. It improved a little bit, and that's because the browsing and action taking data just wasn't in the data set. So I think for Apple to do a good job at this, they will probably have to be good at data. They'll probably have to be good at collecting lots and lots of really high quality samples of data.

Are they going to be hamstrung? Because I mean, they have access to a lot of really potentially useful data that they are committed to not take advantage of.

NF: Well, you can do it. The state-of-the-art way to do this to get high-quality data, the user data engine I think is slightly overrated in the AI space. I think if you took ChatGPT user data away from OpenAI for training, I think they'd still do very well and they've had a willingness to spend a billion dollars a year on high-quality labeled data.

Other labs are doing this too. I think some people saw Scale's fundraising announcement. Scale is one of the companies that's just really capitalizing on this AI wave because they're helping people collect this high-quality data and they're making well over a billion dollars a year now doing that. So Apple could do it, but it requires thousands of labelers, and it's operationally very intensive and extreme attention to detail. I don't know, is that in their DNA, do you think, Daniel?

DG: It's an interesting question. I think when I was working there, and it seems like this is still the case, there is a real extreme obsession of privacy that is genuine, it's not motivated by a business model, it's really motivated by a sense of spirit from people at the top.

That said, to your point, Nat, the idea of the data moat is, it's not clear that it exists in the realm of LLM just because the blast radius of every additional data point is so large, you can get pretty far with a small amount of very high-quality data. They have been collecting data for years for speech recognition for Siri by paying, and Apple has a lot of money and they know how to pay for stuff, and they love paying for things that would preserve user privacy, I think.

There's this question of, "Will you know how to collect the correct amount of information?", which in itself is a little bit more of an art than a science. And I wonder, on the one hand you might think, "Well, maybe Apple doesn't really get ML, and so maybe it's going to be hard for them to do that even though they have a budget", on the other hand, I think one thing we've discussed here previously is, the theology and philosophy of slaving over data is not too different from slaving over pixels. The non-intuitive thing would be to have the Apple design team take a look, understand ML, and then collect the data. But the company has that psychology of sweating details and I think one of the reasons some of the other labs have struggled to date, say in comparison to a company like Mistral, is they don't have a sweating details culture. They have a scale culture which often these things are quite at odds.

So if Apple is able to channel its mythical like, "We're going to care a lot about the animation on the settings page that nobody will see," they may be actually very good at collecting high-quality data. There's a translational aspect to this, meaning the reverence and the importance of a designer at Apple is so much greater than an AI engineer, so the question is, will the organization appreciate the new phenotype?

Enter the organizational pain that will come from the model screwing up. I think that was the most notable thing about the scores they posted about their models, which again, they're from Apple, I'll take them with a grain of salt, but was the rejected responses. Their accuracy was about even to everything they compared it to, but their hostile or whatever it was categorized as was significantly lower.

DG: Refusal.

Refusal, thank you, and that speaks to they don't want to screw up, and if they think they might screw up, they'll dump to OpenAI and let them put their brand on it.

DG: One question that always existed was, "Wow is Apple TV+ going to be possible?", because the brand is such a premium brand, are they going to be able to create risqué content, even a scene with violence in it? That's not the Apple brand that Steve job envisioned. And look, maybe you think For All Mankind it's a lame TV show, but at the end of the day, they do have TV shows that have gore and violence and sex in them, and they somehow manage that pretty well, I think.

NF: I mean, they're walking into a new level of moderation challenge too, or policy challenge, with a conversational AI. We've seen Gemini screw up and we've seen OpenAI take some high voltage shocks at times. How will Apple navigate that?

Just on that, I think you overstate it. I think one thing that is notable is the extent to which they have not even ventured into that territory; their models are basically not generating any text, it's mostly doing the summarization, change of tone. If you want to do any sort of generation, they're kicking that to OpenAI, which by the way as a side note, I think makes it more likely they can get Apple Intelligence into a place like China where they're not very friendly, but they can go in and credibly say, "We're not creating stuff".

Where they are creating stuff is the image generator, but it only does three styles, it won't do photorealistic, they said last night at [John] Gruber's Talk Show, "Yeah, we don't want to be in deep fakes". We've heard through the grapevine that politicians care about deep fakes, everything else is not sort of a big deal, but I think there is a very explicit choice here to actually do very little creation, which is interesting that we're hailing them as a potential well-placed in generative AI and they're doing very little generation.

DG: Yeah, I think that's a fair point and I think, Nat, to your question, a good yardstick would be, "How much did they personally micromanage the scripts for Apple TV+?", you have everything I said on the positive, on the negative side, I don't know if they dropped Jon Stewart because he wanted to talk about China or because the ratings were bad and that was their excuse, but if they were that exacting and it really was like, "We don't like your script," then yeah, the company probably will struggle. To some extent, if you're Apple, you'd say, "Fine, the pressure release valve is, we'll connect with all these other LLMs. Like Google's brand, OpenAI's brand handle the reputational harm of generating tokens that are odd, all we do is control your phone." I think that would be a very palpable stance.

NF: Now that you and Ben said that, I kind of think that's what'll happen. They'll just take action on your behalf and try to do helpful things. They won't give personal advice and they won't generate political essays and that kind of thing. I think that seems like a straightforward approach for them.

Well, they also did, in the State of the Union, they talked a lot about the tools they are delivering to developers to locally optimize external models, and that's the pitch, which is, "If you want to do this stuff, and we'll help you do it locally, but we're not going to actually do that ourselves. We will provide all these frameworks and ways to optimize the model from someone else if you want to put it in your app, but our models are not going to get into that."

DG: Yeah, I think that's the first answer to them.

The second point is, going back to the Talk Show last night, Gruber asked the Apple executives that were on stage about the moderation question which you just raised. They were up there like, "Yeah, we believe that people are creative, we don't want to get in their way, we are just a tool". Number one, stand up and applaud because that's what I want to hear. Number two, classic strategy credit, because if they're not actually generating any content, there's nothing they can be held responsible for.

I posted that note in an Update a few months ago where a mutual friend of ours wrote sort of a framework for thinking about moderation, and it was really about you need to frame these tools in the context of them being tools, and that is how you're going to get societal acceptance and be able to work around these issues; when something's clearly ascribable to an LLM, the LLM is going to incur brand risk. If the action's clearly ascribable to the user, then you can kind of do whatever you want. And it feels like that is exactly Apple's approach.

DG: I think that's correct, and I think the great testament to Apple, is that they did not launch something sooner. I was always told when I was there that you play in the fourth quarter of the game and the role is to be best not first. So you see that now. I mean, they benefited so much from watching these other companies flail around in the public eye.

Also just the tremendous amount that's happened as far as optimization and making this stuff efficient. You go back to what ChatGPT launched compared to GPT-4o, which "o" seems to be for "optimized", that's my takeaway. How are they actually doing this at scale? That seems to be the biggest breakthrough as far as that goes.

NF: Can I throw in one more thing about Apple? Just to make one bear statement at least, because we've been so effusive. So the bear statement is that these models basically have proven themselves to be competent mostly in generation. Apple is betting on them in this tool-use function calling world, and while there have been some demonstrations that that can work, no one's done it at the level, the scale, the reliability that Apple now needs actually for it to work. So I think it very well can, so I don't think this is a losing battle necessarily, but that's the frontier. They now have to push in order for their vision to play out and it's not where the models have demonstrated the most success so far.

Yep, that's a great point and they have to actually innovate in this arena where all this stuff we're talking about is fast following, this is actually doing something new that no one's demonstrated.

OpenAI

And where's OpenAI in this? I analogized them to FedEx and UPS relative to Amazon, where Amazon just dumps the worst tasks on them that Amazon doesn't want to do and they take all the easy stuff. But at the same time, one of my long-running theses is is that OpenAI has the opportunity to be a consumer tech company and they just got the biggest distribution deal of all time. Where do you perceive their position today as opposed to last week?

DG: I don't fully understand the value of the distribution from the Apple deal. Maybe it makes sense, maybe it's the Yahoo-Google deal. I think the question in AI is, if you're working on enterprise, that's one thing. If you're working on consumer, the old rules of capitalism apply and you need a disruptive user interface such that people remember to use your product versus the incumbents and maybe that was chat.openai.com.

Which is now chatgpt.com, by the way.

DG: Chatgpt.com, or maybe that's not enough. I think you saw a hint, not necessarily of just how OpenAI, but all of these labs sort of see themselves going in their product announcement where they created a thing that you just talk to, and it's quite possible that maybe that is sufficient to be a revolutionary new user interface to the point where they can create their own hardware, they can basically command the attention of customers.

But I sort of think the general rule in the handbook is, if you're going to be in consumer, you want to be at the top of the value chain. I mean, certainly it's a mighty and impressive company, but the deal with Apple doesn't really signal top of value chain. So the question is, really the ancient question we've been asking ourselves on this podcast for years now, which is, "What is the new revolutionary user interface that actually causes a change in user behavior?".

Does that mean that Google is the most well-placed? They have all the smartphone attributes that Apple does, they should have better technology as far as models go. Does it matter that they're worse at product or trust, like they don't have the flexible organization that you were detailing before? We spent a lot of time on Google the last time we talked, has anything shifted your view of their potential?

DG: I think it really all depends on whether you can make an experience, and it always has depended on whether you can make an experience that's good enough to justify a change in user behavior.

I'd argue for example, that there was a period in time where even though the actual interface was pretty simple, generating high-quality images was enough to cause a dramatic shift in user behavior. Midjourney is Midjourney not because it has some beautiful angled bar to pinch-and-zoom thing. It's just like that was the remarkable miracle that it had. It made really good images, and it gave it some sticking power. So it's this tension between defaults and inferior product and new revolutionary experiences, and whether they have enough to break the calcification of the incumbent.

It's quite possible that if no one has any new brilliant ideas that Google, even though the models don't seem to be as excellent, at least to the consumer's eye, that they survive just because they have some Android user base, they certainly have Google.com. I will say the thing that has been surprising to me is while the technical capabilities of Google's model seem impressive, the consumer implementation is actually I think worse than, "Just okay". I thought their integration of language models into search was abysmal, sorry, to be totally frank. It was referencing Reddit comments that weren't real facts, it's not that hard to fix this sort of thing. So they need to be doing the bare minimum I think to maintain their status in the hierarchy. It's possible they don't do that, it's possible that a new revolutionary user interface is also created, it's also possible that they catch up and they bumble their way through it and they're just fine.

But this is, I think the main question to the challenger labs, if they're going in the direction of a consumer product is, "How do you make something that is so great that people actually leave the defaults?", and I think we always underestimate how excellent you need to be. Enterprise things are a little bit different, by the way, and OpenAI is a very good lemonade stand just on enterprise dynamics, but consumer is in a way easier to reason about. You just have to have a miracle product and if that doesn't happen, then yeah, maybe you should be long Google and Apple and the existing incumbents.

NF: Yeah, on OpenAI it's really hard to count them out, they have so many advantages. They don't have Meta's or Apple's or Microsoft's distribution, that's true. But it does look like, at least from the SimilarWeb panels, that ChatGPT growth was reignited maybe by the 4o Launch, and if that's true, that's very interesting.

Then on the modeling side, they are still ahead, and their demonstration of 4o I thought was really impressive. The voice capabilities are great, the fact that they have such a small and cheap model they were able to make available for free to everyone with those capabilities was impressive. 4o was a new pre-train, this was not another kind of fine-tuned GPT-4, this was from scratch, it had to be for the multimodality that they delivered. It's clearly a small model based on its price and performance and that the fact that they're offering it for free and so I have to imagine that they're training a bigger one. It's possible that GPT-4o was a kind of checkpoint or maybe failed run of the GPT-5 training, and so they're probably going to train a bigger one. It's not clearly smarter than GPT-4 as people have used it.

GPT-4 is still smarter in my estimation, but 4o is so much faster. And to your point, the fact that it's free, there are product attributes that matter more than just answering esoteric questions correctly.

NF: Yeah, exactly. So if they have a bigger 4o that's much more intelligent soon, or 5 delivers, we know they've got something amazing with Q* that we haven't seen yet, and they have this combination of executing on research, on scaling, and on product better than most and then they have distribution deals with Microsoft and now with Apple. Then they have, let's just be honest, unbelievably strong leadership in many important ways that matter, ability to raise capital, ability to execute and redirect the organization. You watch those 4o demo videos from the 4o team and it's clear none of them had slept before those videos were recorded and so I think that intensity is not to be discounted.

They're the Kleenex of AI right now. It was one of the underrated issues when the whole brouhaha happened last year, which is, "Ok fine, the whole OpenAI team goes to Microsoft", actually losing the name ChatGPT would be an astronomical cost and I think how much of the value of tech companies is bound up in the sort of intangibles, but I think it's a pretty big one. Any distribution you can command on your own without having to go through any of the platforms is meaningful and they have it on the order of what, 100 million or more people?

NF: Yeah, I think that's right. On the other hand, they face challenges. They're basically a small company, relative to these big companies we've been discussing, subscale product distribution. There's clearly tension with Microsoft, we saw the Inflection deal, we saw reports that Microsoft was upset about the Apple deal, we see Mustafa and the Microsoft AI team now working hard to match and maybe beat GPT-4.

Is there any way to think of Inflection other than as a hedge against OpenAI that's directly downstream from last November?

NF: I think it is clearly that. It also could be seen as new product leadership and AI-native leadership for Copilot and for their own integration of OpenAI, Satya may have thought he needed that for some reason, so I think it's probably both. I'm sure they're partnering with OpenAI and I'm sure Satya wants, based on just the governance and the fragility of that partnership, I think it's interesting, Satya has not tweeted about OpenAI, last I checked anyway, since early January. He's tweeted about Cohere models and the Mistral models.

This is a big thing. When I interviewed him a couple of weeks ago, he was all about, "Oh, partnerships, different levels of the stack. We believe in the modular approach." That was in direct contrast to an earnings call in October of last year, one month before the OpenAI stuff went down, where he was talking about integration, "We're integrating from the model all the way down to our infrastructure down to we're now designing our chips that are being built with OpenAI in mind". There has been a 180 in the communication that happened, with Satya Nadella, it happened very professionally and subtly, but it is very different than the way he was talking not that long ago.

NF: So I don't know what's in his head or what's going on. OpenAI and Microsoft need each other right now, that's a really important relationship and I think it'll continue but clearly he wants to build his own first-party capabilities too, and that's what he's doing.

On Google, yeah, I don't think they've got it yet, but I'm pretty impressed by — a friend of ours was saying recently that when they ship things that aren't good, that's actually a bull sign for Google because it means they're able to break through the organizational hesitance about shipping things, and they're willing to make mistakes and they want to win so badly that they're willing to iterate.

Could that be simplistic though, where actually the reason, to Daniel's point, maybe the reason it's bad is because of the organization, and so in a desperate flailing attempt to overcome that, you just ship it and then you realize, "Crap, I shipped a glue on pizza."

NF: Yeah, the seams in the organization are tough, but, and you could see it as flailing. Maybe it's flailing, but if they learn from it and improve, then flailing's okay it's like a rocket blowing up on a pad until it finally starts landing on the ship.

I do think Gemini is a good model — if you looked at the I/O demos that they showed, the multimodal demos, they had a demo of glasses, which I thought was interesting, there's some vision there too. I think the question really is execution, and then the sort of structural problem with the business model. Basically the stuff really does compete with Google search in a meaningful way, but they're out there and they're trying and they could be much more hesitant and that would make me — they're not Apple, Apple taking their time here meant they were going to get it right.

Right, because Apple's business model wasn't at risk. So they had all the time in the world.

NF: Yeah. So I don't know, I think Google's still very much in the game and I think we should probably mostly on net update positively towards their willingness to ship and iterate quickly.

The Bubble Question

There was something [Nvidia CEO] Jensen Huang said on the most recent Nvidia earnings call, that I'm going to quote here:

> Let me give you an example of time being really valuable, why this idea of standing up a data center instantaneously is so valuable and getting this thing called time to train is so valuable. The reason for that is because the next company who reaches the next major plateau gets to announce a groundbreaking AI. And the second one after that gets to announce something that's 0.3% better. And so the question is, do you want to be repeatedly the company delivering groundbreaking AI or the company delivering 0.3% better? And that's the reason why this race, as in all technology races,

> the race is so important. And you're seeing this race across multiple companies because this is so vital to have technology leadership, for companies to trust the leadership and want to build on your platform and know that the platform that they're building on is going to get better and better. And so leadership matters a great deal. Time to train matters a great deal. The difference between time to train that is three months earlier just to get it done, in order to get time to train on three-months project, getting started three months earlier is everything. And so it's the reason why we're standing up Hopper systems like mad right now because the next plateau is just around the corner.

This is basically saying to the point about Midjourney, Midjourney was there on time or early, and so they commanded this huge amount of usage because at the time it was novel, and it doesn't matter that today there are lots of models that can generate those types of images. Now, once you actually figure out how to learn Discord and you're using Midjourney, you're fine, and you're sort of locked in and Daniel, you talked earlier about the astronomical difficulty in getting people to shift once they have use cases in mind and that was basically Huang's point about why we have no problem pre-announcing our products, because there is so much desperation to get out now because if you're a little bit better, but a month later you've lost.

I thought this was such a fascinating observation because number one, it sounds correct and number two, I felt when he said that I suddenly had this visceral understanding of the dot-com era, where it's easy to look back on these things and say, "How could you not see this coming?". And the answer is, "No, you could see it coming, but you could see it coming and you couldn't do anything but invest", and you look at the Microsoft and Google and Meta and all these sorts of things. I was talking to different folks at Computex, and this was sort of the takeaway. It's like, "Yeah, it's probably a bubble, but we have no choice, we have to push forward, I will lose my job if I don't". I don't know, maybe we're in a bubble, maybe we're not. My update to my mental model is it doesn't matter if we are or not, the same decisions are going to be made.

NF: We're in a bubble, in my opinion, no question. Like the early Internet bubble in some ways, not like it in other ways. But yeah, just look at the funding rounds and the capital intensity of all this, it's crazy.

But bubbles are not bad for consumers, they're bad for the investors who lose money in them, but they're great for consumers, because you perform this big distributed search over what works and find out what does and even the failed companies leave behind some little sedimentary layer of progress for everyone else.

The example I love to give his Webvan, which was a grocery delivery service in the Internet bubble, and because they didn't have mobile, they had to build their own warehouses because they couldn't dispatch pickers to grocery stores, and they tried to automate those warehouses, and then because the Internet was so small, they didn't have that much demand. There were not that many people ordering groceries on the web and so they failed and they incinerated a ton of capital and you could regard that as a total failure, except that some of the people at Webvan who worked on those warehouses, went off to found Kiva Systems, which did warehouse automation robots, which Amazon bought, and then built tens of thousands of them, and so Webvan's robot heritage is powering Amazon warehouses and some of those executives ended up running Amazon Fresh and they eventually bought Whole Foods and so all that led to a lot of progress for other people.

The other thing, of course, is that a lot of money gets incinerated and a lot of companies fail, the technology moves forward, the user — putting URLs at the end of movie trailers, people learned about URLs, but some great companies are built in the process and it's always a minority. It's always a small minority, but it does happen. So yeah, I think we're clearly in some kind of bubble, but I don't think it's unjustified. AI is a huge revolution and incredible progress will be made, and we should be grateful to venture capital for philanthropically funding a lot of the progress that we'll all enjoy for decades.

Well thank you to both of you. Go ahead, Daniel.

DG: Well, we don't see ourselves as philanthropists, Nat, unless it comes to scrolls, but I think, yeah, it is completely logical to believe that we are in a bubble, and that many people will lose their shirts, and that historic companies like Google and Amazon will be formed, I think all of those things are true.

One thing I'm never quite sure how to think of is why sometimes markets are very forward-looking and sometimes they're not. So for example, the market is telegraphing — the market is pretty AGI-pilled if you look at Nvidia stock price and you project and extrapolate future earnings. On the other hand, the Apple stock price did not move until the day after the event, the only piece of new information that existed in my world was Ben Thompson's post about it.

I think that's probably overstated, but I'll take credit, sure. What's the Apple equivalent of Roaring Kitty? That's me.

DG: But it is actually much more of a market gap than Roaring Kitty, it must be 10 times, 20 times Roaring Kitty. But yeah, that's the thing I always wonder is what drives some of those dynamics. In some aspects, I feel like AI is almost overheated, in some aspects I feel like it's definitely mis-priced and too cheap, and I don't know what drives those dislocations in the market. I do think it's quite possible that in dot-com you had this telco bubble, maybe a trillion of spend, CapEx build out, whatnot and the issue that happened was obviously it took a little bit longer for consumers to arrive than the market anticipated. I don't think we'll have that issue here because the world is so fully interconnected and whatnot.

Yeah, when Apple rolls this out, it's shipping out to a billion people.

DG: Yeah, and it's just going to be there so the thing we always wonder is what exactly would cause a plateau or a slowdown in exuberance?

Is the entire stock market resting on the quality of GPT-5?

DG: Right, so there are a couple of events, and that I think has another one or two that are I think, very good ideas he could share. But definitely the next revolution of frontier systems being just okay, I think could cause some kind of temporary slowdown and that would be obviously fantastic news I think for some investors who are careful with their bets, you'd get a much better entry price. But American capitalism is a wonderful system that expects to be fed miracles once a quarter and if it's not fed a miracle, it tends to get very depressed and gloomy. So it's possible that GPT-5 is only a modest miracle or the other similar models are only a modest miracle. It's possible that the one after it is only modest and that would cause a slowdown. Nat, are there any other things you think would cause just a temporary slowdown, I guess, at least in exuberance?

NF: I think it would be two things, and I don't know if it's an "and" or an "or". One is, yeah, basically the model capabilities don't improve enough and we're in between generations right now. It takes about three years to roll new models, new big pre-trainings and so we're kind of in the post-4, pre-5 era, and during this time, innovation has shifted to post-training and we've actually learned that you can do a lot in post-training, it can improve a lot. We've seen some multimodality progress, now we're starting to see some UI progress. Okay, one thing would be if capabilities just don't improve that much, and GPT-5 just kind of feels like a chatbot, GPT-4.5, that's how it feels, I think that could at least reduce the CapEx investment.

DG: Why would that happen? What would be the fundamental constriction that would cause that to happen?

NF: Well, it could be one of three things, I think. One is that we're reaching the limits of scaling. The log-log graph on which we want to draw a straight line, it starts to bend, and why would it bend? It could bend because for some reason we're not able to extract more intelligence from even more data, it could bend because we're running out of new data that's out of distribution, we just aren't learning that much from each token. Or it could bend because maybe there's not that many players operating at this frontier and we've seen time and time again, even the very best companies can screw up these pre-training runs, they're tough to do. We know OpenAI has been delayed in getting all the compute they want, we know these clusters are hard to keep running, we know that you can make bad parameter choices and architectural decisions, so it could bend just because of the top two or three companies they each make just mistakes and so GPT-5 is just a little disappointing for that reason.

Is there a role here for, there is $75 billion or whatever of CapEx and a few billion dollars of revenue, and that revenue doesn't show up? Isn't that sort of the more likely scenario?

NF: Well, that's the other thing that could slow things down. Maybe capabilities keep improving, but for some reason they're not translating to economic value that's captured at the application layer, and so you're an application, you're one of these hyperscalers, and it's like, "Why are we going to spend a $100 billion on a cluster if we can't break five or six billion in revenue on these products?", and yeah, I don't know.

I think one of the canaries in the coal mine there may be Microsoft Copilot revenue, I'm sure there's a lot of adoption of it right now because companies are eager to apply AI. If they don't like it and it doesn't grow, I think that would be tough.

If you look at who's making money right now, Nvidia's making money, CoreWeave's making money, Scale's making money, and there are dozens or so startups that are making $50 million plus, but it's not 100 or at least I don't know of 100 of them. So yeah, I do think it has to translate to application level revenue, and that hasn't fully been made clear to us yet. So I think either of those two things could slow down investment. I think the 2025 CapEx is just baked in, I would be really surprised if these big companies canceled their 2025 orders.

That money has already flowed through to TSMC.

NF: Basically yeah, that's happening. I think a question is 2026. If there were a small winter, it could be temporary, there could be little flat spots.

I'm going to jump in on that one. I think you guys are mistaken about this being a little thing if there is a slowdown. I think the implication of a bubble is that the overcorrection is, in many respects, even larger. Like 2001-2002 in tech was dark and you look back and you say, "Oh yeah, that is when all the great companies were built" — they were built because engineers could not get jobs and they were very cheap, and you could buy up all this dark fiber like Google did. It's interesting, because I think that one of the questions with the dark fiber bit is that was such a great asset because you could make the fiber better by updating the endpoints, the fiber was the fiber and so that was a tremendous asset. The big question is if there is a bubble and everything just goes to crap and everyone goes out of business and you guys, I'm having to start pay you to show up in my interview because you need a job.

NF: We're going to be okay, Ben, don't worry.

(laughing) What’s going to happen to all these GPUs? The theory is well GPUs wear out and the energy gains from new generations are so significant that the old ones are just not even worth running. That is I think a question, how useful are these GPUs?

But it just occurred to me in this conversation, the build out that will probably matter, that will persist, that will be the sort of dark fiber is going to happen on the TSMC layer and the Intel layer. We have a massive glut of leading-edge manufacturing coming because Intel is pell-mell rushing into this. TSMC needs to respond to them, they need to respond to Nvidia and they need to respond and Apple just basically said, “We’re massively increasing the amount of huge chips that we’re getting,” and so I could see a scenario where if there is this bad bubble, it does go badly, the GPUs do end up being fairly worthless, maybe that’s overstated, but we have huge amounts of leading-edge fabs that their marginal costs are basically zero and suddenly just chips in general get pretty cheap and the poor TSMC shareholders are in trouble.

NF: I’m not worried about that, I mean we’ll find uses for the chips, one of two things. One is if you’re Meta, for example, [CEO] Mark [Zuckerberg] has said this publicly, chips they don’t use for AI they can use for Instagram Reels for ranking and recommendations and for better ads and that kind of thing and there’s a lot of traditional ML out there that can make use of dark tensors or dark matmul or dark FLOPS so I think that’s going to be fine. The other thing is that experiments in research are currently compute-limited, there’s a lot more ideas for research than there is compute to run it, so if for some reason we’re doing less inference, it means we can do more research and I just believe we have some of the smartest people in the world working on AI. The number of those people has never been higher, no one’s out of research ideas, there are so many ideas to pursue. So I think a small winter leads to more experimentation, I don’t think all the researchers lose their jobs and I don’t think the GPUs run idle. If they’re cheaper and they’re not used for inference or scaled training, then they’ll be used for research and there’s so much more to find.

A big difference like you said is we have these huge companies now that can disperse new things to the entire world immediately and they can also afford to lose billions of dollars. Their stock won’t like it, but they will be fine as entities.

NF: Yeah, they’ll be fine. No one’s made, other than the smaller companies, no one’s made an existential bet here of the big tech companies, and all these discoveries are super simple and obvious in retrospect, and I think that trend will probably continue and so we may just have a couple of years of experimentation. And look, it’s guaranteed that AI in five years will be so much smarter and so much better than it is now. The shape of the curve from here to there, nobody knows for sure, but I don’t think a small winter is catastrophic for that five-year view.

What do you think, Daniel? We’ve been having a little debate over here.

DG: So suppose the price of the H100 system drops in half and suppose it’s now the headlines are there’s not, LLMs are pretty much all we have and they’re not that good. I think at the bare minimum, because there’s already been shown a fair amount of value, companies will snap those up if only to just try and scale up existing techniques because we’re not sure of the ratio between CapEx spend and intelligence increase, but even if that ratio is half what we expected today, the bounty is so large that I think people will, if I had to guess, the folks that have cash or have the ability to raise debt will probably still lean in. Just because again, I think this is a little bit closer to the age of discovery than it is dot-com and you’ve had one or two ships come back from the new world with some precious metals. So I think it makes sense to continue exploring even if you have a bad year where nothing comes back that’s valuable.

I think it's quite possible that the rate at which new economic value comes online is completely wrong in terms of the slope, there should be not just summers and springs, but winters and falls. Things getting as bleak as they were in 2000, I guess the reason I struggle with that analogy a little bit is one issue you had when the Internet bubble crashed is it was just going to take time for economic demands to come online. You couldn't really accelerate the base rate of that no matter how many AOL CDs you handed out at Fry's, whereas here you're going to be really bottlenecked by the rate at which you can make model improvements, which I think you can control much more.

Well, there's so much more demand available now. There was a demand problem in 2001, that doesn't exist today.

DG: Yeah, exactly. Look, what we're trying to make here is a factory for intelligence, there's clear market demand for intelligence. There's an execution risk problem, I think less of a market risk problem. Maybe that's a good way of thinking the distinction between those two and so there may be moments where people wonder, "Is it possible at all to even execute?", and I would think, by the way, the bleakest scenario we could look at that, I think it might be somewhat relevant is autonomy. In 2011, you could have walked around San Francisco, people would've told you there's going to be not just self-driving but flying cars around the city when 2024 comes around and obviously that hasn't happened, but Waymo did see it through and it does exist now in San Francisco and it is a fully self-driving car. So I think that's a very bleak outcome, but it's not a dot-com real winter.

What's Next

So I was talking to a friend of mine on Wall Street a couple of weeks ago, and he said he was sort of increasingly of the view that models are going to be a commodity, but the source of his greatest misgiving is that that was becoming the consensus view, and I think there is a bit throughout this entire conversation about Apple being well-placed — you mentioned at the top that one of the reasons to be more bullish on Apple is that all those leading models have not completely converged but are by and large in the same ballpark.

Is there a bit where, "Okay, maybe that's true for LLMs", but there is something else going on? Is there something with Sora**? Is there something real there? Is there something with Q*, something we don't know about that is actually closer than it seems and maybe we've actually seen some evidence, like this idea of using video, of actually understanding the real world. Is it a possibility we talk again in four or five months and the fundamental premise of today's talk is obsolete?**

NF: Yeah, I interviewed the Sora creators yesterday, the day before on stage at an event and it was super interesting to hear their point of view. I think we see Sora as this media production tool, that's not their view, that's a side effect. Their view is that it is a world simulator and that in fact it can sort of simulate any kind of behavior in the world, including going as far as saying, "Let's create a video with Ben and Daniel and Nat and have them discuss this," and then see where the conversation goes. And their view is also that Sora today is a GPT-1 scale, not a lot of data, not a lot of compute, and so we should expect absolutely dramatic improvement in the future as they simply scale it up and thirdly that there's just a lot more video data than there is text data on the Internet. I think estimates are that YouTube has about an exabyte of data, Common Crawl is orders of magnitude smaller and it's just text. And then Andrej Karpathy, I was talking to him the other day too, and he said, "There's something strange going on-"

And a picture is worth a thousand words by the way, so the number of tokens there is just astronomically larger.

NF: He was exploring this idea that the world model and image and video models actually might better than in text models. You ask it for a car engine, someone fixing a carburetor and just the level of detail that can be in there is extraordinary, and maybe we made a mistake by training on the text extracted from Common Crawl and what we should do instead. I asked him for his most unhinged research idea. He said what we should do instead is train on pictures of web pages and when you ask the model a question, it outputs a picture of a web page with the answer and maybe we'd get way more intelligence and better results from that.

So I think this is just one of the research paths forward, is image and video just a way better, a richer source of information? Are those models somehow smarter? Q* is another, can we use synthetic data to generate reasoning? There's test-time compute. We've seen Devin and Cognition, the agentic demos there. I think the demos are very impressive and in general there seem to be a lot of paths forward and so they don't all have to work out. If one works out, we get much more intelligence.

I have an observation out of that. This gets at one of my favorite Daniel observations in this sort of series, which is that the fundamental flaw of text is it is the end state of entire series of thoughts and actually the level of intelligence you can acquire from the end state is shockingly small because in the overall value chain of generating that output, it's the top layer, you're missing all the intervening layers.

It is interesting to think about in the context of human intelligence, like to what extent you look at a baby, you look at a kid and how they acquire knowledge. I'm most inspired to do more research on babies that are blind or babies that are deaf, how do they handle that decrease in incoming information in building their view of the world and model of the world? Is there a bit where we started out with the less capable models, but when we do add images, when we do add videos, is there just an unlock there that we're underestimating because we've overestimated text all along? I'm repeating what you said, Nat.

NF: Yeah, Daniel was way ahead on this. I think Daniel said that in our first conversation together, and this is a really active area of research now, is how can we synthesize the chain of the internal monologue, the thinking and the dead ends and the chain of thought that leads to the answer that's encoded in the text on the Internet.

There was the Quiet-STaR paper and the STaR paper from [Eric] Zelikman who's now at xAI. I don't know what relation if any of that bears to Q*, but that's basically what he did is to use current models to synthesize chains of reasoning that lead to the right answers where you already know the answer and then take the best ones and fine-tune those and you get a lot more intelligence out of the models when you do that. By the way, that's one of the things the labs are spending money on generating is, "Can I get a lawyer to sit down and generate their reasoning traces for the conclusions that they write and can that be fed into the training data for a model and then make the models better at legal reasoning because it sees the whole process and not just the final answer?" — so chain of thought was an important discovery and yet it's not reflected in our training data as widely as it could be.

The biological aspect of all this is so interesting. To go back to the bit about data, and Daniel you're making this point of if you really are careful about the data that you put in, there's an analogy here to like the nutrients you consume as a human, what is the relationship between inherent genetic capability and the environmental impacts that go in — we barely understand this. In fact, it feels like every day you read something new, the degree to which we don't understand this with humans, but it would make sense to the extent there is an analogy and reasoning and predicting the next token is kind of what we do a lot of the time. There may be way more analogies than we expected.

DG: Yeah, I think babies are born with the ability, infants have the ability to detect snakes and spiders. So there's always this question of how much knowledge is encoded in DNA versus as learned experience. But yeah, I think we don't have to necessarily teach language models or AI to learn the way we learn, it's certainly not how we achieved flight, but it is sort of interesting that when you train a model on the Common Crawl, it is only seeing the final layer of human cognition and so it's missing all those intermediate steps and there should be plenty of ways to retrieve that, but I think that may explain some of the shortcomings of some of the models today.

It's just the product question then, it's actually a fool's errand to be attempting to recreate a human chain of reasoning and the whole breakthrough we're waiting for, the "feed moment" we're waiting for is to approach problems in a fundamentally different way that only an AI can do. This is the, "juggle a messaging thread and an email thread and a FlightAware data to give something useful", is that the whole conceptual trap we're stuck in? Our modern version of putting ads next to an article is trying to imitate the human mind and actually we should be creating an alien, it should think completely different.

DG: I certainly think that's possible. I also think, and maybe to touch back on our previous conversation, if you literally stopped all AI progress today and you just tried to make valuable products with the existing technologies, I think you could and I'm not suggesting that progress should be stopped, I'm just saying there's a lot of latent economic value in the existing capabilities of these models, even if there's no deeper breakthrough.

NF: A hundred percent agree.

DG: Obviously, hopefully there'll be better breakthroughs and the models will become easier to use, but you could also take today's models and make awesome stuff out of them and why that hasn't been happening, by the way, despite the best attempts of this podcast, I think is a very deep and interesting question. What is going on with Silicon Valley in general and is that innovation pipeline from research breakthrough to iPhone waning a little bit? I don't know.

Or is Apple just fulfilling the role that they've always fulfilled, and maybe it's full circle? Microsoft didn't miss mobile, they were trying to build a compelling smartphone for nearly a decade before the iPhone came out and it took Apple saying, "Oh, this is how you actually — this is the interface, this is what it is", and then you have this explosion that went on for 15 years and we're still living in it.

DG: Yeah, but why are those moments so rare and what is the underlying quirk that explains that? Is it a good organization of creative people that's really hard to do? Is it literally a Great Man Theory thing? Are the new cohort of founders as good as their ancestors were at doing this sort of stuff? I guess this is Patrick Collison and Tyler Cowen's idea of measuring progress in our little local world, but that's the thing I've sort of been really surprised by, and I think Nat probably has been really surprised by, is we could sit around on this podcast and come up with five ideas of billion dollar companies that haven't been created and they don't exactly know why they're not getting created.

NF: Yeah, I've been wondering this myself, but I think my conclusion is it just takes longer, it just takes a while. I think the first wave of AI startups that we saw were mostly researchers who weren't product thinkers and mostly wanted to do research and were mimicking OpenAI and basically creating a tech company/research lab with research freedom and tech companies salaries. Then we've started, I think only recently to really get some of the best entrepreneurs and best product thinkers off the bench. They were busy — people who are really amazing, are busy.

I think there is some effect where whatever people who are really good do goes pretty well, and so they get caught in local maxima a lot and so there's probably a lot of great people who should be working on this who aren't, they're doing something else that's important and rewarding and that seems to be changing, I would say.

So yeah, it's like why did the feed take so long? If you look at Tim Berners-Lee's original web, it sort of had some Web 2.0 features in it, it had generated web content, and yet it took many years before that became something widely done on the Internet. So yeah, I hate to say it, but I guess we need to be patient.

That is the theme, patient. Everyone will have to wait patiently for us to do this again. It will happen sooner than you think, even though we'll feel like it's a long time.

DG: Yeah. I mean, it goes back to the bubble question, which is, it's quite easy, it would seem to me, if you look at the history of some of the largest macroeconomic shifts at the time and people that trade them successfully, it would seem easy to overestimate the rate of progress on Earth versus in one's own mind and so I think timing is probably, even for us who are running these ten-year funds, timing is probably everything. The first few cohorts of mobile were actually not that great, I think it took about three or four years for Uber and Instacart to come around.

Obviously the first few quarters of dot-com seemed good and then weren't good and I guess the same will be the case here. Maybe this is an ancient thing people have been wondering after the invention of, I don't know, the abacus, people were wondering, "Well, why aren't we using that to — we just invented trigonometry, why isn't our farming technique better in Egypt?", I don't know, it's quite possible that was an ancient debate. But yeah, it's our job to both have the right idea and the right time horizon, I think, to invest on.

NF: Ben and I are old, and so we both have experience carrying luggage through airports in our hands. At some point someone was like, "Why don't we just put wheels on these things?". Wheels existed, we had rollerblades, we had skateboards, somebody just hadn't thought of it for some reason.

It's both encouraging and kind of depressing, for sure.

NF: Yeah, we're waiting for the AI-wheeled suitcase.

Nat, Daniel, thanks for coming on. Look forward to talking again.

NF: Thanks, Ben.

DG: Thanks, Ben.

This Daily Update Interview is also available as a podcast. To receive it in your podcast player, visit Stratechery.

The Daily Update is intended for a single recipient, but occasional forwarding is totally fine! If you would like to order multiple subscriptions for your team with a group discount (minimum 5), please contact me directly.

Thanks for being a supporter, and have a great day!

← Apple Intelligence is Right On Time

Subscriber's Daily Update

Tuesday, June 18, 2024
FTC Sues Adobe, The Legal Question, The Value of Doing Right
Monday, June 17, 2024
The Apple-OpenAI "Deal", Japan and the App Store, Apple's AI Developer Risk
Thursday, June 13, 2024
An Interview with Daniel Gross and Nat Friedman About Apple and AI
Wednesday, June 12, 2024

How Apple Intelligence Works, Apple’s AI Bear Case, The Talk Show Live