An Interview with Scale AI CEO Alex Wang About the Data Pillar for AI

Thursday, June 20, 2024

Good morning,

This Stratechery Interview is another installment of the Stratechery Founder Series; as a reminder, one of the challenges in covering startups is the lack of available data. My solution is to go in the opposite direction and interview founders directly, letting them give their subjective overview of their companies, while pressing them on their business model, background, and long-term potential.

Today's interview is with Scale AI founder and CEO Alex Wang. Scale AI is, outside of OpenAI, arguably the biggest startup success story in the AI space; the company just raised $1 billion at a $13.8 billion valuation. The irony — as we discuss in the interview — is that Scale AI is in many respects the least-AI company possible: the company's core offering is data for training AI models, which is ultimately sourced from hundreds of thousands of human contributors.

In this interview we discuss Wang's background and how he arrived at the idea of Scale AI, how the company has evolved both in terms of the problems to be solved and the ways in which to solve them, and why data generation is actually the most important factor in the future development of AI. We also discuss Wang's belief around U.S.-China competition in AI, and throughout the conversation get into important questions about where long-term differentiation in AI will arise.

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 Scale AI CEO Alex Wang About the Data Pillar for AI

This interview is lightly edited for clarity.

Topics:

Background

Alex Wang, welcome to Stratechery.

AW: Hey, Ben. Excited to be here. Huge fan of the podcast.

We're going to get to Scale AI in a moment, but first, I always like to start these conversations by learning more about how a founder got to where they are. This is a particularly fraught discussion for me because your story starts

in 1997, as in that is when you were born and I was already a junior in high school. But let's start there, I think it's actually quite interesting, where were you born and tell me your life story?

AW: Yeah, I was born in Los Alamos, New Mexico. The Oppenheimer movie came out last year, so I think now more people are familiar with Los Alamos as the original place of the Manhattan Project. There's still a national lab there, which is where both my parents worked and why I was born there — I believe it has the highest PhDs per capita in the country. Roughly speaking, everybody there is either a scientist or in a family with a scientist who works at the national lab.

Did your parents meet at the lab, or did they go there together?

AW: No, they met at their PhDs, and then they went there together. They're both physicists, and physics is the ruling field of study in Los Alamos, so my parents were very respectable scientists. When I was growing up, I think first my parents started teaching me physics when I must've been before elementary school so when I was in kindergarten. I remember they would teach me basic mechanics, and then it was sort of interleaved concepts about electromagnetism — it was definitely an unusual childhood.

I think my parents instilled a genuine love of math and science in me, I channeled it into competition. I did competitive math, competitive coding, competitive physics, and I started doing these at the state level and then moved on, did these at the national level. There were these National Olympiad Team selection processes, and I had a unique honor in the sense that I didn't get into any of the teams, but I got into the step right before that in all three fields: in math, computer science, and physics.

What is the step before that? Are you training the team, or it's the final selection or something like that?

AW: It's the final selection. So in physics, it's like the final 12, in computer science, roughly the final 12.

Got it. So, you're the Olympiad generalist, as it were.

AW: Exactly, jack-of-all-trades, master of none.

Master being very stringently applied here. You did make the final selection list, but yes.

AW: I'm sure it's happened since I did these competitions, but definitely at the time, I don't think anyone else did as broadly without doing much better in any one area.

So then I was in high school, I was bored, so I left high school a year early to to Silicon Valley to work, and this is because there were in competitive coding, there were a lot of competitive coders who were working at a small handful of tech companies in the Bay Area, one of these being Quora, and so when I was 17, I moved to Silicon Valley, I worked as an engineer at Quora for a year working mostly on speed optimizations, and then that's where I think initially cut my teeth on engineering and the tech industry.

What was different about engineering on an actual product as opposed to Olympiad Computer Science, whatever that entailed?

AW: I think the main insight is that in Olympiad you're basically told what to work on and in engineering on a product, you have free reign to pick whatever you want to do, and in my head, I basically realized very quickly that the problem selection, what you choose to work on, has ten times more effect on your ultimate impact than how well you solve any

individual problem. I got obsessed with this idea like, "Oh, you have to be in general, very good at problem selection, and that's basically the governor of ultimately how much impact you'll end up having".

What really made that land for you? Because I think that's a pretty profound lesson to learn at 17 with only a year in a startup. Was there a specific lesson that drove this into your head?

AW: So one of the things that was great about Quora was that they published internally all the A/B tests that were being done, and so you could see on these mailing lists, every single A/B test that was being run and the effect size of all these A/B tests, and Quora had a pretty great internal A/B testing infrastructure. You look at these experiments that come through, and obviously there's the hardest problems in terms of technical difficulty, there's the recommendation system changes or changes to the feed ranking algorithm, those are the most technically difficult, and then the easiest things were changing the color of a button. When you looked at the effect sizes of all these various experiments, some of the most impactful were the most simple: changing button colors is actually really important, or slightly changing button stylings has a lot of impact.

So Google gets a bad rap for being criticized for A/B testing a zillion shades of blue!

AW: It's actually crazy how impactful it is relative to — I was working on speed and optimization, changing a button color was the similar amount of impact to making the site 20% faster, and making the site 20% faster is a lot harder. You just look at these results and you're like, "Oh, huh. Actually...", there's very easy ways to have a lot of impact and there's very hard ways to have a lot of impact.

The Data Pillar

You left Quora then to go to MIT, is that the right sequence of events?

AW: Yep, I left Quora to go to MIT. I think at this time, right before I went back to MIT — it's funny, this camp hosted by this group of mostly effective altruists called Spark, but one of the people organized this camp is Paul Christiano, who was the inventor of RLHF [Reinforcement Learning from Human Feedback], and now is actually in the Commerce Department working at the US AI Safety Institute. But basically, it was this group of incredibly smart people who were running the summer camp and fully, they were all about deep learning, and this was in basically 2014 era, this was at the beginnings of the modern era of deep learning, and it was right after we had just gotten convolutional neural networks to work, so we just had image recognition starting to work on these neural networks.

Everyone who I thought was much smarter than me and older than me and more experienced than me, was saying that deep learning was clearly one of the most important things to work on, and even at that time, they were already starting to have debates on AI safety. They were like, "Oh, if this deep learning thing continues, then you're going to have AGI, and we have AGI, safety is one of the big most important problems", and obviously that core argument is what ultimately led to OpenAI being started.

So all these people were interested in deep learning, and I was like, "Oh, I don't really know that much about it," but thankfully I'm going back to college so I can just study it. I spent a year studying deep learning pretty intensely, I trained a bunch of neural networks in my dorm room to do various things, and then this led to, I think the core insight that I think has fueled Scale, which is that these models are made up of three things: there's the compute, the algorithms, and the data, and there were very smart people working on compute, Nvidia and other companies, there were very smart people working the algorithms, there were very few smart people working on the data. Then to the point on problem

selection in the past, it matters a lot what problem you select, less so that it's like the sexiest, most technically interesting thing. So I realized that, "Hey, there needed to be a company focused on data".

Yeah, I don't remember an International Olympiad about data labeling. Maybe I missed that one.

AW: I think it's in the future! Data labeling is getting pretty complicated.

When you saw that there was going to be a third pillar, yet no one was there, did you have any particular insights on how that would work, or was it just a matter of, "There's a problem space that needs solving, and we'll figure out how to solve it in the future"?

AW: Yeah. Probably the most formative, immediate experience was that I was training one of a neural network at this time on a single GPU in Google Cloud and using TensorFlow, and it was a neural network that detected emotion based on a photo of someone's face, and all I did basically was I took the tutorial for ImageNet, so basically literally the tutorial code for a very different image recognition algorithm, and then I just swapped out the data set and then pressed "Enter". Then 12 hours later, I had a neural network that smashed any of the other methods on this problem of recognizing emotion from images.

So the takeaway there is actually, data is what matters most.

AW: Yeah. From problem to problem, data is the only thing that varies, is maybe the better way to put it, and as a programmer, you kind of realize, "Oh, actually data is what's doing all the actual programming and my insight into the problem doesn't actually matter, it's just all embedded in the data set that the model ends up getting trained on".

So I think, A) I knew that data was very important. I remember this realization, the model ended at some performance, and I was like, "Okay, I've got to make this model better," and so then I was like, "Okay, how am I going to improve on this data set?", and then there was the second light bulb, which is that this is an incredibly painful process. You open up all the images and then you go through and you just look at, "Okay, are the labels for all the images correct?", and then you're like, "Okay, what new images should I get to pull into this?", and then, "How am I going to get those labeled?", and so all of the core operations, so to speak, of updating or changing or improving the data set were incredibly painful.

So I started the company in 2016, and this was an era where there was a broad-based recognition that platforms, particularly developer platforms that made very ugly things very easy were good businesses. It was already clear that AWS was ridiculously successful as a business, the most successful enterprise business that had ever existed, and then Stripe, it was also clearly recognized that Stripe was very successful, and so as a student of those companies realized that, "Hey, we should take this incredibly messy and complicated thing that exists today, and then figure out how to turn that into a beautiful developer UX and if we can accomplish that, then there's a lot of value to be had here".

There's a lot to unpack there. Just as a broader philosophical point, do you think that insight about data still holds? So it's not just that there's three pillars, compute, algorithm, and data, but actually data is the most important, and just like you saw before, is it more complicated now or is even more the case?

AW: Yeah, I think it's proving to be more and more the case. I was at an event with a lot of other AI CEOs recently, and one of the dinner conversations is, "Okay, compute, power, data: which do you run out of first?", and the consensus answer around the room is data, and I think the data wall has become over the past few months, a pretty commonly debated topics. "Are we hitting a data wall in LLM development, or are we just fundamentally coming against the limits of data?" Even the most liberal assumptions around, let's assume that you really did train on all human-generated text,

which no sensible person does because you filter out all the bullshit, but if you did train on all human-generated texts, even then we will run out by 2027, 2028.

So just overall in terms of the sheer amount of data that's necessary to keep up with scaling, we're very clearly hitting some meaningful wall, and then if you look at, I think a lot of the model performance improvements as of late, or sort of the big gains in models, my personal reason, I think a lot of that actually boils down to data, and innovations on how to use data, and innovations on basically the data-intensive parts of the AI stack.

Building ScaleAI

Well, let's get into how that data comes about. When you started Scale AI, you see that data makes a huge difference, there's no one in this space, there's an opportunity to solve messy problems that are complicated and expensive, and you mentioned the AWS and Stripe analogies there — how did you think that data labeling would work when you started, and how has that reality differed from your assumptions over time?

AW: Yeah, I think at its core, the basic idea, data labeling, if you boil it down, is like you have a bunch of data and you want to blend that data with human cognition. So you want real cognitive human output to combine with that data to serve as basically labels for models or other AI systems to learn from.

And that kind of output could be as simple as just saying what is in the image or something along those lines.

AW: Yeah, exactly. Some of the most basic data sets are like what's in the image or explaining the image or answering questions about an image, and then it's obviously gotten a lot more complex given that now, and we'll get into all of that but this core process, if you really think about it, the way I thought about was if you do this naively in terms of you just literally have images, you show those to people, and then you just have them manually label every single one, you have some level of efficiency, and if you do a good job of building great tools, if you do a good job of actually optimizing and finding the best people for this kind of work, if you do a good job of optimizing this automation in this process, you are able to automate a lot of the workflow using specialized algorithms to do so.

If you're able to do all of these different pieces, then you can drive dramatic efficiency gains or on the flip side, dramatic quality gains on top of the production of these data sets, and I always thought about as a Pareto curve, cost versus quality as the two axes of your Pareto curve, and your whole game is given a fixed budget or given a fixed amount of money, how high quality can you actually accomplish?

So when you started out, did you assume that — I mean you're coming from a background, you had been an engineer — was your assumption that this was going to be largely an engineering problem, or was it clear to you from the get-go that this was going to be a large human resources problem as well?

AW: This is a funny one. I think I actually remember ideally thinking when I built the initial version of the app, I was like, "Oh yeah, now I've built the app and this is actually going to be a shockingly easy problem, because now I just have to get people to use this app and then all the problems will solve themselves", and this was the most naive assumption ever, because the reality is the dispersion on what humans will do on a platform is just so crazy, and the level of operational difficulty in managing this is very challenging.

So to your point, we realized pretty quickly that this is actually a hugely operational problem and that, let's say 70% of the battle is not all the sexy stuff I just described to you, which is building automated tooling and building the best tooling and building more automation to the process, 70% of the battle was just the actual operational management of,

“How do you ensure that all of these people are well-trained? How do you ensure that they do great work? How do you ensure that incentivized to do the right work? How do you ensure that you’re able to communicate with them effectively?”, these became very quickly, 70% or so of the battle.

I’m surprised 70% is that low. Well, I mean, first off, how many people are we talking about? What is a nice big round number that would encapsulate the current number of people that are helping Scale AI do what it does?

AW: I think roughly speaking on a lot of the, what we call expert work, which is really where a lot of the work in improving these LLMs has trended towards, it’s probably on the order of a hundred thousand people.

But there’s a lot of non-expert work as well, correct?

AW: Yes. So it is worth talking about the history. Where we originally got started, again, started the company in 2016, and some of this goes to at the heart of AI as an industry, because AI at its core is such a general purpose technology, and so I’ve seen many waves of excitement on different forms that the technology has taken on.

But when we got started in 2016, more or less, all of the funding was going into autonomous vehicles and autonomous driving and nothing else really was getting funded super well in AI. Autonomous driving was getting astronomical amounts of funding, billions and billions of dollars of funding, not too dissimilar honestly to the modern generative AI companies, and the problems that you care about for autonomous driving are mostly ones of image recognition or object detection and path planning for these cars. You just need these cars to be extremely accurate at recognizing what’s around you, and so here the problem was you have a bunch of Lidar scans and images and videos and radar scans, and how you make sure that you’re actually among all the sensor data, grokking all the people and the pedestrians and the bicyclists and the posts and the construction signs and all that stuff.

So this we could do pretty effectively with people all around the world, and it didn’t require too much specialized knowledge or expertise. Then I would say our first foray into work that required specialized knowledge —

I’m sorry, curious about this. Was this a stage where Scale AI or your various subsidiaries, it was almost more like a marketplace where you have the connections with the companies that need this data, and then you are maybe biting off more than you can chew in terms of operational requirements, discovering why this is a sparse area? No one wants to have a company predicated on finding contractors all over the world to say, “That’s a stop sign”, “That’s a pedestrian”, or whatever it might be.

AW: I think actually in all the phases, and we’ll go to it, Scale is a pseudo-marketplace. We’re not quite a full marketplace, because obviously it’s not like people just transact on a platform, we do facilitate all the transactions, but it is a pseudo-marketplace in the sense that we have these model developers or AI developers on one side of the system, and then we basically convert their data needs into work and tasks that are done by a group of contributors, and the shape of what those contributors need to do or what they need to look like has changed over time. But at its core, it’s about ensuring that we have a big enough and broad enough span of qualified contributors to ensure that the work that needs to be done or the sort of data that needs to be built is able to be built for the model developers.

Did you just have to step-by-step realize, “Oh, we have to do this”, “We have to add this”, what was the balance there? You talked about the efficiencies and actually getting processes to do this, but was there just a huge amount of operational build up that had to happen? What was the balance between being an engineering company or a tech company versus being an operational company, like outsourcing at scale, to a certain extent?

AW: Yeah, I would say stage zero, so to speak, for any kind of business like this, is you solve every problem operationally, and I think if you stay in this stage where you're solving everything operationally, then I don't think you'll end up getting very far, but you can kind of get started and you can get a few initial contracts.

It strikes as that is actually a pretty significant moat. To the extent that it's just a really hard and messy problem, if you actually solve those hard and messy problems, no one is going to want to put forth the effort to reinvent the wheel in that regard.

AW: Oh, definitely. I remember there's a slide in one of our early pitch decks where we actually, we just put on a slide a systems design diagram that showed all the little problems that had to be solved and all the little systems that to work together, whether they're operational or software or all the various systems. Everywhere from how the whole entire quality control mechanism worked to how the entire recruitment machine worked to how the entire performance management system worked to how the entire training system worked, and it was the messiest diagram that you could possibly imagine, and that was actually the point. The point of the slide is to show-

(laughing) This is a picture of our moat.

AW: Yeah, exactly, and that this entire system is so complicated that yeah, sure, other people can try undertaking this, but regardless they're going to have to solve this entire big messy problem set and there's no way to get around the intrinsic messiness of the problem.

The AI War (with China)

How have the needs of the market shifted then? You mentioned that you were getting at this before and I interrupted. You start out with images for self-driving cars, today it's all about these text-based models. What is entailed in going from images to text?

AW: We had an interesting mid-step here, which is broadly speaking, I think the shift as the models have increased in intelligence is towards greater levels of expertise. But basically, we started autonomous vehicles and then starting about 2020 we actually started working with the government, the US government and this was driven because I grew up in Los Almos and realized that AI is likely a very important technology for our security.

We can do a side bit here, you wrote a very interesting piece on Substack in 2022, The AI War and How to Win It. Give me your thesis here and why you think it's a big deal.

AW: Yeah, I think that the basic gist is first, if you look at the long arc of human history, it is punctuated by war. In some sense, human history is all about war, and then if you look at the history of war, then the history of war in some sense is all about technology. If you look at particularly the transitions from World War I to World War II to future wars, the Gulf War for example, the most significant bit so to speak, or the largest factor in how these wars end up playing out really, is access to technology. Obviously this is deep to my upbringing, grew up in Los Alamos, basically every year you have a multi-day history lesson on Los Alamos National Lab and the origins thereof.

So then you think about, "Okay, what are the relevant technologies today that are being built?", and there's a host of technologies I think are important, hypersonic missiles, space technology, et cetera. But AI is, you could very easily make the case, that it is the most important. If you could solve problem solving, then all of a sudden you have this incredibly powerful advantage.

If you believe that AI is really important for hard power, for American hard power, which is very important for I think ensuring that our way of life continues, then the most shocking thing for me was looking at, was going through and looking at the things that the CCP [Chinese Communist Party] were saying about AI, and there are CCP officials who have very literally said, "We believe that AI is our opportunity to become the military superpower of the world". That we believe that roughly speaking, they said, "Hey, the Americans are not going to invest enough into AI, and so we'll disrupt them by investing more into AI proportionally, and if we do so, even though we spend a lot less on our military, we will leapfrog them in capability". This is, I think as a startup person, this is the core Innovator's Dilemma or the core disruptive thesis that the CCP had basically a disruptive thesis on war powered by artificial intelligence.

This is basically the idea that you're going to have these autonomous vehicles, drones, whatever, of all types controlled by AI, versus the US having these very sophisticated but operated by humans sort of systems, and the US will fall into the trap of seeking to augment those systems instead of starting from scratch with the assumption of fully disposable hardware.

AW: Yeah, I think there is at its core two main theses. One is perfect surveillance and intelligence in the sort of CIA form of intelligence, and this I think is not that hard to believe. Obviously, in China, they implemented cross-country facial recognition software as their first killer AI app, and it doesn't take that much to think, "Okay, if you have that, then just extend the line and you have more or less full information about what's happening in the world" and so that I think is not too hard to imagine.

Then the hot war scenarios is to your point, yeah, autonomous drone swarms of in land, air or sea that are able to coordinate perfectly and outperform any human.

I think when people hear AI, they think about the generative AI, LLMs, OpenAI, whatever it might be, and assume that's a US company, Google's a US company, et cetera, and so the US is ahead. This is obviously thinking about AI more broadly as an autonomous operator. Is the US ahead or what's your perception there?

AW: I think that on a pure technology basis, yes, the US is ahead. China's caught up very quickly. There's two very good open source models from China. One is YiLarge, which is the model from Kai-Fu Lee's company, 01.ai. And then the other one is Qwen 2, which is out of Alibaba and these are two of the best open source models in the world and they're actually pretty good.

Do they use Scale AI data?

AW: No, we don't serve any Chinese companies for basically the same reasons that we're working with the US military. YiLarge is basically a GPT-4 level model that they open-sourced and actually performs pretty well, so I think that on the technology plane, I think the US is ahead and by default I think the US will be maintaining a lead.

There's an issue which Leopold Aschenbrenner recently called a lot of attention to, which is lab security. So we have a lead, but it doesn't matter if, it can all be espionaged away basically and there's this case recently of this engineer from Google, Linwei Ding who stole the secrets of TPU v6 and all these other secrets.

And wasn't discovered for six months.

AW: Yeah, it wasn't discovered for six months and also the way he did it was that he copy-pasted the code into Apple Notes and then exported to a PDF, and that was able to circumvent all the security controls.

Creating Expertise

So how does this tie into this middle stage for you of starting to sign government contracts? What were those about?

AW: Yeah, so I basically realized, and the punchline of what I was going through was that the United States was, by default, going to be bad at integrating AI into national security and into the military and a lot of this is driven by, for a while — this is less true now, but for a while — tech companies actively did not want to help the DOD and did not actively want to help US military capabilities based on ideology and whatnot, and even now the DOD and the US government are not really that great at being innovative and have a lot of bureaucracy that prevent this. So I decided basically like, "Hey, Scale, we're an AI company, we should help the US government".

We started helping them and we started working with them on all of their data problems that they needed to train specialized image detectors or specialized image detection algorithms for their various use cases, and this was the first foray into an area that required a lot of expertise to be able to do effectively, because at its core, the US government has a lot of data types and a lot of data that are very, very specialized. These are specialized sensors that they pay for, they're looking at things that generally speaking the general population doesn't care about, but they care a lot about — movement of foreign troops and the kinds of things that you might imagine military cares about — and so required data that was reflective of all of the tradecraft and nuance and capabilities that were necessary, so this was one of the first areas.

We actually have a facility in St. Louis, which have people who are by and large trained to understand all this military data to do this labeling.

So this was a clear separation then from your worldwide workforce?

AW: Yeah, exactly. It was a clear break in the sense that we were doing problems that almost anyone in the world could, with enough effort, do effectively and do well, to almost like the Uber driver, a very broad marketplace view, to something that required niche expertise and niche capability to do extremely well.

This sort of phase transition of data — there's sort of a realization for us that, "Oh, actually in the limit almost all of the data labeling, almost all the data annotation is going to be in the specialized form", because the arc of the technology is, first we're going to build up all this generalized capability, and this will be the initial phase building of all these general capability, but then all the economic value is going to come from specializing it into all these individual specific use cases and industries and capabilities and it flowing into all the niches of the economy.

Well, all that first problem though, you just talked about, it's super operationally difficult and you solved a lot of problems there that people wouldn't have to do. Is this second part, is it operationally difficult as well, or is it just a matter of having the right experts? What makes that new problem, those niche problems difficult?

AW: No, I think it makes it actually doubly difficult because you have all the same problems that you had before, which is you have all these operational problems and getting complex groups of humans to do high quality work that will go into an algorithm, and ensuring that the data quality at the end of the day is very high. But then now you have the secondary problem, which is that you have to go recruit this very fragmented marketplace and ensure that you have this very broad expert network that encompasses every possible human language, every possible field of study, different jobs and expertises, that you're able to actually cover the broad span of human knowledge.

I think one analogy that I think about is, if you think about Airbnb, we think about Airbnb as one big marketplace, but actually it's this amalgamation of a bunch of sub-marketplaces or very fragmented sub-marketplaces and one of the reasons it's so hard to disrupt Airbnb is because you can't just aggregate one subsection of their supply and then be able to feasibly compete, you actually need to figure out how to aggregate across all of the niches of supply that they've managed to recruit hosts onto, and so I think if anything it's made the business, from a pure study of moats and defensibility, a lot more interesting.

What is bigger, more important for differentiation then? Because you could go to Airbnb, you could say, "Well, the reason they can agglomerate all those different sub-niches is because they have consumer demand", so that is ultimately the pull that brings suppliers on to their platform. Certainly there's a chicken-and-egg aspect to this, but when it comes to Scale, is it that, "Look, we have demand and that may be manifests, and we have money to pay these experts", and that's something that they can be aware of? Or is it that, "We've got the experts and then we can go, we secured supply, and so we can acquire demand"? What direction does that flow?

AW: I think it's a little bit of both. In general, I think we aggregate a huge portion of the demand in the marketplace, and so therefore we're able to, we have the resources and the capability, to go out and recruit experts across every language, every field, every job class, every job, family, and are able to build that breadth of the expert network. But then it's a self-reinforcing cycle because once you have this sort of breadth and depth of the network of contributors, then it's very easy for new model developers to very quickly get spin up, basically new data sets and get access to new data utilizing the same expert network. So it's hard to say, which is the progenitor of the cycle, but the cycle is alive and well.

So if it matters, if data is actually the biggest differentiator in the long run and you're the king of data, isn't there a sense that you're commoditizing the marketplace generally, because everyone's going to end up with the same data? Is there a big motivation that if someone wants to build a highly differentiated model, difficult as it may seem, they will need to recreate some aspects of what you're doing just so they can have something unique?

AW: Well, I think there's two thoughts here. The first is this marketplace of experts and their ability to perform high quality work, I think that, by and large, I think there's a huge amount of cost that's gone into building this, and it's very difficult to fully replicate exactly what we've built, and I don't think a lot of our customers are super motivated to do that. I think what a lot of our customers are motivated by are, their plane of differentiation really is, "Okay, there exists all these human experts and we can get them to produce data in a whole variety of ways, what are the best methods for them to create data that are most beneficial to our algorithms?", and really innovating on the methods of different ways to utilize these experts to produce data for their systems.

Do they give you those methods to go implement or how does that relationship work?

AW: It's a combination of both. I mean, ultimately all these, in a nutshell, all of these methods involve some tight integration between a model from the customer, so generally a very advanced large language model they want to improve in some ways, and the ways that the humans give input.

So the early version of this was what I referred to earlier in the Interview, RLHF, Reinforcement Learning from Human Feedback, which is a very naive method at its core, but it still is in some ways state of the art. This is where you show human experts, the model produces two versions of an answer, so let's say you ask the question, "Where should I go in Italy?", or, "Where should I go in Florence?", or whatnot, and it shows two versions of the answer, and then a human expert will just pick the one that they think is better, and you just do that over and over and over again. This clearly

involves a tight integration with the model from the customer, and with enough samples, the model learns at honestly a pretty coarse level what humans prefer and what humans think is better and then does reinforcement learning optimizes along that curve. That's the simplest mode, I think there's been a lot of innovation in this direction.

So one of the things that was published is called Process Supervision, this is out of OpenAI, where what you do is you have the model — let's say you ask the model a math question or whatnot, you produce its full chain of thought. So, "If I have a right triangle with legs of 3 and 4, what's the last leg?", and then it's sort of reasons through, and it's like, "Okay, it's a right triangle, that means you can apply Pythagorean's Theorem, if you apply the Pythagorean Theorem, then it's three-squared plus four-squared equals X-squared, solve for X". Anyway, that's the stage of steps, and then a math expert-

And now we've achieved the capability of a 18-month-old Alex Wang, but yes.

AW: (laughing) And then a math expert will go through and say, "Oh, you're making a mistake at this step and this is the mistake that you're making", and then the labs are all racing towards much greater levels of innovation along this curve of, "What are the best ways to extract or utilize human experts?", and I think that's the relevant plane of differentiation.

So just to make sure I understand the sequence, you start out with very naive labeling, mostly of images, working with the government, drives more of a shift towards expertise as far as labelers go and then, so this step three then, is this kind a step up from data labeling or data generation to this other side, of being in the RLHF process? Is that where a lot of the new work is, or are you still doing a lot of original data generation as well?

AW: Yeah, at this point the lines kind of blur because I think our customers still think about it all as data labeling or data generation. But to your point, the complexity of the process is dramatically increasing over time, and I think the complexity of this process will keep increasing and there'll continue to be innovation in terms of what is the exact right way to leverage — I mean, the maybe zoomed out way to think about this is for every ounce of human brainpower, brainpower from an expert, how do you improve the model the most, with every ounce of human brainpower, let's say for every watt of human brainpower spent, how does that drive the biggest improvement of the model?

Is that something that you are figuring out or the model makers are figuring out, or what, where's your role in this process as far as applying expertise to this problem versus having this operational model that again, I think is underrated in terms of its defensibility, who wants to actually do all this work, but is there a ceiling where you go and then the model makers are taking care of that? How's that balance?

AW: Oh, yeah, I would say that at its core, all of them are very collaborative relationships, but the way that we view it is, roughly speaking, their job is to innovate on the algorithmic methods and exactly how they want to approach these problems, and then our job is to make sure they have the right primitives to ensure that they can accomplish those things. So kind of like AWS to the app developer paradigm, the app developer is constantly iterating to find unique ways to build apps, we just have to supply the right primitives to enable them to do so.

So where does synthetic data come into this?

AW: Yeah, synthetic is super fascinating. So I think that this has become super popular because we're hitting a data wall, in some ways the most seductive answer to the data wall is, "Oh, we'll just generate data to blow past the data wall", generate data synthetically using models themselves. I think the basic results are that, at a very high level, synthetic data is useful, but it has a pretty clear ceiling because at it's core you're using one model to produce data for another model, so it's hard to blow past the ceiling of your original model at a very fundamental level.

It's a compressed version of what went into the original model.

AW: Yeah, exactly. It's a very good way to compress insight from one model to get to another model, but it's not a way to push the frontier of AI, so to speak.

Is this a story though, of one of the reasons why it's easier to have big advancements in small models? Because in this case, it's sort of like if you want to build a really optimized small model, synthetic data is actually maybe better than the original data because it's gone through that compression step in some regards, but it's maybe not suitable for a large model. Is that an appropriate distinction to draw?

AW: I think if you want to make a model good at something that some other model is already good at, then synthetic data is a big part of that story. There's still a lot of nuance because you want — in the actual process of producing synthetic data, usually you need a fair amount of human data as well to ensure that you get the right kinds of outcomes. But broadly speaking, yeah, if you want to make your model good at something that another model is good at, synthetic data is a part of the story.

I think that the intellectual laziness comes in when it's like, "Okay, how are we going to push the frontier of these models?", because these models today are not near the capability that we want them to be in the future, and this is where you have to believe that there's going to be a huge amount of — this is where we have to have to truly innovate and where the approach that we talk about a lot is hybrid, human AI, synthetic data, and it's a continuation of the curve that we just talked about, how do you get the biggest bang for your buck in terms of output from a person?

In some ways AI is a great productivity tool. So is there a way that you can use AI as a productivity tool to enable a person to produce more data more quickly? Simple example is if you wanted to produce the perfect answer to a prompt, let's say, your prompt is like, "What's the morality of X, Y or Z?", then the model produces a first draft. The human expert will read through that, critique it a few times and then will edit after it's been critiqued a few times. So if you were to measure the wall speed of how long that process will take, it's going to be a lot faster if the human uses the AI as an aid.

Right. Well, this is the selling pitch for AI generally. I have a friend in graphic design who is actually very enthusiastic about AI, the image generation or the idea generation capabilities are what's compelling to him, and is this a similar case where you can get that first draft out sooner and the human value as you see it is in the editing process as opposed to the generation process.

AW: One thing that I think is important to get across is actually the biggest barrier to AI progress today is what we call frontier data, but to push the frontier of the models, you need frontier data and this frontier data looks very different usually from the data that exists on the internet. So what is actually really needed to push, let's say, agent capability? Everyone wants to build these models into agents that work super well, what's the biggest barrier to that? Well, it's a lot of data that shows what models should be doing as agents. So what should their chains of thoughts be? What do they do when they get stuck? How do they correct their own errors? What are the tools they're supposed to be using? How do they use a scratch pad? This data of how models are supposed to behave as agents just does not exist.

The Internet is all completed output. You're missing all the intervening steps and how it got there in the first place.

AW: Exactly, and we as humans, we go through these intermediate steps, but we don't record it ever, almost ever, we never really fully explain and show all of our work, and so a lot of the data that is missing to actually bootstrap the models to the next layers of capability is frontier data, agent behavior data, reasoning chains, whatever it may be that's

needed to actually get the model to that level of performance, and so I think that the means of production of data are a really, really critical thing and just in the same way that I think in the industry, we spend a lot of time thinking about the fabs and we spend a lot of time thinking about chip production and compute production and power production and these other key ingredients, we need to think a lot about data means of production, how do we ensure that we have all the ability, the maximal ability, and the most competitive ability to produce new frontier data to go into the models.

The Data Foundry

So say a model maker comes to you and they work through this and they develop some new RLHF process or RLHF 2.0 or whatever it is to develop and discover this process. Then the next model maker comes along and they say, "Scale AI, can you help us out?", who owns that process? Who figured out how to do that? Is that the TSMC, "Look, we figured out how to make a faster chip, everyone benefits?", or is that a, "We're Apple where we have specific IP that went into our chips, that's ours, that's not a Lego building block that can be shared?"

AW: By and large, we take the approach — again, we're a platform provider, our approach is we don't share. If one lab innovates, we don't share that with anyone else unless they choose to publish it or reveal it of their own accord.

And how many of these breakthroughs are operational in nature as opposed to some sort of novel IP breakthrough?

AW: Oh, if there's a breakthrough at what we call the primitive layer, which is how to organize the people or how to ensure that they get trained appropriately, how to ensure that you're able to performance manage them effectively, those obviously get embedded into our overall platform.

Right. But let's say you figured out how to navigate a UI effectively or how to train a model to do that, it's hard to draw the line, what is the insight there versus what is the operational problem of getting people to train a model to navigate a UI?

AW: Yeah, good point. I mean, the key is basically if you think about the human data primitives, it's really around like, "Okay, what are the various ways to dispatch human experts to produce data for these models?, and, "What are the right ways to operate in those interfaces?". We won't reveal an interface from one customer to another, if that makes sense. So if one customer figures out, to your point, RLHF 2.0, some really great way to do alignment and posturing of these models, then we're not going to reveal that interface that we built to another customer by default.

Are they providing some of these interfaces then and you're just providing the people? Or is everything running on your software? I'm just curious about the mechanics of this, how that works.

AW: Yeah. By and large, it runs on our software, but these interfaces, the design and build out of these interfaces is usually a very collaborative process. Usually, they're like, "Hey, this is kind of the thing we want to do, this is kind of what we think that should look like", and we're like, "Okay, yeah, based on our expertise, this is actually how we would design it appropriately". Again, I think the analogy to cloud is pretty good because ultimately for every customer who goes on AWS, they don't actually make the decisions on how to design their cloud, there's cloud consultants who help them actually make all the underlying design decisions.

So you just raised $1 billion at a $14 billion valuation, double your previous valuation, which had fallen a bit in the secondary markets a couple years ago. Is this just the AI hype train? Is there a meaningful shift in the business going forward? Put your investor hat on. You've been talking to plenty of them recently, what is the opportunity that they are seeing going forward?

AW: I think at the core, if I distill it into a simple tagline, it's "Nvidia for data". It's obviously very clearly demonstrated how profitable and how big of a business Nvidia is and we are a parallel pillar to the AI stack. There's three things you need. You need compute data algorithms and our role is to be the Nvidia for the data pillar, so be the platform player at the data layer.

So are you experts or contributors, is that your TSMC? Or what's the relationship there?

AW: Well, our contributors are really critical. Yeah, I think you could think about them as TSMC. I mean, obviously at some point the analogy—

(laughing) It does fall apart pretty quickly, to be fair.

AW: Breaks down. But at its core, I mean, I think that if you look into the future, data is — we talk a lot about the compute intensity and the compute bottlenecks of these models, but the data bottleneck I think is going to become more and more clear over time.

So basically this is huge problem everyone is running into, it's incredibly hard to solve and so someone is going to need to solve it and you've been working on it for eight to ten years or however long it's been. The thesis seems pretty fairly straightforward, even if the margins are not necessarily going to be Nvidia-style margins, given that you have to use hundreds of thousands of humans to do that.

AW: Yeah and I think the other key nuance here, the other interesting thing, is today our revenue is 1% of Nvidia's because, by and large, the budgets are mostly allocated towards compute. I think as with any portfolio optimization problem, in time, if data is actually the biggest problem, the percent of budgets that are allocated to data versus compute will slowly shift over time. So we don't have to be half the budgets, even if we get to 5% of the budgets or 10% of the budgets versus 1% of the budgets, then there's a pretty incredible growth story for data.

So what's the bigger irony or paradox? Is it that the Olympiad competitor raised by physicists is actually managing an astronomical people problem? Or is it that Scale AI is — you think about that, you think about the big models, but no, it's actually just a lot of grunt work.

AW: One our investors, actually Peter Thiel, has said this when I meet with him about the company, and one of his investment theses — he's a very smart investor, so this is kind of a joke — but one of his investment theses is that, "It's just a really good name, it's Scale AI, it's a really good name".

But I think it's actually the core of it, which is that I think we — well, first, we named the company before this idea of scale became all that was needed for these models, so that worked out. But in general, just like Nvidia and TSMC and the entire supply chain have been able to produce massive scale for compute, I think our job is to ensure that we have massive scale on data, and the first step of that in LLMs was just taking the whole Internet, but the next step has to be some means of production.

Anything you would do different if you were to do this path all over again? I mean, you started the company so young. You had to figure out this worldwide sourcing problem, then you have a US sourcing problem. What would you do differently if you had to start all over again?

AW: I would say one thing that we did well, but I honestly would double down on even more, is that AI as a technology is so self-disruptive, in the sense that if I look at what we were doing in 2016 when we got started around convolutional

neural nets, that in some sense is so yesterday and so unimportant relative to large language models today, and I think the same thing's going to happen.

Six years from now, we're going to look back and it's going to be a totally different model paradigm, a totally different underlying AI paradigm, and so one of my take-a-step-back realizations of AI is that the technology is so nascent and progress in technology is so fast that it's just going to disrupt itself a bunch of times over, and so what that means for us is we actually should invest a greater percentage of our resources, and we do in fact, towards what are the new methods, what are the new upstart AI approaches, and how do we make sure that we're always serving those effectively.

Not too dissimilar, by the way, from I think how [Nvidia CEO] Jensen [Huang] and Nvidia think about their business. They were a gaming business, and then they leaned hard into this AI training thing because I think it was a fascinating intellectual problem and then that built the moat to fuel them being probably the biggest company in the world.

Well, I mean, Nvidia's very sexy, I think the most attractive part of your business is its non-sexy aspects, the parts that people don't like to think about, don't want to talk about, but that is precisely why there was no one in that area and that seems to be a pretty big advantage.

AW: Yeah, it's funny, a lot of people who I did either math competitions with or who I knew from my math competition days have gone on to become very prominent people in the AI industry, like Paul Christiano or many others. [OpenAI President] Greg Brockman gave a talk at this camp, and I think it's incredible the work that they've all done at the research level. But I'm very happy, we're very happy, being in the background, quietly doing the dirty work to fuel the industry.

Alex Wang, it was great to talk to you and looking forward to talking to you again soon.

AW: Yeah. Thanks so much for having me on.

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
Summer Break: Week of July 1st →

Subscriber's Daily Update

Wednesday, June 26, 2024
An Interview with Marques Brownlee (MKBHD) About Being a YouTube Star
Tuesday, June 25, 2024
European Commission Charges Apple, Apple Delays New Features for E.U.
Monday, June 24, 2024
Perplexity and Robots.txt, Perplexity's Defense, Google and Competition