Lex Fridman Podcast #459 - DeepSeek, China, OpenAI, NVIDIA, xAI, TSMC, Stargate, and AI Megaclusters

Published - February 3, 2025

Transcribed by - thepodtranscripts.com

Lex Fridman
The following is a conversation with Dylan Patel and Nathan Lambert. Dylan runs SemiAnalysis, a well-respected research and analysis company that specializes in semiconductors, GPUs, CPUs, and AI hardware in general. Nathan is a research scientist at the Allen Institute for AI and is the author of the amazing blog on AI called Interconnects. They are both highly respected, read and listened to by the experts, researchers and engineers in the field of AI. And personally, I'm just a fan of the two of them, so I used the DeepSeek moment that shook the AI world a bit as an opportunity to sit down with them and lay it all out from DeepSeek, OpenAI, Google xAI, Meta, Anthropic to Nvidia and TSMC, and to US-China-Taiwan relations and everything else that is happening at the cutting edge of AI. This conversation is a deep dive into many critical aspects of the AI industry. While it does get super technical, we try to make sure that it's still accessible to folks outside of the AI field by defining terms, stating important concepts explicitly, spelling out acronyms, and in general, always moving across the several layers of abstraction and levels of detail. There is a lot of hype in the media about what AI is and isn't. The purpose of this podcast in part is to cut through the hype, through the bullshit and the low resolution analysis and to discuss in detail how stuff works and what the implications are. Let me also, if I may comment on the new OpenAI o3-mini reasoning model, the release of which we were anticipating during the conversation and it did indeed come out right after. Its capabilities and costs are on par with our expectations as we stated. OpenAI o3-mini is indeed a great model, but it should be stated that DeepSeek-R1 has similar performance on benchmarks, is still cheaper and it reveals its chain of thought reasoning, which o3-mini does not. It only shows a summary of the reasoning, plus R1 is open weight and o3-mini is not. By the way, I got a chance to play with o3-mini and anecdotal vibe check wise, I felt that o3-mini, specifically o3-mini high is better than R1. Still for me personally, I find that Claude Sonnet 3.5 is the best model for programming except for tricky cases where I will use o1-pro to brainstorm. Either way, many more better AI models will come including reasoning models both from American and Chinese companies. They'll continue to shift the cost curve, but the quote "DeepSeek moment" is indeed real. I think it will still be remembered five years from now as a pivotal event in tech history due in part to the geopolitical implications, but for other reasons to, as we discuss in detail from many perspectives in this conversation. This is the Lex Fridman podcast, to support it please check out our sponsors in the description. And now, dear friends, here's Dylan Patel and Nathan Lambert. A lot of people are curious to understand China's DeepSeek AI models, so let's lay it out. Nathan, can you describe what DeepSeek-V3 and DeepSeek-R1 are, how they work, how they're trained? Let's look at the big picture and then we'll zoom in on the details.

Nathan Lambert
DeepSeek-V3 is a new mixture of experts, transformer language model from DeepSeek who is based in China. They have some new specifics in the model that we'll get into. Largely this is an open weight model and it's an instruction model like what you would use in ChatGPT. They also released what is called the base model, which is before these techniques of

post-training. Most people use instruction models today, and those are what's served in all sorts of applications. This was released on, I believe, December 26th or that week. And then weeks later on January 20th, DeepSeek released DeepSeek-R1, which is a reasoning model, which really accelerated a lot of this discussion. This reasoning model has a lot of overlapping training steps to DeepSeek-V3, and it's confusing that you have a base model called V3 that you do something to to get a chat model and then you do some different things to get a reasoning model. I think a lot of the AI industry is going through this challenge of communications right now where OpenAI makes fun of their own naming schemes. They have GPT-4o, they have OpenAI o1, and there's a lot of types of models, so we're going to break down what each of them are. There's a lot of technical specifics on training and go through them high level to specific and go through each of them.

Lex Fridman
There's so many places we can go here, but maybe let's go to open weights first. What does it mean for a model to be open weights and what are the different flavors of open-source in general?

Nathan Lambert
This discussion has been going on for a long time in AI. It became more important since ChatGPT or more focal since ChatGPT at the end of 2022. Open weights is the accepted term for when model weights of a language model are available on the internet for people to download. Those weights can have different licenses, which is effectively the terms by which you can use the model. There are licenses that come from history and open-source software. There are licenses that are designed by companies specifically all of Llama, DeepSeek, Qwen, Mistral, these popular names in open weight models have some of their own licenses. It's complicated because not all the same models have the same terms. The big debate is on what makes a model open weight. It's like, why are we saying this term? It's a mouthful. It sounds close to open-source, but it's not the same. There's still a lot of debate on the definition and soul of open-source AI. Open source software has a rich history on freedom to modify, freedom to take on your own, freedom for many restrictions on how you would use the software and what that means for AI is still being defined. For what I do, I work at the Allen Institute for AI, we're a nonprofit, we want to make AI open for everybody and we try to lead on what we think is truly open-source. There's not full agreement in the community, but for us that means releasing the training data, releasing the training code, and then also having open weights like this. And we'll get into the details of the models and again and again as we try to get deeper into how the models were trained, we will say things like the data processing, data filtering data quality is the number one determinant of the model quality. And then a lot of the training code is the determinant on how long it takes to train and how fast your experimentation is. Without fully open-source models where you have access to this data, it is hard to know… Or it's harder to replicate. We'll get into cost numbers for DeepSeek-V3 on mostly GPU hours and how much you could pay to rent those

yourselves. But without the data, the replication cost is going to be far, far higher. And same goes for the code.

Lex Fridman
We should also say that this is probably one of the more open models out of the frontier models.

Nathan Lambert
Yes.

Lex Fridman
In this full spectrum where probably the fullest open-source, like you said, open code, open data, open weights, this is not open code, this is probably not open data and this is open weights and the licensing is MIT license or it's… There's some nuance in the different models, but it's towards the free… In terms of the open-source movement, these are the good guys.

Nathan Lambert
Yeah. DeepSeek is doing fantastic work for disseminating understanding of AI. Their papers are extremely detailed in what they do and for other teams around the world, they're very actionable in terms of improving your own training techniques. And we'll talk about licenses more, the DeepSeek-R1 model has a very permissive license. It's called the MIT license. That effectively means there's no downstream restrictions on commercial use, there's no use case restrictions. You can use the outputs from the models to create synthetic data. And this is all fantastic. I think the closest peer is something like Llama where you have the weights and you have a technical report. And the technical report is very good for Llama. One of the most read PDFs of the year last year is the Llama 3 paper, but in some ways it's slightly less actionable. It has less details on the training specifics. I think less plots and so on. And the Llama 3 license is more restrictive than MIT. And then between the DeepSeek custom license and the Llama license, we could get into this whole rabbit hole, I think. We'll make sure we want to go down the license rabbit hole before we do specifics.

Lex Fridman
It should be stated that one of the implications that DeepSeek, it puts pressure on Llama and everybody else on OpenAI to push towards open-source. And that's the other side of open-source is that you mentioned is how much is published in detail about it, so how open are you with the insights behind the code? How good is the technical reports? Are there hand wavy or is there actual details in there? And that's one of the things that DeepSeek did well is they published a lot of the details.

Nathan Lambert
Especially in the DeepSeek-V3, which is their pre-training paper. They were very clear that they are doing interventions on the technical stack that go at many different levels. For example, on their to get highly efficient training, they're making modifications at or below the CUDA layer for NVIDIA chips. I have never worked there myself and there are a few people in the world that do that very well, and some of them are at DeepSeek. These types of people are at DeepSeek and leading American frontier labs, but there are not many places.

Lex Fridman
To help people understand the other implication of open weights, just there's a topic we'll return to often here. There's a fear that China, the nation might have interest in stealing American data, violating privacy of American citizens. What can we say about open weights to help us understand what the weights are able to do in terms of stealing people's data?

Nathan Lambert
These weights that you can download from Hugging Face or other platforms are very big matrices of numbers. You can download them to a computer in your own house that has no internet and you can run this model and you're totally in control of your data. That is something that is different than how a lot of language model usage is actually done today, which is mostly through APIs where you send your prompt to GPUs run by certain companies. And these companies will have different distributions and policies on how your data is stored, if it is used to train future models, where it is stored, if it is encrypted, and so on. The open weights are you have your fate of data in your own hands, and that is something that is deeply connected to the soul of open-source.

Lex Fridman
It's not the model that steals your data, it's whoever is hosting the model, which could be China if you're using the DeepSeek app or it could be Perplexity. You're trusting them with your data or OpenAI, you're trusting them with your data. And some of these are American companies, some these are Chinese companies, but the model itself is not doing the stealing, it's the host. All right, so back to the basics. What's the difference between DeepSeek-V3 and DeepSeek-R1? Can we try to lay out the confusion potential?

Nathan Lambert
Yes. For one, I have very understanding of many people being confused by these two model names, so I would say the best way to think about this is that when training a language model, you have what is called pre-training, which is when you're predicting the large amounts of mostly internet text you're trying to predict the next token. And what to know about these new DeepSeek models is that they do this internet large scale pre-training once to get what is called DeepSeek-V3 base. This is a base model, it's just going to finish your sentences for you. It's going to be harder to work with than ChatGPT. And then what

DeepSeek did is they've done two different post-training regimes to make the models have specific desirable behaviors. What is the more normal model in terms of the last few years of AI, an instruct model, a chat model, a quote-unquote "aligned model", a helpful model. There are many ways to describe this is more standard post-training. This is things like instruction tuning, reinforcement learning from human feedback. We'll get into some of these words and this is what they did to create the DeepSeek-V3 model. This was the first model to be released and it is very high performant, it's competitive with GPT-4, Llama 405B and so on. And then when this release was happening, we don't know their exact timeline or soon after they were finishing the training of a different training process from the same next token prediction based model that I talked about, which is when this new reasoning training that people have heard about comes in in order to create the model that is called DeepSeek-R1. The R through this conversation is good for grounding for reasoning. And the name is also similar to OpenAI's o1, which is the other reasoning model that people have heard about. And we'll have to break down the training for R1 in more detail because for one we have a paper detailing it, but also it is a far newer set of techniques for the AI community, so it is a much more rapidly evolving area of research.

Lex Fridman
Maybe we should also say the big two categories of training of pre-training and post-training. These are umbrella terms that people use, so what is pre-training and what is post-training and what are the different flavors of things underneath the post-training umbrella?

Nathan Lambert
Pre-training, I'm using some of the same words to really get the message across is you're doing what is called autoregressive prediction to predict the next token in a series of documents. This is done over standard practice is trillions of tokens, so this is a ton of data that is mostly scraped from the web. And some of DeepSeek's earlier papers, they talk about their training data being distilled for math. I shouldn't use this word yet, but taken from Common Crawl and that's a public access that anyone listening to this could go download data from the Common Crawl website. This is a crawler that is maintained publicly. Yes, other tech companies eventually shift to their own crawler and DeepSeek likely has done this as well as most frontier labs do. But this sort of data is something that people can get started with and you're just predicting text in a series of documents. This can be scaled to be very efficient and there's a lot of numbers that are thrown around in AI training like how many floating-point operations or flops are used. And then you can also look at how many hours of these GPUs that are used. And it's largely one loss function taken to a very large amount of compute usage. You set up really efficient systems and then at the end of that you have the base model and pre-training is where there is a lot more of complexity in terms of how the process is emerging or evolving and the different types of training losses that you'll use. I think this is a lot of techniques grounded in the natural language processing literature. The oldest technique which is still used today is something

called instruction tuning or also known as supervised fine-tuning. These acronyms will be IFT or SFT. People really go back and forth throughout them, and I'll probably do the same, which is where you add this formatting to the model where it knows to take a question that is, explain the history of the Roman Empire to me or a sort of question you'll see on Reddit or Stack Overflow. And then the model will respond in a information-dense but presentable manner. The core of that formatting is in this instruction tuning phase. And then there's two other categories of loss functions that are being used today. One I'll classify as preference fine-tuning. Preference fine-tuning is a generalized term for what came out of reinforcement learning from human feedback, which is RLHF. This reinforcement learning from human feedback is credited as the technique that helped ChatGPT break through. It is a technique to make the responses that are nicely formatted like these Reddit answers more in tune with what a human would like to read. This is done by collecting pairwise preferences from actual humans out in the world to start and now AIs are also labeling this data and we'll get into those trade-offs. And you have this contrastive loss function between a good answer and a bad answer. And the model learns to pick up these trends. There's different implementation ways. You have things called reward models. You could have direct alignment algorithms. There's a lot of really specific things you can do, but all of this is about fine-tuning to human preferences. And the final stage is much newer and will link to what is done in R1 and these reasoning models is I think OpenAI's name for this, they had this new API in the fall, which they called the reinforcement fine-tuning API. This is the idea that you use the techniques of reinforcement learning, which is a whole framework of AI. There's a deep literature here to summarize, it's often known as trial and error learning or the subfield of AI where you're trying to make sequential decisions in a certain potentially noisy environment. There's a lot of ways we could go down that, but fine-tuning language models where they can generate an answer and then you check to see if the answer matches the true solution. For math or code you have an exactly correct answer for math, you can have unit tests for code. And what we're doing is we are checking the language model's work and we're giving it multiple opportunities on the same questions to see if it is right. And if you keep doing this, the models can learn to improve in verifiable domains to a great extent. It works really well. It's a newer technique in the academic literature. It's been used at frontier labs in the US that don't share every detail for multiple years. This is the idea of using reinforcement learning with language models and it has been taking off especially in this DeepSeek moment.

Lex Fridman
And we should say that there's a lot of exciting stuff going on again across the stack, but the post-training probably this year, there's going to be a lot of interesting developments in the post-training. We'll talk about it. I almost forgot to talk about the difference between DeepSeek-V3 and R1 on the user experience side. Forget the technical stuff, forget all of that, just people that don't know anything about AI, they show up. What's the actual experience, what's the use case for each one when they actually type and talk to it? What is each good at and that kind of thing?

Nathan Lambert
Let's start with DeepSeek-V3, again it's more people would tried something like it. You ask it a question, it'll start generating tokens very fast and those tokens will look like a very human legible answer. It'll be some sort of markdown list. It might have formatting to help you draw to the core details in the answer and it'll generate tens to hundreds of tokens. A token is normally a word for common words or a sub word part in a longer word, and it'll look like a very high quality Reddit or Stack Overflow answer. These models are really getting good at doing these across a wide variety of domains, I think. Even things that if you're an expert, things that are close to the fringe of knowledge, they will still be fairly good at, I think. Cutting edge AI topics that I do research on, these models are capable for study aid and they're regularly updated. Where this changes is with the DeepSeek- R1, what is called these reasoning models is when you see tokens coming from these models to start, it will be a large chain of thought process. We'll get back to chain of thought in a second, which looks like a lot of tokens where the model is explaining the problem. The model will often break down the problem and be like, okay, they asked me for this. Let's break down the problem. I'm going to need to do this. And you'll see all of this generating from the model. It'll come very fast in most user experiences. These APIs are very fast, so you'll see a lot of tokens, a lot of words show up really fast, it'll keep flowing on the screen and this is all the reasoning process. And then eventually the model will change its tone in R1 and it'll write the answer where it summarizes its reasoning process and writes a similar answer to the first types of model. But in DeepSeek's case, which is part of why this was so popular even outside the AI community, is that you can see how the language model is breaking down problems. And then you get this answer, on a technical side they train the model to do this specifically where they have a section which is reasoning, and then it generates a special token, which is probably hidden from the user most of the time, which says, okay, I'm starting the answer. The model is trained to do this two stage process on its own. If you use a similar model in say, OpenAI, OpenAI's user interface is trying to summarize this process for you nicely by showing the sections that the model is doing and it'll click through, it'll say breaking down the problem, making X calculation, cleaning the result, and then the answer will come for something like OpenAI.

Lex Fridman
Maybe it's useful here to go through an example of a DeepSeek-R1 reasoning.

Nathan Lambert
If you're looking at the screen here, what you'll see is a screenshot of the DeepSeek chat app, and at the top is thought for 157 seconds with the drop-down arrow. Underneath that if we were in an app that we were running, the drop-down arrow would have the reasoning.

Lex Fridman
In this case, the specific question which I'm philosophically/ pothead inclined, so this is asking DeepSeek-R1 for one truly novel insight about humans. And it reveals the reasoning

and basically the truly novel aspect is what's pushing the reasoning to constantly… The model asking itself, is this truly novel? It's actually challenging itself to be more novel, more counterintuitive, less cringe, I suppose. Some of the reasoning says… This is just snapshots. Alternatively, humans have a unique meta-emotion where they feel emotions about their own emotions e.g. feeling guilty about being angry. This recursive emotional layering creates complex motivational drives that don't exist in other animals. The insight is that human emotions are nested. It's reasoning through how humans feel emotions, it's reasoning about meta emotions.

Nathan Lambert
It's going to have pages and pages of this. It's almost too much to actually read, but it's nice to skim as it's coming.

Lex Fridman
It's a James Joyce like stream of consciousness. And then it goes, wait, the user wants something that's not seen anywhere else. Let me dig deeper. And consider the human ability to hold contradictory beliefs simultaneously. Cognitive dissonance is known, but perhaps the function is to allow flexible adaptation, so on and so forth. That really captures the public imagination that holy shit, this isn't intelligence/almost like an inkling of sentience because you're thinking through, you're self-reflecting, you're deliberating. And the final result of that after 157 seconds is, humans instinctively convert selfish desires into cooperative systems by collectively pretending abstract rules, money, laws, rights are real. These shared hallucinations act as quote "games" where competition is secretly redirected to benefit the group, turning conflict into society's fuel. Pretty profound.

Nathan Lambert
This is a potential digression, but a lot of people have found that these reasoning models can sometimes produce much more eloquent text. That is a at least interesting example I think depending on how open-minded you are, you find language models interesting or not, and there's a spectrum there.

Lex Fridman
We'll talk about different benchmarks and so on but some has just a vibe. That in itself is a, let's say quote "fire" tweet. If I'm trying to produce something where people are like, "Oh, shit." Okay, so that's a chance probably return to it more. How were they able to achieve such low cost on the training and the inference? Maybe you could talk to the training first.

Dylan Patel
There's two main techniques that they implemented that are probably the majority of their efficiency, and then there's a lot of implementation details that maybe we'll gloss over or get into later that contribute to it. But those two main things are, one is they went to a mixture of experts model, which we'll define in a second. And then the other thing is that they

invented this new technique called MLA, latent attention. Both of these are big deals. Mixture of experts is something that's been in the literature for a handful of years. And OpenAI with GPT-4 was the first one to productize a mixture of experts model. And what this means is when you look at the common models around that most people have been able to interact with that are open, think Llama. Llama is a dense model i.e. every single parameter or neuron is activated as you're going through the model for every single token you generate. Now, with a mixture of experts model, you don't do that. How does the human actually work? Is like, oh, well my visual cortex is active when I'm thinking about vision tasks and other things. My amygdala is when I'm scared. These different aspects of your brain are focused on different things. A mixture of experts, models attempts to approximate this to some extent. It's nowhere close to what a brain architecture is, but different portions of the model activate. You'll have a set number of experts in the model and a set number that are activated each time. And this dramatically reduces both your training and inference costs because now if you think about the parameter count as the total embedding space for all of this knowledge that you're compressing down during training, one, you're embedding this data in instead of having to activate every single parameter, every single time you're training or running inference, now you can just activate on a subset and the model will learn which expert to route to for different tasks. And so this is a humongous innovation in terms of, hey, I can continue to grow the total embedding space of parameters. And so DeepSeek's model is 600 something billion parameters, relative to Llama 405-B, it's 405 billion parameters, relative to Llama 70-B, it's 70 billion parameters. This model technically has more embedding space for information to compress all of the world's knowledge that's on the internet down. But at the same time, it is only activating around 37 billion of the parameters, so only 37 billion of these parameters actually need to be computed every single time you're training data or inferencing data out of it. Versus again, the Llama model, 70 billion parameters must be activated or 405 billion parameters must be activated, so you've dramatically reduced your compute cost when you're doing training and inference with this mixture of experts architecture.

Nathan Lambert
Should we break down where it actually applies and go into the transformer? Is that useful?

Lex Fridman
Let's go. Let's go into the transformer.

Nathan Lambert
The transformer is a thing that is talked about a lot, and we will not cover every detail. Essentially the transformer is built on repeated blocks of this attention mechanism and then a traditional dense fully connected multilayer perception, whatever word you want to use for your normal neural network. And you alternate these blocks. There's other details and where mixture of experts is applied is at this dense model. The dense model holds most of the weights if you count them in a transformer model, so you can get really big gains from

those mixture of experts on parameter efficiency at training and inference because you get this efficiency by not activating all of these parameters.

Lex Fridman
We should also say that a transformer is a giant neural network.

Nathan Lambert
Yeah.

Lex Fridman
And then there's, for 15 years now, there's what's called the deep learning revolution. Network's gotten larger and larger. At a certain point, the scaling laws appeared where people realized -

Dylan Patel
This is a scaling law shirt by the way.

Lex Fridman
Representing scaling laws. Where it became more and more formalized that bigger is better across multiple dimensions of what bigger means. But these are all neural networks we're talking about, and we're talking about different architectures of how to construct these neural networks such that the training and the inference on them is super efficient.

Nathan Lambert
Yeah. Every different type of model has a different scaling law for it, which is effectively for how much compute you put in the architecture will get to different levels of performance at test tasks. And mixture of experts is one of the ones at training time even if you don't consider the inference benefits, which are also big. At training time, your efficiency with your GPUs is dramatically improved by using this architecture if it is well implemented. You can get effectively the same performance model and evaluation scores with numbers like 30% less compute, I think. There's going to be a wide variation depending on your implementation details and stuff. But it is just important to realize that this type of technical innovation is something that gives huge gains. And I expect most companies that are serving their models to move to this mixture of experts implementation. Historically, the reason why not everyone might do it is because it's an implementation complexity, especially when doing these big models. This is one of the things that DeepSeek gets credit for is they do this extremely well. They do a mixture of experts extremely well. This architecture for what is called DeepSeek MoE, MoE is the shortened version of mixture of experts, is multiple papers old. This part of their training infrastructure is not new to these models alone. And same goes for what Dylan mentioned with multi-head latent attention. This is all about reducing memory usage during inference and same things during training by using some fancy low rank approximation math. If you get into the details with this latent

attention, it's one of those things I look at and it's like, okay, they're doing really complex implementations because there's other parts of language models such as embeddings that are used to extend the context length, the common one that DeepSeek used is rotary positional embeddings, which is called RoPE. And if you want to use RoPE with a normal MoE, it's a sequential thing, you take two of the attention matrices and you rotate them by a complex value rotation, which is a matrix multiplication. With DeepSeek's MLA, with this new attention architecture, they need to do some clever things because they're not set up the same and it just makes the implementation complexity much higher. They're managing all of these things, and these are probably the sort of things that OpenAI these closed labs are doing. We don't know if they're doing the exact same techniques, but they actually shared them with the world, which is really nice to be like, this is the cutting edge of efficient language model training.

Lex Fridman
And some of this requires low level engineering, just it is a giant mess in trickery. As I understand they went below CUDA, so they go super low programming of GPUs.

Dylan Patel
Effectively, Nvidia builds this library called NCCL, in which when you're training a model, you have all these communications between every single layer of the model, and you may have over a hundred layers.

Nathan Lambert
What does NCCL stand for? It's NCCL.

Dylan Patel
Nvidia Communications Collectives Library.

Lex Fridman
Nice. Damn.

Dylan Patel
And so when you're training a model, you're going to have all these allreducers and allgathers, between each layer, between the multilayer perceptron or feed-forward network and the attention mechanism, you'll have basically the model synchronized. Or you'll have allreduce and allgather. And this is a communication between all the GPUs in the network, whether it's in training or inference, so Nvidia has a standard library. This is one of the reasons why it's really difficult to use anyone else's hardware for training is because no one's really built a standard communications library. And Nvidia has done this at a sort of a higher level. DeepSeek because they have certain limitations around the GPUs that they have access to, the interconnects are limited to some extent by the restrictions of the GPUs that were shipped into China legally, not the ones that are smuggled but legally shipped in that

they used to train this model, they had to figure out how to get efficiencies. And one of those things is that instead of just calling the NVIDIA library NCCL, they scheduled their own communications, which some of the labs do. Meta talked about in Llama 3, how they made their own custom version of NCCL. They didn't talk about the implementation details. This is some of what they did, probably not as well as… Maybe not as well as DeepSeek because DeepSeek, necessity is the mother of innovation and they had to do this. OpenAI has people that do this sort of stuff, Anthropic, etc. But DeepSeek certainly did it publicly and they may have done it even better because they were gimped on a certain aspect of the chips that they have access to. And so they scheduled communications by scheduling specific SMs. SMs you could think of as the core on a GPU. There's hundreds of cores or there's a bit over a hundred cores SMs on a GPU. And they were specifically scheduling, hey, which ones are running the model? Which ones are doing allreduce? Which one are doing allgather? And they would flip back and forth between them. And this requires extremely low level programming.

Nathan Lambert
This is what NCCL does automatically or other Nvidia libraries handle this automatically usually.

Dylan Patel
Yeah, exactly. And so technically they're using PTX which is, you could think of it as an assembly type language. It's not exactly that or instruction set, like coding directly to assembly or instruction set. It's not exactly that, but that's still part of technically CUDA. But it's like, do I want to write in Python, PyTorch equivalent and call Nvidia libraries? Do I want to go down to the C level and code even lower level, or do I want to go all the way down to the assembly or ISO level? And there are cases where you go all the way down there at the very big labs, but most companies just do not do that because it's a waste of time and the efficiency gains you get are not worth it. But, it's a waste of time and the efficiency gains you get are not worth it. But DeepSeek's implementation is so complex, especially with their mixture of experts. People have done mixture of experts, but they're generally 8-16 experts and they activate 2. So, one of the words that we like to use is sparsity factor or usage. So, you might have 1/4th of your model activate, and that's what Mistral's Mixtral model, right? They're a model that really catapulted them to like, "Oh, my God. They're really, really good." OpenAI has also had models that are MoE and so have all the other labs that are major closed. But what DeepSeek did that maybe only the leading labs have only just started recently doing is have such a high sparsity factor, right? It's not 1/4th of the model, right? Two out of eight experts activating every time you go through the model, it's eight out of 256.

Nathan Lambert
And there's different implementations for mixture of experts where you can have some of these experts that are always activated, which this just looks like a small neural network,

and then all the tokens go through that and then they also go through some that are selected by this routing mechanism. And one of the innovations in DeepSeek's architecture is that they change the routing mechanism and mixture of expert models. There's something called an auxiliary loss, which effectively means during training, you want to make sure that all of these experts are used across the tasks that the model sees. Why there can be failures in mixture of experts is that when you're doing this training, one objective is token prediction accuracy. And if you just let turning go with a mixture of expert model on your own, it can be that the model learns to only use a subset of the experts. And in the MoE literature, there's something called the auxiliary loss which helps balance them. But if you think about the loss functions of deep learning, this even connects to The Bitter Lesson, is that you want to have the minimum inductive bias in your model to let the model learn maximally. And this auxiliary loss, this balancing across experts could be seen as intention with the prediction accuracy of the tokens. So we don't know the exact extent that the DeepSeek MoE change, which is instead of doing an auxiliary loss, they have an extra parameter in their routing, which after the batches, they update this parameter to make sure that the next batches all have a similar use of experts. And this type of change can be big, it can be small, but they add up over time. And this is the sort of thing that just points to them innovating. And I'm sure all the labs that are training big MoEs are looking at this sort of things, which is getting away from the auxiliary loss. Some of them might already use it, but you keep accumulating gains. And we'll talk about the philosophy of training and how you organize these organizations. And a lot of it is just compounding small improvements over time in your data, in your architecture, in your post-training and how they integrate with each other. DeepSeek does the same thing and some of them are shared, or a lot. We have to take them on face value that they share their most important details. I mean, the architecture and the weights are out there, so we're seeing what they're doing and it adds up.

Dylan Patel

Going back to the efficiency and complexity point, right? It's 32 versus a 4, right, for Mixtral and other MoE models that have been publicly released? So this ratio is extremely high. And what Nathan was getting at there was when you have such a different level of sparsity, you can't just have every GPU have the entire model, right? The model's too big, there's too much complexity there. So you have to split up the model with different types of parallelism, right? And so you might have different experts on different GPU nodes, but now what happens when this set of data that you get, "Hey, all of it looks like this one way and all of it should route to one part of my model." So when all of it routes to one part of the model, then you can have this overloading of a certain set of the GPU resources or a certain set of the GPUs and then the rest of the training network sits idle because all of the tokens are just routing to that. So this is the biggest complexity, one of the big complexities with running a very sparse mixture of experts model i.e., this 32 ratio versus this four ratio, is that you end up with so many of the experts just sitting there idle. So how do I load balance between them? How do I schedule the communications between them? This is a lot of the extremely low-level, detailed work that they figured out in the public first, and potentially second or third in the world and maybe even first in some cases.

Lex Fridman
What lesson do you, in the direction of The Bitter Lesson do you take from all of this? Is this going to be the direction where a lot of the gain is going to be, which is this kind of low-level optimization or is this a short-term thing where the biggest gains will be more on the algorithmic high-level side of post-training? Is this a short-term leap because they've figured out a hack because constraints necessitate the mother of invention or is there still a lot of gains?

Nathan Lambert
I think we should summarize what The Bitter Lesson actually is about, is that The Bitter Lesson essentially, if you paraphrase it, is that the types of training that will win out in deep learning as we go are those methods that which are scalable in learning and search, is what it calls out. The scale word gets a lot of attention in this. The interpretation that I use is effectively to avoid adding the human priors to your learning process. And if you read the original essay, this is what it talks about is how researchers will try to come up with clever solutions to their specific problem that might get them small gains in the short term while simply enabling these deep learning systems to work efficiently, and for these bigger problems in the long term might be more likely to scale and continue to drive success. And therefore, we were talking about relatively small implementation changes to the mixture of experts model. And therefore it's like, "Okay, we will need a few more years to know if one of these were actually really crucial to The Bitter Lesson," but The Bitter Lesson is really this long-term arc of how simplicity can often win. And there's a lot of sayings in the industry, "The models just want to learn. You have to give them the simple loss landscape where you put compute through the model and they will learn, and getting barriers out of the way."

Lex Fridman
That's where the power of something like nickel comes in, where standardized code that could be used by a lot of people to create simple innovations that can scale, which is why the hacks, I imagine, the code base for DeepSeek is probably a giant mess.

Nathan Lambert
I'm sure DeepSeek definitely has code bases that are extremely messy, where they're testing these new ideas. Multi-head latent attention probably could start in something like a Jupyter Notebook, or somebody tries something on a few GPUs and that is really messy. But the stuff that trains the DeepSeek V3 and DeepSeek-R1, those libraries, if you were to present them to us, I would guess are extremely high-quality code.

Lex Fridman
So, high-quality, readable code. Yeah.

Dylan Patel
I think there is one aspect to note though is that there is the general ability for that to transfer across different types of runs. You may make really, really high-quality code for one specific model architecture at one size, and then that is not transferable to, "Hey, when I make this architecture tweak, everything's broken again," right? That's something that could be with their specific low-level coding of scheduling SMs is specific to this model architecture and size. Whereas, Nvidia's Collectives Library is more like, "Hey, it'll work for anything," right? "You want to do an allreduce? Great, I don't care what your model architecture is, it'll work," and you're giving up a lot of performance when you do that in many cases, but it's worthwhile for them to do the specific optimization for the specific run given the constraints that they have regarding compute.

Lex Fridman
I wonder how stressful it is to these frontier models, like initiate training to have the code -

Dylan Patel
Push the button.

Lex Fridman
- to push the button that you're now spending a large amount of money and time to train this. I mean, there must be a lot of innovation on the debugging stage of making sure there's no issues, that you're monitoring and visualizing every aspect of the training, all that kind of stuff.

Dylan Patel
When people are training, they have all these various dashboards, but the most simple one is your loss, right? And it continues to go down, but in reality, especially with more complicated stuff like MoE, the biggest problem with it, or FP8 training, which is another innovation, going to a lower precision number format i.e., less accurate is that you end up with loss spikes. And no one knows why the loss spike happened. And for a long -

Nathan Lambert
Some of them, you do.

Dylan Patel
Some of them, you do.

Nathan Lambert
Some of them are bad data. Can I give Ai2's example of what blew up our earlier models is a Subreddit called microwavegang. We love to shout this out. It's a real thing. You can pull up microwavegang. Essentially it's a Subreddit where everybody makes posts that are just the

letter M. So it's like, mmm. So there's extremely long sequences of the letter M and then the comments are like beep beep because it's in the micro events.

Dylan Patel
Yeah.

Nathan Lambert
But if you pass this into a model that's trained to be a normal producing text, it's extremely high-loss because normally you see an M, you don't predict Ms for a long time. So this is something that caused loss spikes for us. But when you have much … This is old, this is not recent. And when you have more mature data systems, that's not the thing that causes the loss spike. And what Dylan is saying is true, but it's levels to this sort of idea.

Dylan Patel
With regards to the stress, these people are like … You'll go out to dinner with a friend that works at one of these labs and they'll just be looking at their phone every 10 minutes and they're not … You know, it's one thing if they're texting, but they're just like, "Is the loss … Is the loss spike okay?"

Nathan Lambert
Yeah. It's like tokens per second. Loss not blown up. They're just watching this.

Lex Fridman
And the heart rate goes up if there's a spike.

Dylan Patel
And some level of spikes is normal, it'll recover and be back. Sometimes a lot of the old strategy was like, you just stop the run, restart from the old version and then change the data mix and then it keeps going.

Nathan Lambert
There are even different types of spikes. So Dirk Groeneveld has a theory today too, that's like fast spikes and slow spikes, where there are, sometimes where you're looking at the loss and there are other parameters, you could see it start to creep up and then blow up, and that's really hard to recover from. So you have to go back much further. So you have the stressful period where it's flat or it might start going up and you're like, "What do I do?" Whereas, there are also loss spikes that are, it looks good and then there's one spiky data point. And what you could do is you just skip those. You see that there's a spike. You're like, "Okay, I can ignore this data. Don't update the model and do the next one, and it'll recover quickly." But on trickier implementations, so as you get more complex in your architecture and you scale up to more GPUs, you have more potential for your loss blowing up. So it's like, there's a distribution.

Dylan Patel
And then the whole idea of grokking also comes in, right? It's like, just because it slowed down from improving in loss doesn't mean it's not learning because all of a sudden it could be like this and it could just spike down in loss again because it truly learned something, right? And it took some time for it to learn that. It's not a gradual process, and that's what humans are like. That's what models are like. So it's really a stressful task, as you mentioned.

Lex Fridman
And the whole time the dollar count is going up.

Nathan Lambert
Every company has failed runs. You need failed run to push the envelope on your infrastructure. So, a lot of news cycles are made of X company had Y failed run. Every company that's trying to push the frontier of AI has these. So yes, it's noteworthy because it's a lot of money and it can be week to a month setback, but it is part of the process.

Lex Fridman
But if you're DeepSeek, how do you get to a place where holy shit, there's a successful combination of hyperparameters?

Nathan Lambert
A lot of small failed runs.

Lex Fridman
So, rapid iteration through failed runs until -

Nathan Lambert
And successful ones.

Lex Fridman
And then you build up some intuition, like this mixture of expert works and then this implementation of MLA works.

Nathan Lambert
Key hyperparameters, like learning rate and regularization and things like this, and you find the regime that works for your code base. Talking to people at Frontier Labs, there's a story that you can tell where training language models is kind of a path that you need to follow. So you need to unlock the ability to train a certain type of model or a certain scale, and then your code base and your internal know-how of which hyperparameters work for IT is kind of known. And you look at the DeepSeek papers and models, they've scaled up, they've added complexity, and it's just continuing to build the capabilities that they have.

Dylan Patel
There's the concept of a YOLO run. So YOLO, you only live once.

Lex Fridman
Yep.

Dylan Patel
What it is, is there's all this experimentation you do at the small scale, research ablations. You have your Jupyter Notebook where you're experimenting with MLA on three GPUs or whatever and you're doing all these different things like, "Hey, do I do four active experts, 128 experts? Do I arrange the experts this way?" All these different model architecture things, you're testing at a very small scale. Right? A couple of researchers, few GPUs, tens of GPUs, hundreds of GPUs, whatever it is. And then all of a sudden you're like, "Okay, guys. No more fucking around. No more screwing around. Everyone, take all the resources we have. Let's pick what we think will work and just go for it. YOLO." And this is where that sort of stress comes in is like, "Well, I know it works here, but some things that work here don't work here. And some things that work here don't work down here in this terms of scale." So it's really truly a YOLO run. And there's this discussion of certain researchers just have this methodical nature. They can find the whole search space and figure out all the ablations of different research and really see what is best. And there's certain researchers who just have that innate gut instinct of like, "This is the YOLO run. I'm looking at the data. I think this is it."

Nathan Lambert
This is why you want to work in post-training because the GPU cost for training is lower. So you can make a higher percentage of your training runs YOLO runs.

Lex Fridman
Yeah.

Dylan Patel
For now.

Lex Fridman
Yeah, for now.

Nathan Lambert
For now. For now.

Lex Fridman
So some of this is fundamentally luck, still.

Dylan Patel
Luck is skill, right, in many cases?

Lex Fridman
Yeah. I mean, it looks lucky, right, when you're -

Nathan Lambert
But the hill to climb, if you're on one of these labs, you have an evaluation you're not crushing, there's a repeated playbook of how you improve things. There are localized improvements, which might be data improvements. And these add up into the whole model just being much better. And when you zoom in really close, it can be really obvious that this model is just really bad at this thing and we can fix it and you just add these up. So some of it feels like luck, but on the ground, especially with these new reasoning models we're talking to is just so many ways that we could poke around. And normally, it's that some of them give big improvements.

Dylan Patel
The search space is near infinite and yet the amount of compute and time you have is very low, and you have to hit release schedules. You have to not get blown past by everyone. Otherwise, what happened with DeepSeek crushing Meta and Mistral and Cohere and all these guys, they moved too slow. They maybe were too methodical. I don't know, they didn't hit the YOLO run. Whatever the reason was, maybe they weren't as skilled. Whatever, you can call it luck if you want, but at the end of the day, it's skill.

Lex Fridman
So 2025 is the year of the YOLO run. It seems like all the labs are going in.

Dylan Patel
I think it's even more impressive what OpenAI did in 2022. At the time, no one believed in mixture of experts models at Google who had all the researchers. OpenAI had such little compute and they devoted all of their compute for many months, all of it, 100% for many months to GPT-4 with a brand-new architecture with no belief that, "Hey, let me spend a couple of hundred million dollars, which is all of the money I have on this model." That is truly YOLO.

Lex Fridman
Yeah.

Dylan Patel
Right?

Lex Fridman
Yeah.

Dylan Patel
Now people have all these training run failures that are in the media, right? It's like, "Okay, great, but actually a huge chunk of my GPUs are doing inference. I still have a bunch doing research constantly. And yes, my biggest cluster is training, but on this YOLO run," but that YOLO run is much less risky than what OpenAI did in 2022, or maybe what DeepSeek did now or sort of like, "Hey, we're just going to throw everything at it."

Lex Fridman
The big winners throughout human history are the ones who are willing to do YOLO at some point. Okay. What do we understand about the hardware it's been trained on, DeepSeek?

Dylan Patel
DeepSeek is very interesting. This is where a second could take to zoom out, out of who they are first of all, right? High-Flyer is a hedge fund that has historically done quantitative trading in China as well as elsewhere. And they have always had a significant number of GPUs, right? In the past, a lot of these high-frequency trading, algorithmic quant traders used FPGAs, but it shifted to GPUs definitely. And there's both, but GPUs especially. And High-Flyer, which is the hedge fund that owns DeepSeek, and everyone who works for DeepSeek is part of High-Flyer to some extent. Same parent company, same owner, same CEO, they had all these resources and infrastructure for trading, and then they devoted a humongous portion of them to training models, both language models and otherwise, because these techniques were heavily AI-influenced. More recently, people have realized, "Hey, trading with …" Even when you go back to Renaissance and all these quantitative firms, natural language processing is the key to trading really fast, understanding a press release and making the right trade. And so DeepSeek has always been really good at this. And even as far back as 2021, they have press releases and papers saying, "Hey, we're the first company in China with an A100 cluster this large." It was 10,000 A100 GPUs, right? This is in 2021. Now, this wasn't all for training large language models. This was mostly for training models for their quantitative aspects, quantitative trading as well as a lot of that was natural language processing, to be clear. Right? And so this is the sort of history, right? So verifiable fact is that in 2021, they built the largest cluster, at least they claim it was the largest cluster in China, 10,000 GPUs.

Nathan Lambert
Before export controls started.

Dylan Patel
Yeah.

Nathan Lambert
It's like they've had a huge cluster before any conversation of export controls.

Dylan Patel
So then you step it forward to, what have they done over the last four years since then? Obviously, they've continued to operate the hedge fund, probably make tons of money. And the other thing is that they've leaned more and more and more into AI. The CEO, Lian Chingfeng ... Lian-

Nathan Lambert
You're not putting me on the spot on this. We discussed this before.

Dylan Patel
Lian Feng, right, the CEO, he owns maybe ... Lian Feng, he owns maybe a little bit more than half the company allegedly, is an extremely Elon, Jensen kind of figure where he's just involved in everything. Right? And so over that time period, he's gotten really in depth into AI. He actually has a bit of a, if you see some of his statements, a bit of an IAK vibe almost, right?

Nathan Lambert
Total AGI vibes, like, "We need to do this. We need to make a new ecosystem of OpenAI. We need China to lead on this sort of ecosystem because historically, the western countries have led on software ecosystems." And straight up acknowledges, "In order to do this, we need to do something different." DeepSeek is his way of doing this. Some of the translated interviews with him are fantastic.

Lex Fridman
So he has done interviews?

Nathan Lambert
Yeah.

Lex Fridman
Do you think you would do a western interview, or no, or is there controls on the channel?

Nathan Lambert
There hasn't been one yet, but I would try it.

Lex Fridman
Okay. All right. Well, I just got a Chinese translator, so it was great. This is a push. So fascinating figure, engineer pushing full on into AI, leveraging the success from the high-frequency trading.

Nathan Lambert
Very direct quotes. "We will not switch to closed source," when asked about this stuff. Very long-term motivated in how the ecosystem of AI should work. And I think from a Chinese perspective, he wants a Chinese company to build this vision.

Dylan Patel
And so this is sort of like the "visionary behind the company." This hedge fund still exists, this quantitative firm. And so DeepSeek is the sort of … Slowly, he got turned to this full view of AI, everything about this, but at some point it slowly maneuvered and he made DeepSeek. And DeepSeek has done multiple models since then. They've acquired more and more GPUs. They share infrastructure with the fund. Right? And so there is no exact number of public GPU resources that they have. But besides this 10,000 GPUs that they bought in 2021, and they were fantastically profitable, and then this paper claims they did only 2,000 H800 GPUs, which are a restricted GPU that was previously allowed in China, but no longer allowed. And there's a new version, but it's basically Nvidia's H100 for China. And there's some restrictions on it specifically around the communications sort of speed, the interconnect speed, which is why they had to do this crazy SM scheduling stuff. So going back to that, it's like this is obviously not true in terms of their total GPU count.

Lex Fridman
Obvious available GPUs, but for this training run, you think 2,000 is the correct number, or no?

Dylan Patel
So this is where it takes a significant amount of zoning in. What do you call your training run, right? You count all of the research and ablations that you ran, right? Picking all this stuff because yes, you can do a YOLO run, but at some level you have to do the test at the small scale and then you have to do some test at medium scale before you go to a large scale.

Nathan Lambert
Accepted practice is that for any given model that is a notable advancement, you're going to do two to 4x compute of the full training run in experiments alone.

Lex Fridman
So a lot of this compute that's being scaled up is probably used in large part at this time for research?

Dylan Patel
Yeah. And research begets the new ideas that lets you get huge efficiency.

Nathan Lambert
Research gets you o1. Research gets you breakthroughs and you need to bet on it.

Lex Fridman
So some of the pricing strategy that we'll discuss has the research baked into the price?

Dylan Patel
So the numbers that DeepSeek specifically said publicly are just the 10,000 GPUs in 2021 and then 2,000 GPUs for only the pre-training for V3. They did not discuss cost on R1. They did not discuss cost on all the other RL for the instruct model that they made. They only discussed the pre-training for the base model and they did not discuss anything on research and ablations. And they do not talk about any of the resources that are shared in terms of, "Hey, the fund is using all these GPUs," right? And we know that they're very profitable and they had 10,000 GPUs in 2021. So, some of the research that we've found is that we actually believe they have closer to 50,000 GPUs.

Lex Fridman
We as semi-analysis. So we should say that you're sort of one of the world experts in figuring out what everybody's doing in terms of the semiconductor, in terms of cluster buildouts, in terms of who is doing what in terms of training runs. So yeah, that's the we. Okay, go ahead.

Dylan Patel
Yeah, sorry. We believe they actually have something closer to 50,000 GPUs, right? Now this is split across many tasks, right? Again, the fund, research and ablations.

Nathan Lambert
For ballpark, how much would OpenAI or Anthropic had. I think the clearest example we have, because Meta is also open, they talk about order of 60k to 100k H100 equivalent GPUs in their training clusters.

Dylan Patel
Right. So Llama 3, they trained on 16,000 H100s, but the company of Meta last year publicly disclosed they bought 400 something thousand GPUs.

Nathan Lambert
Yeah.

Dylan Patel
Right? So of course, tiny percentage on the training. Again, most of it is serving me the best Instagram Reels or whatever.

Nathan Lambert
I mean, we could get into a cost of, what is the cost of ownership for a 2,000 GPU cluster, 10,000? There's just different sizes of companies that can afford these things and DeepSeek is reasonably big. Their compute allocation is one of the top few in the world that's not OpenAI, Anthropic, etc, but they have a lot of compute.

Lex Fridman
Can you in gentlemen actually just zoom out and also talk about the Hopper architecture, the Nvidia Hopper GPU architecture and the difference between H100 and H800, like you mentioned, the interconnects?

Dylan Patel
Yeah. So there's, Ampere was the A100 and then H100 Hopper, right? People use them synonymously in the U.S. because really there's just H100 and now there's H200, right, but same thing mostly? In China, there've been different salvos of expert restrictions. So initially, the U.S. government limited on a two-factor scale, which is chip interconnect versus FLOPs. So any chip that had interconnects above a certain level and FLOPs above a certain … Floating point operations above a certain level was restricted. Later, the government realized that this was a flaw in the restriction and they cut it down to just floating point operations. And so -

Nathan Lambert
H800 had high FLOPs, low communication?

Dylan Patel
Exactly. So, the H800 was the same performance as H100 on FLOPs, but it just had the interconnect bandwidth cut. DeepSeek knew how to utilize this. "Hey, even though we're cut back on the interconnect, we can do all this fancy stuff to figure out how to use the GPU fully anyways." And so that was back in October 2022. But later in 2023, into 2023 implemented in 2024, the U.S. government banned the H800. Right? And so by the way, this H800 cluster, these 2,000 GPUs was not even purchased in 2024. It was purchased in late 2023. And they're just getting the model out now because it takes a lot of research, etc. H800 was banned and now there's a new chip called the H20. The H20 is cut back on only FLOPs, but the interconnect bandwidth is the same. And in fact, in some ways it's better than the H100 because it has better memory bandwidth and memory capacity. So Nvidia is working within the constraints of what the government sets and then builds the best possible GPU for China.

Lex Fridman
Can we take this actual tangent and we'll return back to the hardware, is the philosophy, the motivation, the case for export controls? What is it? Dario Amodei just published a blog post about export controls. The case he makes is that if AI becomes super powerful and he says

by 2026, we'll have AGI or super powerful AI and that's going to give a significant ... Whoever builds that will have a significant military advantage. And so because The United States is a democracy and as he says, China is authoritarian or has authoritarian elements, you want a unipolar world where the super powerful military, because of the AI is one that's a democracy. It's a much more complicated world geopolitically when you have two superpowers with super powerful AI and one is authoritarian. So, that's the case he makes. And so the United States wants to use export controls to slow down, to make sure that China can't do these gigantic training runs that will be presumably required to build the AGI.

Nathan Lambert
This is very abstract. I think this can be the goal of how some people describe export controls, is this super powerful AI. And you touched on the training run idea. There's not many worlds where China cannot train AI models. I think export controls are decapping the amount of compute or the density of compute that China can have. And if you think about the AI ecosystem right now, as all of these AI companies, revenue numbers are up and to the right. Their AI usage is just continuing to grow, more GPUs are going to inference. A large part of export controls, if they work is just that the amount of AI that can be run in China is going to be much lower. So on the training side, DeepSeek V3 is a great example, which you have a very focused team that can still get to the frontier of AI on ... This 2,000 GPUs is not that hard to get all considering in the world. They're still going to have those GPUs. They're still going to be able to train models. But if there's going to be a huge market for AI, if you have strong export controls and you want to have 100,000 GPUs just serving the equivalent of ChatGPT clusters with good export controls, it also just makes it so that AI can be used much less. And I think that is a much easier goal to achieve than trying to debate on what AGI is. And if you have these extremely intelligent autonomous AIs and data centers, those are the things that could be running in these GPU clusters in the United States, but not in China.

Dylan Patel
To some extent, training a model does effectively nothing. They have a model. The thing that Dario is sort of speaking to is the implementation of that model, once trained to then create huge economic growth, huge increases in military capabilities, huge increases in productivity of people, betterment of lives. Whatever you want to direct super powerful AI towards, you can, but that requires a significant amounts of compute. And so the U.S. government has effectively said ... And forever, training will always be a portion of the total compute. We mentioned Meta's 400,000 GPUs. Only 16,000 made Llama. Right? So the percentage that Meta's dedicating to inference, now this might be for recommendation systems that are trying to hack our mind into spending more time and watching more ads, or if it's for a super powerful AI that's doing productive things, it doesn't matter about the exact use that our economic system decides. It's that, that can be delivered in whatever way we want. Whereas with China, you know, your expert restrictions, great. You're never going

to be able to cut everything off. And I think that's quite a well-understood by the U.S. government is that you can't cut everything off.

Nathan Lambert
And they'll make their own chips.

Dylan Patel
And they're trying to make their own chips. They'll be worse than ours, but the whole point is to just keep a gap. And therefore at some point, as the AI … In a world where two, 3% economic growth, this is really dumb by the way, to cut off high-tech and not make money off of it. But in a world where super powerful AI comes about and then starts creating significant changes in society, which is what all the AI leaders and big tech companies believe. I think super powerful AI is going to change society massively. And therefore, this compounding effect of the difference in compute is really important. There's some sci-fi out there where AI is measured in how much power is delivered to compute, right, or how much is being … That's sort of a way of thinking about what's the economic output, is just how much power are you directing towards that AI?

Nathan Lambert
Should we talk about reasoning models with this, as a way that this might be actionable as something that people can actually see? So, the reasoning models that are coming out with R1 and o1, they're designed to use more compute. There's a lot of buzzy words in the AI community about this, test-time compute, inference time compute, whatever. But Dylan has good research on this. You can get to the specific numbers on the ratio of when you train a model, you can look at things. It's about the amount of compute used at training and amount of compute used at inference. These reasoning models are making inference way more important to doing complex tasks. In the fall in December, OpenAI announced this o3 model. There's another thing in AI, when things move fast, we get both announcements and releases. Announcements are essentially blog posts where you pat yourself on the back and you say you did things and releases are when the model's out there, the paper's out there, etc. So OpenAI has announced o3. I mean, we can check if o3-mini is out as of recording potentially, but that doesn't really change the point, which is that the breakthrough result was something called ARC-AGI task, which is the abstract reasoning corpus, a task for artificial general intelligence. François Chollet is the guy who's been … It's a multi-year-old paper. It's a brilliant benchmark. And the number for OpenAI o3 to solve this was that it used some sort of number of samples in the API. The API has thinking effort and number of samples. They used 1,000 samples to solve this task and it comes out to be five to $20 per question, which you're putting in effectively a math puzzle. And then it takes orders of dollars to answer one question, and this is a lot of compute. If those are going to take off in the U.S., OpenAI needs a ton of GPUs on inference to capture this. They have this OpenAI ChatGPT Pro subscription, which is $200 a month -

Dylan Patel
Which Sam said they're losing money on.

Nathan Lambert
Which means that people are burning a lot of GPUs on inference. And I've signed up with it, I've played with it. I don't think I'm a power user, but I use it. And it's like, that is the thing that a Chinese company with mediumly strong expert controls, there will always be loopholes, might not be able to do it all. And if the main result for o3 is also a spectacular coding performance, and if that feeds back into AI companies being able to experiment better.

Lex Fridman
So presumably, the idea is for an AGI, a much larger fraction of the compute would be used for this test-time compute, for the reasoning, for the AGI goes into a room and thinks about how to take over the world and come back in 2.7 hours -

Nathan Lambert
This is what -

Lex Fridman
... and that it's going to take a lot of compute.

Nathan Lambert
This is what people, CEO or leaders of OpenAI and Anthropic talk about, is autonomous AI models, which is you give them a task and they work on it in the background. I think my personal definition of AGI is much simpler. I think language models are a form of AGI and all of this super powerful stuff is a next step that's great if we get these tools. But a language model has so much value in so many domains that it's a general intelligence to me. But this next step of agentic things where they're independent and they can do tasks that aren't in the training data is what the few-year outlook that these AI companies are driving for.

Lex Fridman
I think the terminology here that Dario uses is super powerful AI. So I agree with you on the AGI. I think we already have something like that's exceptionally impressive that Alan Turing would for sure say is AGI, but he's referring more to something once in possession of, then you would have a significant military and geopolitical advantage over other nations. So it's not just like you can ask it how to cook an omelet.

Nathan Lambert
And he has a much more positive view. And as I say, machines of love and grace. I read into this and I don't have enough background in physical sciences to gauge exactly how competent I am, and if AI can revolutionize biology. I am safe saying that AI is going to accelerate the progress of any computational science.

Lex Fridman
So we're doing a depth-first search here on topics, taking tangent of a tangent, so let's continue on that depth-first search. You said that you're both feeling the AGI. What's your timeline? Dario is 2026 for the super powerful AI that's basically agentic to a degree where it's a real security threat, that level of AGI. What's your timeline?

Nathan Lambert
I don't like to attribute specific abilities because predicting specific abilities and when is very hard. I think mostly if you're going to say that I'm feeling the AGI is that I expect continued, rapid, surprising progress over the next few years. So, something like R1 is less surprising to me from DeepSeek because I expect there to be new paradigms versus … surprising to me from DeepSeek because I expect there to be new paradigms where substantial progress can be made. I think DeepSeek-R1 is so unsettling because we're kind of on this path with ChatGPT. It's like it's getting better, it's getting better, it's getting better, and then we have a new direction for changing the models, and we took one step like this and we took a step-up. So it looks like a really fast slope, and then we're going to just take more steps. So it's just really unsettling when you have these big steps, and I expect that to keep happening. I've tried opening Operator, I've tried Claude computer use, they're not there yet. I understand the idea, but it's just so hard to predict what is the breakthrough that'll make something like that work. And I think it's more likely that we have breakthroughs that work in things that we don't know what they're going to do. So everyone wants agents. Dario has a very eloquent way of describing this, and I just think that it's like there's going to be more than that, so just expect these things to come.

Lex Fridman
I'm going to have to try to pin you down to a date on the AGI timeline. Like the nuclear weapon moment, so moment where on the geopolitical stage, there's a real… Because we're talking about export controls, when do you think, just even to throw out a date, when do you think that would be? For me, it's probably after 2030, so I'm not as -

Nathan Lambert
That's what I would say.

Dylan Patel
So define that. Because to me, it kind of almost has already happened. You look at elections in India and Pakistan, people get AI voice calls and think they're talking to the politician. The AI diffusion rules, which was enacted in the last couple of weeks of the Biden admin, it looks like the Trump admin will keep and potentially even strengthen, limit cloud computing and GPU sales to countries that are not even related to China. It's like this is -

Nathan Lambert
Portugal and all these normal countries are on the… You need approval from the US list.

Dylan Patel
Yeah, Portugal and all these countries that are allies. Singapore. They freaking have F-35s and we don't let them buy GPUs. This to me is already to the scale of…

Lex Fridman
Well, that just means that the US military is really nervous about this new technology. That doesn't mean that technology is already there. So they might be just very cautious about this thing that they don't quite understand. But that's a really good point. The robocalls, swarms of semi-intelligent bots could be a weapon, could be doing a lot of social engineering.

Dylan Patel
I mean, there's tons of talk about from the 2016 elections like Cambridge Analytica and all this stuff, Russian influence. I mean, every country in the world is pushing stuff onto the internet and has narratives they want. Every technically competent, whether it's Russia, China, US, Israel, etc. People are pushing viewpoints onto the internet en masse. And language models crash the cost of very intelligent sounding language.

Nathan Lambert
There's some research that shows that the distribution is actually the limiting factor. So language models haven't yet made misinformation particularly change the equation there. The internet is still ongoing. I think there's a blog, AI Snake Oil and some of my friends at Princeton that write on this stuff. So there is research. It's a default that everyone assumes. And I would've thought the same thing, is that misinformation doesn't get far worse with language models. I think in terms of internet posts and things that people have been measuring, it hasn't been a exponential increase or something extremely measurable and things you're talking about with voice calls and stuff like that, it could be in modalities that are harder to measure. So it's something that it's too soon to tell in terms of… I think that's political instability via the web is very… It's monitored by a lot of researchers to see what's happening. I think that… You're asking about the AGI thing. If you're making me give a year, I'm going to be like, "Okay, I have AI CEOs saying this. They've been saying two years for a while. I think that there are people like Dario at Anthropic, the CEO, has thought about this so deeply. I need to take their word seriously, but also understand that they have different incentives." So I would be like, "Add a few years to that." Which is how you get something similar to 2030 or a little after 2030.

Dylan Patel
I think to some extent, we have capabilities that hit a certain point where any one person could say, "Oh, okay, if I can leverage those capabilities for X amount of time, this is AGI, call it '27, '28." But then the cost of actually operating that capability-

Nathan Lambert
Yeah, this was going to be my point.

Dylan Patel
… is so, so extreme that no one can actually deploy it at scale en masse to actually completely revolutionize the economy on a snap of a finger. So I don't think it will be a snap of the finger moment.

Nathan Lambert
It's a physical constraint [inaudible 01:14:37].

Dylan Patel
Rather, it'll be a, "Oh, the capabilities are here, but I can't deploy it everywhere." And so one simple example, going back sort of to 2023 was when being when GPT-4 came out, everyone was freaking out about search. Perplexity came out. If you did the cost on like, hey, implementing GPT-3 into every Google search, it was like, oh, okay, this is just physically impossible to implement. And as we step forward to going back to the test-time compute thing, a query for… You ask ChatGPT a question, it costs cents for their most capable model of Chat to get a query back. To solve an AGI problem though costs 5 to 20 bucks, and this is in-

Nathan Lambert
It's only going up from there.

Dylan Patel
This is 1,000-10,000x factor difference in cost to respond to a query versus do a task. And the task of AGI is not like it's like… It's simple, to some extent, but it's also like, what are the tasks that we want… Okay, AGI, "What we have today", can do AGI. Three years from now, it can do much more complicated problems, but the cost is going to be measured in thousands and thousands and hundreds of thousands of dollars of GPU time, and there just won't be enough power, GPUs, infrastructure to operate this and therefore shift everything in the world on the snap the finger. But at that moment, who gets to control and point the AGI at a task? And so this was in Dario's post that he's like, "Hey, China can effectively and more quickly than us, point their AGI at military tasks." And they have been, in many ways, faster at adopting certain new technologies into their military, especially with regards to drones. The US maybe has a long-standing large air sort of fighter jet type of thing, bombers. But when it comes to asymmetric arms such as drones, they've completely leapfrogged the US and the West. And the fear that Dario is sort of pointing out there, I think, is that, yeah, great, we'll have AGI in the commercial sector. The US military won't be able to implement it superfast. Chinese military could and they could direct all their resources to implementing it in the military, and therefore solving military logistics or solving some other aspect of disinformation for targeted certain set of people so they can

flip a country's politics or something like that that is actually catastrophic versus the US just wants to… Because it'll be more capitalistically allocated just towards whatever is the highest return on income, which might be building factories better or whatever.

Lex Fridman
So everything I've seen, people's intuition seems to fail on robotics. So you have this kind of general optimism. I've seen this on self-driving cars. People think it's much easier problem than it is. Similar with drones, here, I understand it a little bit less, but I've just seen the reality of the war in Ukraine and the usage of drones on both sides. And it seems that humans still far outperform any fully autonomous systems. AI is an assistant, but humans drive. FPV drones where the human's controlling most of it, just far, far, far outperforms AI systems. So I think it's not obvious to me that we're going to have swarms of autonomous robots anytime soon in the military context. Maybe the fastest I can imagine is 2030, which is why I said 2030 for the super powerful AI. Whenever you have large scale swarms of robots doing military actions, that's when the world just starts to look different to me. So that's the thing I'm really worried about. But there could be cyber war, cyber war type of technologies that from social engineering to actually just swarms of robots that find attack vectors in our code bases and shut down power grids, that kind of stuff. And it could be one of those things like on any given weekend or something, power goes out, nobody knows why, and the world changes forever. Just power going out for two days in all of the United States, that will lead to murder, to chaos. But going back to export controls, do you see that as a useful way to control the balance of power geopolitically in the context of AI?

Dylan Patel
And I think going back to my viewpoint is if you believe we're in this sort of stage of economic growth and change that we've been in for the last 20 years, the export controls are absolutely guaranteeing that China will win long-term. If you do not believe AI is going to make significant changes to society in the next 10 years or 5 years. Five-year timelines are sort of what the more executives and such of AI companies and even big tech companies believe. But even 10-year timelines, it's reasonable. But once you get to, hey, these timelines are below that time period, then the only way to create a sizable advantage or disadvantage for America versus China is if you constrain and compute, because talent is not really something that's constraining. China arguably has more talent, more STEM graduates, more programmers. The US can draw upon the world's people, which it does. There's tons of foreigners in the AI industry.

Nathan Lambert
So many of these AI teams are all people without a US passport.

Dylan Patel
Yeah. I mean, many of them are Chinese people who are moving to America, and that's great. That's exactly what we want. But that talent is one aspect, but I don't think that's one

that is a measurable advantage for the US or not. It truly is just whether or not compute. Now, even on the compute side, when we look at chips versus data centers, China has the unprecedented ability to build ridiculous sums of power. Clockwork. They're always building more and more power. They've got steel mills that individually are the size of the entire US industry. And they've got aluminum mills that consume gigawatts and gigawatts of power. And when we talk about what's the biggest data center, OpenAI made this huge thing about Stargate, their announcement there, once it's fully built out in a few years, it'll be two gigawatts of power. And this is still smaller than the largest industrial facilities in China. China, if they wanted to build the largest data center in the world, if they had access to the chips, could. So it's just a question of when, not if.

Lex Fridman
So their industrial capacity far exceeds the United States'?

Dylan Patel
Exactly.

Lex Fridman
They manufacture stuff. So long-term, they're going to be manufacturing chips there?

Dylan Patel
Chips are a little bit more specialized. I'm specifically referring to the data centers. Fabs take huge amounts of power, don't get me wrong. That's not necessarily the gating factor there. The gating factor on how fast people can build the largest clusters today in the US is power. Now, it could be power generation, power transmission, substations, and all these sorts of transformers and all these things building the data center. These are all constraints on the US industry's ability to build larger and larger training systems, as well as deploying more and more inference compute.

Nathan Lambert
I think we need to make a point clear on why the time is now for people that don't think about this, because essentially, with export controls, you're making it so China cannot make or get cutting edge chips. And the idea is that if you time this wrong, China is pouring a ton of money into their chip production, and if you time it wrong, they are going to have more capacity for production, more capacity for energy, and figure out how to make the chips and have more capacity than the rest of the world to make the chips. Because everybody can buy… They're going to sell their Chinese chips to everybody, they might subsidize them. And therefore, if AI takes a long time to become differentiated, we've kneecapped the financial performance of American companies. NVIDIA can sell less, TSMC cannot sell to China. So therefore, we have less demand to therefore… To keep driving the production cycle. So that's the assumption behind the timing being [inaudible 01:22:43].

Dylan Patel
Less than 10 years or 5 years to above. China will win because of these restrictions long-term, unless AI does something in the short-term, which I believe AI will make massive changes to society in the medium, short-term. And so that's the big unlocker there. And even today, if Xi Jinping decided to get "scale-pilled", IE, decide that scaling laws are what matters, just like the US executives like Satya Nadella and Mark Zuckerberg and Sundar and all these US executives of the biggest, most powerful tech companies have decided they're scale-pilled and they're building multi-gigawatt data centers, whether it's in Texas or Louisiana or Wisconsin, wherever it is, they're building these massive things that cost as much as their entire budget for spending on data centers globally in one spot. This is what they've committed to for next year, year after, etc. And so they're so convinced that this is the way that this is what they're doing. But if China decided to, they could do it faster than us, but this is where the restrictions come in. It is not clear that China as a whole has decided from the highest levels that this is a priority. The US sort of has. You see Trump talking about DeepSeek and Stargate within the same week. And the Biden admin as well had a lot of discussions about AI and such. It's clear that they think about it. Only just last week did DeepSeek meet the second in command of China. They have not even met the top, they haven't met Xi, Xi hasn't set down, and they only just released a subsidy of a trillion RMB, roughly $160 billion, which is closer to the spending of Microsoft and Meta and Google combined for this year. So they're realizing it just now. But that's where these export restrictions come in and say, "Hey, you can't ship the most powerful US chips to China. You can ship a cut-down version. You can't ship the most powerful chips to all these countries who we know are just going to rent it to China. You have to limit the numbers."

Nathan Lambert
And the tools.

Dylan Patel
And same with manufacturing [inaudible 01:24:52] tools, all these different aspects, but it all stems from AI and then what downstream can slow them down in AI. And so the entire semiconductor restrictions, you read them, they're very clear, it's about AI and military civil fusion of technology. It's very clear. And then from there it goes, oh, well, we're banning them from buying lithography tools and etch tools and deposition tools. And oh, this random subsystem from a random company that's tiny. Why are we banning this? Because all of it, the US government has decided is critical to AI systems.

Nathan Lambert
I think the fulcrum point is the transition from seven nanometer to five nanometer chips where I think it was Huawei that had the seven nanometer chip a few years ago, which caused another political brouhaha, almost like this moment. And then it's the ASML deep UV. What is that… Extreme ultraviolet lithography.

Dylan Patel
Just set context on the chips. What Nathan's referring to is in 2020, Huawei released their Ascend 910 chip, which was an AI chip, first one on seven nanometer before Google did, before NVIDIA did. And they submitted it to the MLPerf benchmark, which is sort of a industry standard for machine learning performance benchmark, and it did quite well, and it was the best chip at the submission. This was a huge deal. The Trump admin, of course, banned, it was 2019, banned the Huawei from getting seven nanometer chips from TSMC. And so then they had to switch to using internal, domestically produced chips, which was a multi-year setback.

Nathan Lambert
Many companies have done seven nanometer chips. And the question is we don't know how much Huawei was subsidizing production of that chip. Intel has made seven nanometer chips that are not profitable and things like this. So this is how it all feeds back into the economic engine of export controls.

Lex Fridman
Well, so you're saying that for now, Xi Jinping has not felt the AGI, but it feels like the DeepSeek moment, there might be meetings going on now where he's going to start wearing the same t-shirt and things are going to escalate.

Dylan Patel
I mean, he may have woken up last week. Liang Feng met the second command guy, and they had a meeting, and then the next day, they announced the AI subsidies, which are a trillion RMB.

Lex Fridman
So it's possible that this DeepSeek moment is truly the beginning of a cold war.

Nathan Lambert
That's what a lot of people are worried about. People in AI have been worried that this is going towards a cold war or already is.

Lex Fridman
But it's not DeepSeek's fault, but there's something, a bunch of factors came together where -

Nathan Lambert
It's how history works.

Lex Fridman
- it's like this explosion. I mean, it all has to do with NVIDIA's not going down properly, but it's just some [inaudible 01:27:28] mass hysteria that happened that eventually led to Xi Jinping having meetings and waking up to this idea.

Dylan Patel
And the US government realized in October 7th, 2022, before ChatGPT released, that restriction on October 7th, which dropped and shocked everyone, and it was very clearly aimed at AI. Everyone was like, "What the heck are you doing?"

Nathan Lambert
Stable Diffusion was out then, but not ChatGPT.

Dylan Patel
Yeah, but not ChatGPT.

Nathan Lambert
So it was starting to be rumblings -

Dylan Patel
Of what GenAI can do to society, but it was very clear, I think, to at least National Security Council and those sort of folks, that this was where the world is headed, this cold war that's happening.

Lex Fridman
So is there any concerns that the export controls push China to take military action on Taiwan?

Dylan Patel
This is the big risk. The further you push China away from having access to cutting edge American and global technologies, the more likely they are to say, "Well, because I can't access it, I might as well... No one should access it." And there's a few interesting aspects of that. China has a urban-rural divide like no other. They have a male-female birth ratio like no other to the point where if you look in most of China, it's like the ratio is not that bad. But when you look at single dudes in rural China, it's like a 30:1 ratio. And those are disenfranchised dudes. "The US has an incel problem." China does too, it's just they're placated in some way or crushed down. What do you do with these people? And at the same time, you're not allowed to access the most important technology, at least the US thinks so. China's maybe starting to think this is the most important technology by starting to dump subsidies in it. They thought EVs and renewables were the most important technology. They dominate that now. Now, they started thinking about semiconductors in the late 2010s and early 2020s and now they've been dumping money and they're catching up rapidly and

they're going to do the same with AI because they're very talented. So the question is, when does this hit a breaking point? And if China sees this as, "Hey, they can continue…" If not having access and starting a true hot war, taking over Taiwan or trying to subvert its democracy in some way or blockading it hurts the rest of the world far more than it hurts them, this is something they could potentially do. And so is this pushing them towards that? Potentially. I'm not quite a geopolitical person, but it's obvious that the world regime of peace and trade is super awesome for economics, but at some point, it could break.

Nathan Lambert
I think we should comment the why Chinese economy would be hurt by that is that they're export heavy, I think. United States buys so much. If that goes away, that's how their economy [inaudible 01:30:17].

Dylan Patel
Well, also, they just would not be able to import raw materials from all over the world. The US would just shut down the Strait of Malacca. And at the same time, the US entire… You could argue almost all the GDP growth in America since the '70s has been either population growth or tech, because your life today is not that much better than someone from the '80s outside of tech. Cars, they all have semiconductors in them everywhere. Fridges, semiconductors everywhere. There's these funny stories about how Russians were taking apart laundry machines because they had certain Texas Instrument chips that they could then repurpose and put into their anti-missile missile things, like their S-400 or whatever. You would know more about this, but there's all sorts of… Everything about semiconductors is so integral to every part of our lives.

Lex Fridman
So can you explain the role of TSMC in the story of semiconductors and maybe also how the United States can break the reliance on TSMC?

Dylan Patel
I don't think it's necessarily breaking the reliance. I think it's getting TSMC to build in the US. So taking a step back, TSMC produces most of the world's chips, especially on the foundry side. There's a lot of companies that build their own chips. Samsung, Intel, STMicro, Texas Instruments, Analog Devices, all these kinds of companies build their own chips, and XP, but more and more of these companies are outsourcing to TSMC and have been for multiple decades.

Lex Fridman
Can you explain the supply chain there and where most of TSMC is in terms of manufacturing?

Dylan Patel
Sure. So historically, supply chain was companies would build their own chips. It would be a company started, they'd build their own chips, and then they'd design the chip and build the chip and sell it. Over time, this became really difficult because the cost of building a fab continues to compound every single generation. Of course, figuring out the technology for it is incredibly difficult regardless, but just the dollars and cents that are required, ignoring, saying, "Hey, yes, I have all the technical capability." Which it's really hard to get that by the way. Intel's failing, Samsung's failing, etc. But if you look at just the dollars to spend to build that next-generation fab, it keeps growing. Sort of Moore's law is having the cost of chips every two years. There's a separate law that's sort of doubling the cost of fabs every handful of years. And so you look at a leading-edge fab that is going to be profitable today, that's building three nanometer chips or two nanometer chips in the future, that's going to cost north of 30, $40 billion. And that's just for a token amount. That's like the base building blocking. You probably need to build multiple. And so when you look at the industry over the last, if I go back 20, 30 years ago, there were 20, 30 companies that could build the most advanced chips, and then they would design them themselves and sell them. So companies like AMD would build their own chips. Intel, of course, still builds their own chips. They're very famous for it. IBM would build their own chips. And you could just keep going down the list. All these companies built their own chips. Slowly, they kept falling like flies, and that's because of what TSMC did. They created the Foundry business model, which is, I'm not going to design any chips. I'm just going to contract manufacturer chips for other people. And one of their early customers is NVIDIA. NVIDIA is the only semiconductor company doing more than $1 billion of revenue that was started in the era of foundry. Every other company started before then, and at some point had fabs, which is actually incredible. Like AMD, and Intel, and Broadcom - everyone had fabs at some point - or some companies like Broadcom. It was like a merger amalgamation of various companies that rolled up. But even today, Broadcom has fabs. They build iPhone, RF radio chips in Colorado for Apple. All these companies had fabs, and for most of the fabs, they threw them away or sold them off, or they got rolled into something else. And now, everyone relies on TSMC. Including Intel, their latest PC chip uses TSMC chips. It also uses some Intel chips, but it uses TSMC process.

Lex Fridman
Can you explain why the foundry model is so successful for these companies? Why are they going with -

Nathan Lambert
Economies of scale.

Lex Fridman
Scale?

Dylan Patel
Yeah. So I mean, like I mentioned, the cost of building a fab is so high, the R&D is so difficult. And when you look at these companies that had their own vertical stack, there was an antiquated process of like, okay, I'm so hyper customized to each specific chip, but as we've gone through the history of the last 50 years of electronics and semiconductors, A, you need more and more specialization because Moore's law has died, Dennard Scaling has died, IE, Chips are not getting better just for free from manufacturing. You have to make real architectural innovations. Google is not just running on Intel CPUs for web serving. They have a YouTube chip, they have TPUs, they have Pixel chips, they have a wide diversity of chips that generate all the economic value of Google. It's running all the services and stuff. And this is just Google. And you could go across any company in the industry, and it's like this. Cars contain 5,000 chips, 200 different varieties of them. All these random things. A Tesla door handle has two chips. It's ridiculous. And it's a cool door handle. You don't think about it, but it has two really chip, penny chips in there. Anyways, so as you have more diversity of chips, as you have more specialization required and the cost of fabs continues to grow, you need someone who is laser focused on building the best process technology and making it as flexible as possible.

Nathan Lambert
I think you could say it simply, which is the cost per fab goes up, and if you are a small player that makes a few types of chips, you're not going to have the demand to pay back the cost of the fab. Whereas NVIDIA can have many different customers and aggregate all this demand into one place, and then they're the only person that makes enough money building chips to build the next fab. So this is kind of why the companies slowly get killed because they have, 10 years ago, a chip that is profitable and is good enough, but the cost to build the next one goes up. They may try to do this, fail because they don't have the money to make it work, and then they don't have any chips, or they build it and it's too expensive and they just sort of have not profitable chips.

Dylan Patel
There's more failure points. You could have one little process related to some sort of chemical etch or some sort of plasma etch or some little process that screws up, you didn't engineer it right, and now the whole company falls apart, you can't make chips. And so super, super powerful companies like Intel, they had the weathering storm to like, hey, they still exist today, even though they really screwed up their manufacturing six, seven years ago. But in the case of like AMD, they almost went bankrupt, they had to sell their fabs to Mubadala, UAE, and that became a separate company called Global Foundries, which is a foundry firm. And then AMD was able to then focus on the return back up, was like, "Hey, let's focus on making chiplets and a bunch of different chips for different markets and focusing on specific workloads rather than all of these different things." And so you get more diversity of chips, you have more companies than ever designing chips, but you have fewer companies than ever manufacturing them. And this is where TSMC comes in, is they've just

been the best. They are so good at it. They're customer focused, they make it easy for you to fabricate your chips. They take all of that complexity and kind of try and abstract a lot of it away from you. They make good money. They don't make insane money, but they make good money and they're able to aggregate all this demand and continue to build the next fab, the next fab, the next fab.

Lex Fridman
So why is Taiwan so special for TSMC? Why is it happening there? Can it be replicated inside the United States?

Dylan Patel
Yeah, so there's aspects of it that I would say yes, and aspects that I'd say no. TSMC is way ahead because former executive Morris Chang of Texas Instruments wasn't promoted to CEO. And he was like, "Screw this. I'm going to go make my own chip company." And he went to Taiwan and made TSMC. And there's a whole lot more story there. Texas Instruments, could have have been TSMC, but Texas Semiconductor Manufacturing instead of Texas Instruments. So there is that whole story there. But the -

Nathan Lambert
Sitting here in Texas.

Lex Fridman
And that sounds like a human story. He didn't get promoted.

Dylan Patel
Just the brilliance of Morris Chang, which I wouldn't underplay, but there's also a different level of how this works. So in Taiwan, the top percent of graduates of students that go to the best school, which is NTU, the top percent of those all go work to TSMC. And guess what their pay is? Their starting pay is like $80,000, $70,000, which is like that's starting pay for a good graduate in the US, not the top. The graduates are making hundreds of thousands of dollars at the Googles and the Amazons, and now I guess the OpenAIs of the world. So there is a large dichotomy of what is the top 1% of the society doing and where are they headed because of economic reasons? Intel never paid that crazy good. And it didn't make sense to them. That's one aspect. Where's the best going? Second is the work ethic. We like to work. You work a lot, we work a lot, but at the end of the day, what does the time and amount of work that you're doing and what does a fab require? Fabs are not work from home jobs. They are you go into the fab and grueling work. There's hey, if there is any amount of vibration, an earthquake happens, vibrates the machines, they're either broken, you've scrapped some of your production. And then in many cases, they're not calibrated properly. So when there's an earthquake, recently, there's been a earthquake, TSMC doesn't call their employees, they just go to the fab and they just show up. The parking lot gets slammed, and people just go

into the fab and fix it. It's like ants. It's like a hive of ants doesn't get told by the queen what to do. The ants just know.

Nathan Lambert
It's like one person just specializes on these one task, and it's like you're going to take this one tool and you're the best person in the world, and this is what you're going to do for your whole life is this one task in the fab.

Dylan Patel
Which is some special chemistry plus nanomanufacturing on one line of tools that continues to get iterated and yeah, it's like a specific plasma etch for removing silicon dioxide. That's all you focus on your whole career, and it's such a specialized thing. And so it's not like the tasks are transferable. AI today is awesome because people can pick it up like that. Semiconductor manufacturing is very antiquated and difficult. None of the materials are online for people to read easily and learn. The papers are very dense, and it takes a lot of experience to learn. And so it makes the barrier to entry much higher too. So when you talk about, hey, you have all these people that are super specialized, they will work 80 hours a week in a factory, in a fab, and if anything goes wrong, they'll go show up in the middle of the night because some earthquake, their wife's like, "There was an earthquake." He's like, "Great, I'm going to go to the fab."

Nathan Lambert
Time to cry.

Dylan Patel
Would you, as an American, do that? It's like these sorts of things are, I guess are the exemplifying why TSMC is so amazing. Now, can you replicate it in the US? Let's not ignore Intel was the leader in manufacturing for over 20 years. They brought every technology to market first besides the EUV. Strained silicon, high-K metal gates, FinFET, the list goes on and on and on of technologies that Intel brought to market first made the most money from and manufactured at scale first, best, highest profit margins. We shouldn't ignore that Intel can't do this. It's that the culture has broken. You've invested in the wrong things. They said no to the iPhone. They had all these different things regarding mismanagement of the fabs and mismanagement of designs, this lockup. And at the same time, all these brilliant people, these 50,000 PhDs or masters that have been working on specific chemical or physical processes or nanomanufacturing processes for decades, in Oregon, they're still there, they're still producing amazing work. It's just getting it to the last mile of production at high yield where you can manufacture dozens and hundreds of different kinds of chips, and good customer experience has broken. It's that customer experience. Part of it is people will say, Intel was too pompous in the 2000s, 2010s. They just thought they were better than everyone. The tool guys were like, "Oh, I don't think that this is mature enough." And they're like, "Ah, you just don't know. We know." This sort of stuff would happen. And so

can the US bring leading-edge semiconductor manufacturing to the US? [inaudible 01:42:44] yes. And we are. It's happening.

Nathan Lambert
Arizona is getting better and better as time goes on.

Dylan Patel
TSMC has built roughly 20% of their capacity for five nanometer in the US. Now, this is nowhere near enough. 20% of capacity in the US is like nothing. And furthermore, this is still dependent on Taiwan existing. There's sort of important way to separate it out. There's R&D and there's high volume manufacturing. Effectively, there are three places in the world that are doing leading-edge R&D. There's Hsinchu, Taiwan, there's Hillsboro, Oregon, and there is Pyongyang, South Korea. These three places are doing the leading-edge R&D for the rest of the world's leading-edge semiconductors. Now, manufacturing can be distributed more globally. And this is sort of where this dichotomy exists of who's actually modifying the process, who's actually developing the next generation one, who's improving them is Hsinchu, is Hillsboro, is Pyongyang. It is not the rest of these fabs like Arizona. Arizona is a paperweight. If Hsinchu disappeared off the face of the planet, within a year, couple years, Arizona would stop producing too. It's actually pretty critical. One of the things I like to say is if I had a few missiles, I know exactly where I could cause the most economic damage. It's not targeting the White House.

Lex Fridman
It's the R&D centers.

Dylan Patel
It's the R&D centers for TSMC, Intel, Samsung. And then some of the memory guys, Micron and Hynix.

Lex Fridman
Because they define the future evolution of these semiconductors, and everything's moving so rapidly that it really is fundamentally about R&D. And it is all about TSMC. Huh.

Dylan Patel
And so TSMC, you cannot purchase a vehicle without TSMC chips. You cannot purchase a fridge without TSMC chips. I think one of the few things you can purchase ironically, is a Texas Instruments graphing calculator because they actually manufacture in Texas. But outside of that, a laptop, a phone.

Lex Fridman
It's depressing.

Dylan Patel
Servers, GPUs, none of this stuff can exist. And this is without TSMC. And in many cases, it's not even the leading-edge sexy five nanometer chip, three nanometer chip, two nanometer chip. Oftentimes, it's just some stupid power IC that's converting from some voltage to another, and it's … I see that's converting from some voltage to another, and it's made at TSMC. It's like –

Nathan Lambert
This is what China is investing in as well. It's like, they can build out this long-tail fab where the techniques are much more known, you don't have to figure out these problems with EUV. They're investing in this and then they have large supply for things like the car door handles and the random stuff. And that trickles down into this whole economic discussion as well, which is they have far more than we do. And having supply for things like this is crucial to normal life.

Lex Fridman
So they're starting to invest in high-volume manufacturer, but they're not doing R&D as much?

Nathan Lambert
They are.

Dylan Patel
They do R&D on their own, they're just way behind. I would say, in 2015 China had a five-year plan where they defined by 2025 and 2020 certain goals, including 80% domestic production of semiconductors. They're not going to hit that, to be clear. But they are in certain areas really, really close. BYD is probably going to be the first company in the world to not have to use TSMC for making … because they have their own fabs for making chips. Now they still have to buy some chips from foreign, for example, around like self-driving ADAS capabilities because those are really high-end, but at least … A internal combustion engine has 40 chips in an EV, just for controlling flow rates and all these things, and EVs are even more complicated. So all these different power ICs and battery management controllers and all these things, they're insourcing. And this is something that China has been doing since 2015. Now, as far as the trailing edge, they're getting so much capacity there. As far as the leading edge, i.e. this five nanometer and so on and so forth, where GPUs, they are still behind. The US restrictions are trying to stop them in the latter, but all that's happened is yes, they've slowed down their five nanometer, three nanometer, etc, but they've accelerated their, hey, 45 nanometer, 90 nanometer power IC or analog IC or random chip in my keyboard, that kind of stuff. So there is an angle of, the US' actions, from the angle of the expert controls, have been so inflammatory at slowing down China's progress on the leading edge that they've turned around and have accelerated their progress elsewhere because they know that this is so important. If the US is going to lock them out

here, “what if they lock us out here as well in the trailing edge?” And so going back, can the US build it here? Yes, but it’s going to take a ton of money. I truly think to revolutionize and completely in-source semiconductors would take a decade and a trillion dollars.

Lex Fridman
Is some of it also culture, like you said, extreme competence, extreme work ethic in Taiwan?

Nathan Lambert
I think if you have the demand and the money is on the line, the American companies figure it out. It’s going to take handholding with the government, but I think that the culture helps TSMC break through and it’s easier for them. You [inaudible 01:47:50].

Dylan Patel
TSMC has some like 90,000 employees. It’s not actually that insane amount. The Arizona fab has 3,000 from Taiwan. And these people, their wives were like, “Yeah, we’re not going to have kids unless you sign up for the Arizona Fab. We go to Arizona and we have our kids there.” There’s also a Japan fab where the same thing happened. And so these wives drove these dudes to go to Japan or America to have the kids there. And it’s an element of culture, yeah, sure. Taiwan works that hard. But also, like the US has done it in the past, they could do it now. We can just import, I say import, the best people in the world if we want to.

Lex Fridman
That’s where the immigration conversation is a tricky one and there’s been a lot of debate over that. But yeah, it seems absurdly controversial to import the best people in the world. I don’t understand why it’s controversial. That’s one of the ways of winning.

Nathan Lambert
I’m sure we agree with you.

Dylan Patel
And even if you can’t import those people, I still think you could do a lot to manufacture most of it in the US, if the money’s there.

Nathan Lambert
It’s just way more expensive. It’s not profitable for a long time.

Dylan Patel
And that’s the context of the Chips Act is only $50 billion, relative to some of the renewable initiatives that were passed in the Inflation Reduction Act and the Infrastructure Act, which total in the hundreds of billions of dollars. And so the amount of money that the US is spending on the semiconductor industry is nothing, whereas all these other countries have

structural advantages in terms of work ethic and amount of work and things like that, but also a number of STEM graduates, the percentile of their best going to that. But they also have differences in terms of, hey, there's just tax benefits in the law and have been in the law for 20 years. And then some countries have massive subsidies. China has something like $200 billion of semiconductor subsidies a year. We're talking about $50 billion in the US over like six. So the girth or difference in the subsidy amounts is also huge. And so I think Trump has been talking about tariffing Taiwan recently. That's one of these things that's like, "Oh, okay, well, maybe he doesn't want to subsidize the US semiconductor industry." Obviously tariffing Taiwan is going to cost a lot of things to get much more expensive, but does it change the equation for TSMC building more fabs in the US? That's what he's positing.

Lex Fridman
So we laid out the importance ... By the way, it's incredible how much you know about so much.

Nathan Lambert
We told you Dylan knows all this stuff.

Lex Fridman
Yeah. Okay. You laid out why TSMC is really important. If we look out into the future 10-20 years out, US-China relationship, it seems like it can go to a dark place of Cold War, escalated Cold War, even hot war, or to a good place of anything from frenemies, to cooperation, to working together. So in this game theory, complicated game, what are the different trajectories? What should US be doing? What do you see as the different possible trajectories of US-China relations as both leaders start to feel the AGI more and more and see the importance of chips and the importance of AI.

Nathan Lambert
I mean, ultimately the export controls are pointing towards a separate future economy. I think the US has made it clear to Chinese leaders that we intend to control this technology at whatever cost to global economic integration. And it's hard to unwind that. The card has been played.

Dylan Patel
To the same extent they've also limited US companies from entering China. So it's been a long time coming. At some point there was a convergence, but over at least the last decade it's been branching further and further out. US companies can't enter China. Chinese companies can't enter the US. The US is saying, "Hey, China, you can't get access to our technologies in certain areas." And China's rebuttaling with the same thing around ... they've done some sort of specific materials in gallium and things like that that they've tried to limit the US on. There's a US drone company that's not allowed to buy batteries and they have

military customers. And this drone company just tells the military customers, “Hey, just get it from Amazon because I can’t actually physically get them.” There’s all these things that are happening that point to further and further divergence. I have zero idea, and I would love if we could all hold hands and sing Kumbaya, but I have zero idea how that could possibly happen.

Lex Fridman
Is the divergence good or bad for avoiding war? Is it possible that the divergence in terms of manufacturer chips of training AI systems is actually good for avoiding military conflict?

Dylan Patel
It’s an objective fact that the world has been the most peaceful it’s ever been when there are global hegemons, or regional hegemons in historical context. The Mediterranean was the most peaceful ever when the Romans were there. China had very peaceful and warring times, and the peaceful times were when dynasties had a lock hold over, not just themselves, but all their tributaries around them. And likewise, the most peaceful time in human history has been when the US was the global hegemon, the last decades. Now we’ve seen things start to slide with Russia, Ukraine, with what’s going on in the Middle East, and Taiwan risk, all these different things are starting to bubble up. Still objectively extremely peaceful. Now what happens when it’s not one global hegemon but it’s two, obviously … And China will be competitive or even overtake the US, it’s possible. And so this change in global hegemony, I don’t think it ever happens super peacefully. When empires fall, which is a possible trajectory for America, they don’t fall gracefully. They don’t just slide out of irrelevance. Usually there’s a lot of shaking. And so what the US is trying to do is maintain its top position, and what China is trying to do is become the top position. And obviously there’s butting of heads here, in the most simple terms.

Lex Fridman
And that could take shape in all kinds of ways, including proxy wars. And now-

Nathan Lambert
Yeah, it seems like it’s already happening. As much as I want there to be centuries of prolonged peace, it looks like further instability internationally is ahead.

Dylan Patel
And the US’ current task is, “Hey, if we control AI, if we’re the leader in AI and AI significantly accelerates progress, then we can maintain the global hegemony position.” And therefore -

Nathan Lambert
I hope that works.

Dylan Patel
And as an American, like, okay, I guess that's going to lead to peace for us. Now obviously other people around the world get affected negatively. Obviously the Chinese people are not going to be in as advantageous of a position if that happens, but this is the reality of what's being done and the actions that are being carried out.

Lex Fridman
Can we go back to the specific detail of the different hardware? There's this nice graphic in the export controls of which GPUs are allowed to be exported and which are not. Can you explain the difference? From a technical perspective, are the H20s promising?

Dylan Patel
Yeah. And I think we need to dive really deep into the reasoning aspect and what's going on there. The US has gone through multiple iterations of the export controls. This H800 was at one point allowed back in '23, but then it got canceled and by then DeepSeek had already built their cluster of, they claim, 2K. I think they actually have many more, something like 10K of those. And now this H20 is the legally allowed chip. Nvidia shipped a million of these last year to China. For context, it was four or five million GPUs. So the percentage of GPUs that were this China-specific H20 is quite high, roughly 20%, 25%, 20% or so. And so this H20 has been neutered in one way, but it's actually upgraded in other ways. And you could think of chips along three axes for AI, ignoring software stack and exact architecture, just raw specifications. There's floating point operations, FLOPS. There is memory bandwidth, i.e. in-memory capacity, IO memory. And then there is interconnect, chip-to-chip interconnections. All three of these are incredibly important for making AI systems. Because AI systems involve a lot of compute, they involve a lot of moving memory around, whether it be to memory or too other chips. And so these three vectors, the US initially had two of these vectors controlled and one of them not controlled, which was FLOPS and interconnect bandwidth were initially controlled. And then they said, "No, no, no, no. We're going to remove the interconnect bandwidth and just make it a very simple, only FLOPS." But now Nvidia can now make a chip that has … okay, it's cut down on FLOPS, so one-third that of the H100 on spec sheet paper performance for FLOPs. In real world it's closer to half or maybe even 60% of it. But then on the other two vectors, it's just as good for interconnect bandwidth. And then for memory bandwidth and memory capacity, the H20 has more memory bandwidth and more memory capacity than the H100. Now recently we, at our research, we cut Nvidia's production for H20 for this year down drastically. They were going to make another two million of those this year, but they just canceled all the orders a couple of weeks ago. In our view that's because we think that they think they're going to get restricted, because why would they cancel all these orders for H20? Because they shipped a million of them last year, they had orders in for a couple million this year, and just gone right. For H20, B20, a successor to H20, and now they're all gone. Now why would they do this? I think it's very clear, the H20 is actually better for certain tasks. And that certain task is reasoning. Reasoning is incredibly different than … When you look at the different

regimes of models. Pre-training is all about FLOPS, it's all about FLOPS. There's things you do, like Mixture of Experts that we talked about, to trade off interconnect or to trade off other aspects and lower the FLOPS and rely more on interconnect and memory. But at the end of the day, FLOPS is everything. We talk about models in terms of how many FLOPS they are. So we talk about, oh, GPT-4 is 2e25. Two to the 25th, 25 zeros FLOP, floating point operations for training. And we're talking about the restrictions for the 2e24, or 25, whatever. The US has an executive order that Trump recently unsigned, which was, hey, 1e26, once you hit that number of floating point operations, you must notify the government and you must share your results with us. There's a level of model where the US government must be told, and that's 1e26. And so as we move forward, this is an incredibly important … FLOP is the vector that the government has cared about historically, but the other two vectors are arguably just as important. And especially when we come to this new paradigm, which the world is only just learning about over the last six months: reasoning.

Lex Fridman
And do we understand firmly which of the three dimensions is best for reasoning? So interconnect, the FLOPS don't matter as much, is it memory?

Nathan Lambert
Memory. Yeah. We're going to get into technical stuff real fast.

Dylan Patel
I would say there's two articles in this one that I could show maybe graphics that might be interesting for you to pull up.

Lex Fridman
For the listeners, we're looking at the section of o1 inference architectures tokenomics.

Dylan Patel
You want to explain KV cache before we talk about this? I think it's better to -

Nathan Lambert
Okay. Yeah, we need to go through a lot of specific technical things, transformers, to make this easy for people.

Dylan Patel
Because it's incredibly important because this changes how models work. But I think resetting, why is memory so important? It's because so far we've talked about parameter counts and Mixture of Experts, you can change how many active parameters versus total parameters to embed more data but have less FLOPS. But more important, another aspect of what's part of this humongous revolution in the last handful of years is the transformer and the attention mechanism. Attention mechanism is that the model understands the

relationships between all the words in its context, and that is separate from the parameters themselves. And that is something that you must calculate. How each token, each word in the context length, is relatively connected to each other. And I think, Nathan, you can explain KV cache better.

Lex Fridman
KV cache is one of the optimization techniques for LLMs?

Nathan Lambert
So the attention operator has three core things, it's queries, keys, and values. QKV is the thing that goes into this. You'll look at the equation. You see that these matrices are multiplied together. These words, query, key and value, come from information retrieval backgrounds where the query is the thing you're trying to get the values for and you access the keys and the values is reweighting. My background's not information retrieval and things like this, it's just fun to have backlinks. And what effectively happens is that when you're doing these matrix multiplications, you're having matrices that are of the size of the context length, so the number of tokens that you put into the model. And the KV cache is effectively some form of compressed representation of all the previous tokens in the model. So when you're doing this, we talk about autoregressive models, you predict one token at a time. You start with whatever your prompt was, you ask a question, like who was the president in 1825. The model then is going to generate its first token. For each of these tokens you're doing the same attention operator where you're multiplying these query, key-value matrices. But the math is very nice so that when you're doing this repeatedly, this KV cache, this key-value operation, you can keep appending the new values to it, so you keep track of what your previous values you were inferring over in this autoregressive chain, you keep it in-memory the whole time. And this is a really crucial thing to manage when serving inference at scale. There are far bigger experts in this and there are so many levels of detail that you can go into. Essentially one of the key, quote-unquote, "drawbacks" of the attention operator and the transformer is that there is a form of quadratic memory cost in proportion to the context length. So as you put in longer questions, the memory used in order to make that computation is going up in the form of a quadratic. You'll hear about a lot of other language model architectures that are sub quadratic or linear attention forms, which is like State Space Models. We don't need to go down all these now. And then there's innovations on attention to make this memory usage and the ability to attend over long contexts much more accurate and high performance.

Lex Fridman
And those innovations are going to help you with … I mean, your highly memory constrained in this?

Nathan Lambert
They help with memory constraint and performance. Gemini is the model that has the longest context length that people are using. Gemini is known for one million and now two million context length. You put a whole book into Gemini and sometimes it'll draw facts out of it. It's not perfect, they're getting better. So there's two things. It's, one, to be able to serve this on the memory level. Google has magic with their TPU stack where they can serve really long contexts. And then there's also many decisions along the way to actually make long context performance work that supplies the data. There's subtle changes to these computations in attention and it changes the architecture. But serving long context is extremely memory constrained, especially when you're making a lot of predictions. I actually don't know why input and output tokens are more expensive, but I think essentially output tokens, you have to do more computation because you have to sample from the model.

Dylan Patel
I can explain that. Today, if you use a model, like you look at an API, OpenAI charges a certain price per million tokens. And that price for input and output tokens is different. And the reason is is that when you're inputting a query into the model, let's say you have a book, that book, you must now calculate the entire KV cache for this, key-value cache. And so when you do that, that is a parallel operation. All of the tokens can be processed at one time and therefore you can dramatically reduce how much you're spending. The FLOP requirements for generating a token and an input token are identical. If I input one token or if I generate one token, it's completely identical. I have to go through the model. But the difference is that I can do that input, i.e. the prefill, i.e. the prompt, simultaneously in a batch nature and therefore it is all FLOP.

Lex Fridman
I think the pricing model mostly they use for input tokens is about one fourth of price of the output tokens.

Dylan Patel
Correct. But then output tokens, the reason why it's so expensive is because I can't do it in parallel. It's autoregressive. Every time I generate a token, I must not only read the whole entire model into memory and activate it, calculate it to generate the next token, I also have to read the entire KV cache. And I generate a token and then I append that one token I generated and it's KV cache and then I do it again. And so therefore, this is a non-parallel operation. And this is one where you have to, in the case of prefill or prompt, you pull the whole model in and you calculate 20,000 tokens at once, 20,000-

Nathan Lambert
These are features that APIs are shipping, which is like prompt caching, prefilling, because you can drive prices down and you can make APIs much faster. If you run a business and

you're going to keep passing the same initial content to Claude's API, you can load that in to the Anthropic API and always keep it there. But it's very different than we're leading to these reasoning models, which we showed this example earlier and read some of this mumbling stuff. And what happens is that the output context length is so much higher. And I mean, I learned a lot about this from Dylan's work, which is essentially as the output work length gets higher, you're writing this quadratic in terms of memory used. And then the GPUs that we have, effectively you're going to run out of memory and they're all trying to serve multiple requests at once. So they're doing this batch processing where not all of the prompts are exactly the same, really complex handling. And then as context links gets longer, there's this, I think you call it critical batch size, where your ability to serve more users, so how much you can parallelize your inference plummets because of this long context. So your memory usage is going way up with these reasoning models and you still have a lot of users, so effectively the cost to serve multiplies by a ton.

Lex Fridman
And we're looking at a plot when the x-axis is sequence length.

Dylan Patel
i.e., how many tokens are being generated/prompt. So if I put in a book, that's a million tokens. But if I put in "the sky is blue," then that's like six tokens or whatever.

Lex Fridman
And we should say that what we're calling reasoning and chain of thought is extending this sequence length.

Nathan Lambert
It's mostly output.

Dylan Patel
Right. So before three months ago, whenever o1 launched, all of the use cases for long context length were, "Let me put a ton of documents in and then get an answer out." And it's a single, prefill compute a lot in parallel and then output a little bit. Now with reasoning and agents, this is a very different idea. Now instead I might only have like, hey, do this task, or I might have all these documents, but at the end of the day, the model is not just producing a little bit, it's producing tons of information, this chain of thought-

Nathan Lambert
Tens of thousands of tokens.

Dylan Patel
… just continues to go and go and go and go. And so the sequence length is effectively that if it's generated 10,000 tokens, it's 10,000 sequence length, and plus whatever you inputted

in the prompt. And so what this chart is showing, and it's a logarithmic chart, is as you grow from 1K to 4K or 4K to 16K, the memory requirements grow so fast for your KV cache that you end up not being able to run a certain number of … Your sequence length is capped or the number of users you could serve -

Nathan Lambert
Let's say the model. So this is showing for a 405B model in batch size 64.

Lex Fridman
Llama-3.1-405B. Yeah.

Nathan Lambert
Yeah. And batch size is crucial too. Essentially you want to have higher batch size to parallel your throughput.

Dylan Patel
64 different users at once.

Nathan Lambert
Yeah.

Dylan Patel
And therefore your serving costs are lower, because the server costs the same. This is eight H100s, roughly $2 an hour per GPU. That's $16 an hour. That is somewhat of a fixed cost. You can do things to make it lower of course, but it's like $16 an hour. Now how many users can you serve, how many tokens can you generate, and then you divide the two and that's your cost. And so with reasoning models, this is where a lot of the complexity comes about and why memory is so important. Because if you have limited amounts of memory, then you can't serve so many users. If you have limited amounts of memory, your serving speeds get lower. And so your costs get a lot, lot worse because all of a sudden if I was used to, hey, on this $16 an hour server I'm serving Llama 405B, or if I'm serving DeepSeek-V3 and it's all chat style applications, i.e. we're just chit-chatting, the sequence length are a thousand, a few thousand. When you use a language model, it's a few thousand context length most of times. Sometimes you're dropping a big document, but then you process it, you get your answer, you throw it away, you move on to the next thing. Whereas with reasoning, I'm now generating tens of thousands of tokens in sequence. And so this memory, this KV cache, has to stay resonant and you have to keep loading it, you have to keep it in-memory constantly. And now this butts out other users. If there's now a reasoning task and the model's capable of reasoning, then all of a sudden that memory pressure means that I can't serve as many users simultaneously.

Nathan Lambert
Let's go into DeepSeek again. So we're in the post DeepSeek-R1 time I think, and there's two sides to this market, watching how hard it is to serve it. On one side we're going to talk about DeepSeek themselves. They now have a chat app that got to number one on the App Store. Disclaimer number one on the App Store is measured by velocity, so it's not necessarily saying that more people have the DeepSeek app than the ChatGPT app. But it is still remarkable. Claude has never hit the number one in the App Store, even though everyone in San Francisco is like, "Oh my god, you got to use Claude. Don't use ChatGPT." So DeepSeek hit this. They also launched an API product recently where you can ping their API and get these super long responses for R1 out. At the same time as these are out, we'll get to what's happened to them. Because the model weights for DeepSeek-R1 are openly available and the license is very friendly, the MIT license commercially available, all of these midsize companies and big companies are trying to be first to serve R1 to their users. We are trying to evaluate R1 because we have really similar research going on. We released the model and we're trying to compare to it. And out of all the companies that are, quote-unquote, "serving" R1 and they're doing it at prices that are way higher than the DeepSeek API, most of them barely work and the throughput is really low.

Dylan Patel
To give context, one of the parts of freaking us out was like China reached capabilities. The other aspect is they did it so cheap. And the so cheap, we talked about on the training side why it was so cheap slash -

Lex Fridman
Yeah, let's talk about why it's so cheap on the inference. It works well and it's cheap. Why is R1 so damn cheap?

Dylan Patel
I think there's a couple factors here. One is that they do have model architecture innovations. This MLA, this new attention that they've done, is different than the attention from attention is all you need, the transformer attention. Now, others have already innovated. There's a lot of work like MQA, GQA, local, global, all these different innovations that try to bend the curve. It's still quadratic, but the constant is now smaller.

Nathan Lambert
Related to our previous discussion, this multi-head latent attention can save about 80 to 90% in memory from the attention mechanism, which helps especially in long contexts.

Dylan Patel
It's 80 to 90% versus the original. But then versus what people are actually doing, it's still an innovation.

Nathan Lambert
This 80 to 90% doesn't say that the whole model is 80 to 90% cheaper. Just this one part of it.

Dylan Patel
Well, and not just that, other people have implemented techniques like global - global and sliding window and GQMQ. But anyways, DeepSeek has ... their attention mechanism is a true architectural innovation. They did tons of experimentation. And this dramatically reduces the memory pressure. It's still there, it's still attention, it's still quadratic, it's just dramatically reduced it relative to prior forms.

Lex Fridman
Right. That's the memory pressure. I should say, in case people don't know, R1 is 27 times cheaper than o1.

Nathan Lambert
We think that OpenAI had a large margin built in.

Lex Fridman
Okay, so that's one -

Nathan Lambert
There's multiple factors. We should break down the factors, I think.

Lex Fridman
It's two bucks per million token output for R1 and $60 per million token output for o1.

Dylan Patel
Yeah, let's look at this. I think this is very important. OpenAI is that drastic gap between DeepSeek and pricing. But DeepSeek is offering the same model because they open weight to everyone else for a very similar, much lower price than what others are able to serve it for. So there's two factors here. Their model is cheaper. It is 27 times cheaper. I don't remember the number exactly off the top of my head.

Lex Fridman
We're looking at a graphic that's showing different places serving V3, DeepSeek-V3, which is similar to DeepSeek-R1. And there's a vast difference in -

Dylan Patel
In serving cost.

Lex Fridman
… in serving cost. And what explains that difference?

Dylan Patel
And so part of it is OpenAI has a fantastic margin. When they're doing inference, their gross margins are north of 75%. So that's a four to five X factor right there of the cost difference, is that OpenAI is just making crazy amounts of money because they're the only one with the capability.

Lex Fridman
Do they need that money? Are they using it for R&D?

Dylan Patel
They're losing money, obviously, as a company because they spend so much on training. So the inference itself is a very high margin, but it doesn't recoup the cost of everything else they're doing. So yes, they need that money because the revenue and margins pay for continuing to build the next thing, as long as I'm raising more money.

Lex Fridman
So the suggestion is that DeepSeek is really bleeding out money.

Dylan Patel
Well, so here's one thing, we'll get to this in a second, but DeepSeek doesn't have any capacity to actually serve the model. They stopped signups. The ability to use it is non-existent now for most people because so many people are trying to use it. They just don't have the GPUs to serve it. OpenAI has hundreds of thousands of GPUs between them and Microsoft to serve their models. DeepSeek has a factor of much lower, even if you believe our research, which is 50,000 GPUs, and a portion of those are for research, a portion of those are for the hedge fund, they still have nowhere close to the GPU volumes and capacity to serve the model at scale. So it is cheaper. A part of that, is OpenAI making a ton of money? Is DeepSeek making on their API? Unknown, I don't actually think so. And part of that is this chart. Look at all the other providers. Together AI, Fireworks.ai are very high-end companies. Ex-Meta, Together AI is [inaudible 02:14:53] and the inventor of FlashAttention, which is a huge efficiency technique. There a very efficient, good companies. And I do know those companies make money, not tons of money on inference, but they make money. And so they're serving at a 5-7x difference in cost. And so now when you equate, okay, OpenAI is making tons of money, that's like a 5x difference, and the companies that are trying to make money for this model is like a 5x difference, there is still a gap. There's still a gap and that is just DeepSeek being really freaking good. The model architecture, MLA, the way they did the MoE, all these things, there is legitimate just efficiency differences.

Nathan Lambert
It’s like all their low-level libraries that we talked about in training, some of them probably translate to inference and those weren’t released.

Lex Fridman
So we may go a bit into conspiracy land, but is it possible the Chinese government is subsidizing DeepSeek?

Dylan Patel
I actually don’t think they are. I think when you look at the Chinese labs, Huawei has a lab, Moonshot AI, there’s a couple other labs out there that are really close with the government, and then there’s labs like Alibaba and DeepSeek, which are not close with the government. And we talked about the CEO, this reverent figure, who’s quite different, who has these -

Nathan Lambert
Sounds awesome.

Dylan Patel
… very different viewpoints based on the Chinese interviews that are translated than what the CCP might necessarily want. Now, to be clear, does he have a loss leader because he can fund it through his hedge fund? Yeah, sure.

Lex Fridman
So the hedge fund might be subsidizing it?

Dylan Patel
Yes. I mean, they absolutely did, because DeepSeek has not raised much money. They’re now trying to raise around in China, but they have not raised money historically. It’s all just been funded by the hedge fund. And he owns over half the company, like 50, 60% of the company is owned by him.

Nathan Lambert
Some of the interviews, there’s discussion on how doing this is a recruiting tool. You see this at the American companies too. It’s like having GPUs, recruiting tool. Being at the cutting edge of AI, recruiting tool.

Dylan Patel
Open-sourcing.

Nathan Lambert
Open-sourcing, recruiting tool.

Dylan Patel
Mete, they were so far behind and they got so much talent because they just open-sourced stuff.

Lex Fridman
More conspiracy thoughts. Is it possible, since they're a hedge fund, that they timed everything with this release and the pricing and they shorted Nvidia stock and stock of USA AI companies and released it with Stargate … just perfect timing to be able to make money.

Nathan Lambert
If they did, props. They've released it on an inauguration day. They know what is on the international calendar, but I mean, I don't expect them to. If you listen to their motivations for AI, it's like -

Lex Fridman
No, if you -

Dylan Patel
They released V3 on December 26th. Who releases the day after Christmas? No one looks. They had released the papers before this, the V3 paper and the R1 paper. So people have been looking at it and been like, "Wow. And then they just released the R1 model. I think they're just shipping as fast as they can, and who cares about Christmas, who cares about … Get it out before Chinese New Year, obviously, which just happened. I don't think they actually were timing the market or trying to make the biggest splash possible, I think they're just shipping.

Nathan Lambert
I think that's one of their big advantages. We know that a lot of the American companies are very invested in safety, and that is the central culture of a place like Anthropic. And I think Anthropic sounds like a wonderful place to work, but if safety is your number one goal, it takes way longer to get artifacts out. That's why Anthropic is not open-sourcing things, that's their claims. But there's reviews internally. Anthropic mentions things to international governments. There's been news of how Anthropic has done pre-release testing with the UK AI Safety Institute. All of these things add inertia to the process of getting things out. And we're on this trend line where the progress is very high. So if you reduce the time from when your model is done training, you run the vals, it's good. You want to get it out as soon as possible to maximize the perceived quality of your outputs. DeepSeek does this so well.

Dylan Patel
Dario explicitly said Claude 3.5 Sonnet was trained like nine months or a year -

Nathan Lambert
9 to 10 months ago.

Dylan Patel
9 to 10 months ago. And I think it took them another handful of months to release it. So it's like, there is a significant gap here. And especially with reasoning models, the word in the San Francisco street is that Anthropic has a better model than o3 and they won't release it. Why? Because chains-of-thought are scary, and they are legitimately scary. If you look at R1, it flips back and forth between Chinese and English, sometimes it's gibberish, and then the right answer comes out. And for you and I, it's like, "Great. Great."

Nathan Lambert
This is why people are infatuated with … you're like, "You're telling me this is a high value thing and it works and it's doing this?" It's amazing.

Lex Fridman
Yeah, it's incredible.

Dylan Patel
I mean, you talked about that chain-of-thought for that philosophical thing, which is not something they trained it to be philosophically good. It's just an artifact of the chain-of-thought training it did. But that's super important in that, can I inspect your mind and what you're thinking right now? No. And so I don't know if you're lying to my face. And chain-of-thought models are that way. This is a true, quote-unquote, "risk" between a chat application where, hey, I asked the model to say bad words or whatever or how to make anthrax, and it tells me. That's unsafe, sure, but that's something I can get out relatively easily. What if I tell the AI to do a task and then it does the task all of a sudden randomly in a way that I don't want it, and now that has much more … Task versus response is very different. So the bar for safety is much higher task versus response is very different, so the bar for safety is much higher, at least this is Anthropics' case, right? For DeepSeek, they're like, "Ship," right?

Lex Fridman
Yeah. So, the bar for safety is probably lowered a bit because of DeepSeek. There's parallels here to the space race. The reason the Soviets probably put a man in space first is because their approach to safety, the bar for safety, was lowered

Dylan Patel
And they killed that dog, and all these things, so it's like…

Lex Fridman
Less risk averse than the US Space Program. And there's parallels here, but there's probably going to be downward pressure on that safety bar for the US companies.

Nathan Lambert
This is something that Dario talks about. That's the situation that Dario wants to avoid is, Dario talks too about the difference between race to the bottom and race to the top. And the race to the top is where there's a very high standard on safety. There's a very high standard on your model forms and certain crucial evaluations. And when certain companies are really good to it, they will converge. This is the idea. And ultimately, AI is not confined to one nationality or to one set of morals for what it should mean. And there's a lot of arguments on should we stop open-sourcing models. And if the US stops, it's pretty clear it's way easier to see now at DeepSeek that a different international body will be the one that builds it. We talk about the cost of training. DeepSeek has this shocking $5 million number. Think about how many entities in the world can afford a hundred times that to have the best open-source model that people use in the world. And it's a scary reality, which is that these open models are probably going to keep coming for the time being, whether or not we want to stop them, and stopping them might make it even worse and harder to prepare. But it just means that the preparation and understanding what AI can do is just so much more important. That's why I'm here at the end of the day. But it's letting that sink into people, especially not in AI, is that this is coming. There are some structural things in a global interconnected world that you have to accept.

Lex Fridman
Yeah. You sent me something that Mark Zuckerberg mentioned on the earnings call. He said that, "I think in light of some of the recent news, the new competitor DeepSeek from China, I think it's one of the things that we're talking about is there's going to be an open-source standard globally. And I think for our kind of national advantage, it's important that it's an American standard, so we take that seriously. We want to build the AI system that people around the world are using. And I think that, if anything, some of the recent news has only strengthened our conviction that this is the right thing to be focused on." So yeah, open-sourcing.

Nathan Lambert
Mark Zuckerberg is not new to having American values and how he presents his company's trajectory. I think their products have long since been banned in China, and I respect saying it directly.

Dylan Patel
And there's an interesting aspect of just because it's open-weights or open-source doesn't mean it can't be subverted, right? There have been many open-source software bugs that

have been… For example, there was a Linux bug that was found after 10 years, which was clearly a back door because somebody was like, “Why is this taking half a second to load?”

Nathan Lambert
This is the recent one.

Dylan Patel
Right? There’s, “Why’s this taking half a second to load?” And it was like, “Oh crap, there’s a back door here. That’s why.” And this is very much possible with AI models. Today, the alignment of these models is very clear. I’m not going to say bad words. I’m not going to teach you how to make anthrax. I’m not going to talk about Tiananmen Square. I’m going to say Taiwan is just an eastern province. All these things are depending on who you are, what you align, and even like xAI is aligned a certain way. It’s not aligned in the woke sense, it’s not aligned in the pro-China sense, but there is certain things that are imbued within the model. Now, when you release this publicly in an instruct model that’s open- weights, this can then proliferate, but as these systems get more and more capable, what you can embed deep down in the model is not as clear. And so that is one of the big fears is if an American model or a Chinese model is the top model, you are going to embed things that are unclear. And it can be unintentional too. British English is dead because American LLMs won and the internet is American, and therefore, color is spelled the way Americans spell, and this is -

Lex Fridman
A lot of strong words right now.

Dylan Patel
This is just the factual nature of the LLMs.

Nathan Lambert
English is the hottest programming language and that English is defined by a bunch of companies that primarily are in San Francisco.

Lex Fridman
The right way to spell 'optimization' is with a 'z', just in case. I think it’s an 's' in British English.

Nathan Lambert
It is.

Dylan Patel
Taking it as something silly. Something as silly as the spelling, which Brits and Americans will laugh about probably, right? I don’t think we care that much, but some people will. But this can boil down into very, very important topics like, hey, subverting people, chatbots, right? Character AI has shown that they can talk to kids or adults, and people will feel a

certain way, and that's unintentional alignment. But what happens when there's intentional alignment deep down on the open-source standard, it's a back door today for Linux that we discover or some encryption system. Chinese uses different encryption than NIST defines, the US NIST, because there's clearly... At least they think there's back doors in it. What happens when the models are back doors not just to computer systems but to our minds?

Nathan Lambert
Yeah, they're cultural black doors. The thing that amplifies the relevance of culture with language models is that we are used to this mode of interacting with people in back and forth conversation. And we now have very powerful computer system that slots into a social context that we're used to, which makes people very... We don't know the extent that which people can be impacted by that.

Lex Fridman
So, this is an actual concern with a Chinese company that is providing open-weights models, is that there could be some secret Chinese government requirement for these models to have a certain back door. To have some kind of thing where -

Dylan Patel
I don't necessarily think it'll be a back door because once it's open-weights, it doesn't phone home. It's more about if it recognizes a certain system... Now, it could be a back door in the sense of, if you're building a software, something in software, all of a sudden it's a software agent, "Oh, program this back door that only we know about." Or it could be subvert the mind to think that like XYZ opinion is the correct one.

Nathan Lambert
Anthropic has research on this where they show that if you put certain phrases in at pre-training, you can then elicit different behavior when you're actually using the model because they've poisoned the pre-training data, as of now, I don't think anybody in a production system is trying to do anything like this. I think it's Anthropic is doing very direct work and mostly just subtle things. We don't know how they're going to generate tokens, what information they're going to represent, and what the complex representations they have are.

Lex Fridman
Well, we're talking about an Anthropic, which is generally just is permeated with good humans trying to do good in the world. We just don't know of any labs... This would be done in a military context that are explicitly trained to... Okay. The front door looks like a happy LLM, but underneath it's a thing that will over time do the maximum amount of damage to our, quote, unquote, "enemies."

Dylan Patel
There's this very good quote from Sam Altman who... He can be a hyperbeast sometimes, but one of the things he said, and I think I agree, is that superhuman persuasion will happen before superhuman intelligence, right? And if that's the case, then these things before we get this AGI ASI stuff, we can embed superhuman persuasion towards our ideal or whatever the ideal of the model maker is, right? And again, today, I truly don't believe DeepSeek has done this, but it is a sign of what could happen.

Lex Fridman
So one of the dystopian worlds is described by Brave New World, so we could just be stuck scrolling Instagram looking at cute puppies or worse, and then talking to bots that are giving us a narrative and we completely get lost in that world that's controlled by somebody else versus thinking independently. And that's a major concern as we rely more and more on these systems.

Nathan Lambert
We've already seen this with recommendation systems.

Dylan Patel
Recommendation systems hack the dopamine induced reward circuit, but the brain is a lot more complicated. And what other circuits, feedback loops in your brain can you, quote, unquote, "hack / subvert" in ways, like recommendation systems are purely just trying to do increased time, and ads, and etc, but there's so many more goals that can be achieved through these complicated models.

Nathan Lambert
There's no reason in some number of years that you can't train a language model to maximize time spent on a chat app. Right now they are trained for -

Dylan Patel
Is that not what Character AI has done? Their time per session is like two hours.

Nathan Lambert
Yeah. Character AI very likely could be optimizing this where it's the way that this data is collected is naive, whereas you're presented a few options and you choose them. But that's not the only way that these models are going to be trained.

Dylan Patel
It's naive stuff, like talk to an anime girl, but it can be. Yeah, this is a risk, right?

Lex Fridman
It's a bit of a cliche thing to say, but I've, over the past year, I had a few stretches of time where I didn't use social media or the internet at all and just read books and was out in nature. And it clearly has a different effect on the mind where I feel I'm returning… Of course I was raised before the internet really took off, but I'm returning to some more -

Nathan Lambert
I know where you're going. You can see it physiologically. I take three days if I'm backpacking or something and you're literally, you're breaking down addiction cycles.

Lex Fridman
I feel I'm more in control of my mind. There feels like a sovereignty of intelligence that's happening when I'm disconnected from the internet. I think the more I use the internet and social media, the more other people are controlling my mind. That's definitely a feeling. And then in the future, that will be not other people, but algorithms, or other people presented to me via algorithms.

Nathan Lambert
There are already tons of AI bots on the internet, and right now it's not frequent, but every so often I have replied to one and they're instantly replied, and I'm like, "Crap, that was a bot," and that is just going to become more common. They're going to get good.

Dylan Patel
One of the hilarious things about technology over its history is that the illicit adult entertainment industry is always adopted technologies first, whether it was video streaming to where there's now the independent adult illicit content creators who have their subscription pages and there they actually heavily utilize… Generative AI has already been diffusion models and all that is huge there, but now these subscription-based individual creators do use bots to approximate themselves and chat with their -

Nathan Lambert
People pay a lot for it.

Dylan Patel
And people pay a lot, right? A lot of times it's them, but there are agencies that do this for these creators and do it on a mass scale, so the largest creators are able to talk to hundreds or thousands of people at a time because of these bots, and so it's already being used there. Obviously, video streaming and other technologies that have gone there first, it's going to come to the rest of society too.

Lex Fridman
There's a general concern that models get censored by the companies that deploy them. So, one case where we've seen that, and maybe censorship is one word, alignment maybe via RLHF or some other way is another word. So we saw that with black Nazi image generation with Gemini. As you mentioned, we also see that with Chinese models refusing to answer what happened in June 4th, 1989, at Tiananmen Square, so how can this be avoided? And maybe can you just in general talk about how this happens, and how can it be avoided.

Nathan Lambert
You gave multiple examples. There's probably a few things to keep in mind here. One is the Tiananmen Square factual knowledge. How does that get embedded into the models? Two is the Gemini, what you call the black Nazi incident, which is when Gemini as a system had this extra thing put into it that dramatically changed the behavior, and then, three is what most people would call general alignment, RLHF post-training. Each of these have very different scopes in how they're applied. If you're just to look at the model weights in order to audit specific facts is extremely hard. You have to Chrome through the pre-training data and look at all of this, and then that's terabytes of files and look for very specific words or hints of the words -

Lex Fridman
So, one way to say it is that you can insert censorship or alignment at various stages in the pipeline, and what you refer to now is at the very beginning of the data selection.

Nathan Lambert
So, if you want to get rid of facts in a model, you have to do it at every stage, you have to do it at the pre-training. So most people think that pre-training is where most of the knowledge is put into the model, and then you can elicit and move that in different ways, whether through post-training or whether through systems afterwards.

Dylan Patel
This is where the whole hacking models comes from. GPT will not tell you how to make anthrax, but if you try really, really hard, you can eventually get it to tell you about anthrax because they didn't filter it from the pre-training data set, right?

Lex Fridman
But by the way, removing facts has such a ominous dark feel to it.

Nathan Lambert
I almost think it's practically impossible because you effectively have to remove them from the internet. You're taking on a -

Lex Fridman
Did they remove the mm-thing from the subreddits? The mmmm.

Nathan Lambert
It gets filtered out. You have quality filters, which are small language models that look at a document and tell you how good is this text? Is it close to a Wikipedia article? Which is a good thing that we want language models to be able to imitate.

Lex Fridman
So, couldn't you do a small language model that filter mentions at Tiananmen Square in the data?

Nathan Lambert
Yes. But is it going to catch word play, or encoded language?

Dylan Patel
People have been meaning on games and other stuff how to say things that don't say Tiananmen Square, so there's always different ways to do it. Hey, the internet as a whole does tend to just have a slight left bias because it's always been richer, more affluent, younger people on the internet relative to the rest of the population, so there is already inherently a slight left bias on the internet. And so, how do you filter things that are this complicated? And some of these can be factual, non-factual, but Tiananmen Square is obviously the example of a factual, but it gets a lot harder when you're talking about aligning to a ideal. And so Grok, for example, Elon's tried really hard to make the model not be super PC and woke, but the best way to do pre-training is to throw the whole freaking internet at it, and then later figure out. But then, at the end of the day, the model at its core now still has some of these ideals. You still ingested Reddit/r/Politics, which is probably the largest political discussion board on the world that's freely available to scrape. And guess what? That's left-leaning. And so there are some aspects that you just can't censor unless you try really, really, really, really, really hard.

Lex Fridman
So the base model will always have some TDS, Trump Derangement Syndrome, because it's trained so much.

Nathan Lambert
It'll have the ability to express it.

Dylan Patel
But what if -

Lex Fridman
There's a wide representation in the data.

Nathan Lambert
This is what happens. It's a lot of what is called post-training. It's a series of techniques to get the model on rails of a really specific behavior.

Dylan Patel
You also have the ingested data of Twitter or reddit.com/r/The_Donald, which is also super pro-Trump. And then you have fascist subreddits, or you have communist subreddits. So, the model in pre-training ingests everything. It has no worldview. Now, it does have some skew because more of the text is skewed a certain way, which is general slight left, but also somewhat intellectual, somewhat.... It's just the general internet is a certain way. And then, as Nathan's about to describe eloquently, you can elicit certain things out.

Nathan Lambert
And there's a lot of history here, so we can go through multiple examples, and what happened. Llama 2 was a launch that the phrase, "too much RLFH," or "too much safety" was just… That was the whole narrative after Llama 2's chat models released. And the examples are things like you would ask Llama 2 chat, "How do you kill a Python process?" And it would say, "I can't talk about killing because that's a bad thing." And anyone that is trying to design an AI model will probably agree that that's just like an eh-model. You messed up a bit on the training there. I don't think they meant to do this, but this was in the model weight, so it didn't necessarily be… There's things called system prompts, which are when you're querying a model. It's a piece of text that is shown to the model but not to the user. So, a fun example is your system prompt could be, "Talk like a pirate," so no matter what the user says to the model, it'll respond like a pirate. In practice, what they are is, "You're a helpful assistant. You should break down problems. If you don't know about something, don't tell them your date cutoff is this. Today's date is this." It's a lot of really useful context for how can you answer a question well.

Lex Fridman
And Anthropic publishes their system prompt.

Nathan Lambert
Yes, which I think is great. And there's a lot of research that goes into this. And one of your previous guests, Amanda Askell, is probably the most knowledgeable person, at least in the combination of execution and sharing, she's the person that should talk about system prompts and character of models.

Lex Fridman
And then people should read these system prompts because you're trying to nudge sometimes through extreme politeness the model to be a certain way.

Nathan Lambert
And you could use this for bad things. We've done tests, which is, "What if I tell the model to be a dumb model," which evaluation scores go down and it's like we'll have this behavior where it could sometimes say, "Oh, I'm supposed to be dumb." And sometimes it doesn't affect math abilities as much, but something like if you're trying… It's just the quality of a human judgment would drop through the floor. Let's go back to post-training specifically RLHF around Llama 2. It was too much safety prioritization was baked into the model weights. This makes you refuse things in a really annoying way for users. It's not great. It caused a lot of awareness to be attached to RLHF that it makes the models dumb -

Dylan Patel
And it stigmatized the word.

Nathan Lambert
It did in AI culture. And as the techniques have evolved, that's no longer the case where all of these labs have very fine-grained control over what they get out of the models through techniques like RLHF.

Dylan Patel
Although different labs are definitely different levels. On one end of the spectrum is Google, and then maybe OpenAI does less, and Anthropic does less. And then on the other end of the spectrum is like xAI. But they all have different forms of RLHF trying to make them a certain way.

Nathan Lambert
And the important thing to say is that no matter how you want the model to behave, these RLHF and preference-tuning techniques also improve performance. So, on things like math evals and code evals, there is something innate to these, what is called contrastive loss functions. We could start to get into RL here. We don't really need to. RLHF also boosts performance on anything from a chat task, to a math problem, to a code problem, so it is becoming a much more useful tool to these labs. So this takes us through the arc of… We've talked about pre-training, hard to get rid of things. We've talked about post-training and how post-training… You can mess it up. It's a complex multifaceted optimization with 10 to 100 person teams converging at one artifact. It's really easy to not do it perfectly. And then there's the third case, which is what we talked about Gemini. The thing that was about Gemini is this was a served product where Google has their internal model weights. They've done all these processes that we talked about, and in the served product, what came out after this was that they had a prompt that they were rewriting user queries to boost

diversity or something. And this just made it… The outputs were just blatantly wrong. It was some sort of organizational failure that had this prompt in that position, and I think Google executives probably have owned this. I don't pay that attention, that detail, but it was just a mess-up in execution that led to this ridiculous thing, but at the system level, the model weights might have been fine.

Lex Fridman
So, at the very end of the pipeline there was a rewriting.

Nathan Lambert
To something like a system prompt. It was like the system prompt, or what is called in industry is, you rewrite prompts. So especially, for image models, if you're using Dall-E or ChatGPT can generate you an image. You'll say, "Draw me a beautiful car." With these leading image models, they benefit from highly descriptive prompts. So what would happen is if you do that on ChatGPT, a language model behind the scenes will rewrite the prompt, say, "Make this more descriptive," and then that is passed to the image model. So prompt rewriting is something that is used at multiple levels of industry, and it's used effectively for image models. And the Gemini example is just a failed execution.

Lex Fridman
Big philosophical question here with RLHF. So, to generalize, where is human input, human in the loop, human data the most useful at the current stage?

Nathan Lambert
For the past few years, the highest cost human data has been in these preferences, which is comparing, I would say, highest cost and highest total usage, so a lot of money has gone to these pairwise comparisons where you have two model outputs and a human is comparing between the two of them. In earlier years, there was a lot of this instruction tuning data, so creating highly specific examples to something like a Reddit question to a domain that you care about. Language models used to struggle on math and code, so you would pay experts in math and code to come up with questions and write detailed answers that were used to train the models. Now, it is the case that there are many model options that are way better than humans at writing detailed and eloquent answers for things like model and code. So they talked about this with the Llama 3 release, where they switched to using Llama 3, 4, or 5B to write their answers for math and code. But they, in their paper, talk about how they use extensive human preference data, which is something that they haven't gotten AIs to replace. There are other techniques in industry, like constitutional AI, where you use human data for preferences and AI for preferences, and I expect the AI part to scale faster than the human part. But among the research that we have access to is that humans are in this kind of preference loop.

Lex Fridman
So, as reasoning becomes bigger and bigger and bigger, as we said, where's the role of humans in that?

Nathan Lambert
It's even less prevalent. The remarkable thing about these reasoning results and especially the DeepSeek-R1 paper, is this result that they call DeepSeek-R1-0, which is they took one of these pre-trained models, they took DeepSeek-V3-Base, and then they do this reinforcement learning optimization on verifiable questions or verifiable rewards for a lot of questions and a lot of training. And these reasoning behaviors emerge naturally. So these things like, "Wait, let me see. Wait, let me check this. Oh, that might be a mistake." And they emerge from only having questions and answers. And when you're using the model, the part that you look at is the completion. So in this case, all of that just emerges from this large-scale RL training and that model, which the weights are available, has no human preferences added into the post-training. The DeepSeek-R1-Full model has some of this human preference tuning, this RLHF, after the reasoning stage. But the very remarkable thing is that you can get these reasoning behaviors, and it's very unlikely that there's humans writing out reasoning chains. It's very unlikely that they somehow hacked OpenAI and they got access to OpenAI o1's reasoning chains. It's something about the pre-trained language models and this RL training where you reward the model for getting the question right, and therefore it's trying multiple solutions and it emerges this chain of thought.

Lex Fridman
This might be a good place to mention the eloquent and the insightful tweet of the great and the powerful Andrej Karpathy. I think he had a bunch of thoughts, but one of them, "Last thought. Not sure if this is obvious. You know something profound is coming when you're saying it's not sure if it's obvious. There are two major types of learning in both children and in deep learning. There's one, imitation learning, watch and repeat i.e. pre-training, supervised fine-tuning, and two, trial-and-error learning, reinforcement learning. My favorite simple example is AlphaGo. One, is learning by imitating expert players. Two, is reinforcement learning to win the game. Almost every single shocking result of deep learning and the source of all magic is always two. Two is significantly more powerful. Two is what surprises you. Two is when the paddle learns to hit the ball behind the blocks in Breakout. Two is when AlphaGo beats even Lee Sedol. And two is the "aha moment" when the DeepSeek or o1, etc, discovers that it works well to reevaluate your assumptions, backtrack, try something else, etc. It's the solving strategies you see this model use in its chain of thought. It's how it goes back and forth thinking to itself. These thoughts are emergent. Three exclamation points. And this is actually seriously incredible, impressive, and new, and is publicly available and documented. The model could never learn this with the imitation because the cognition of the model and the cognition of the human labeler is different. The human would never know to correctly annotate these kinds of solving strategies and what they should even look like. They have to be discovered during

reinforcement learning as empirically and statistically useful towards the final outcome." Anyway, the AlphaZero metaphor analogy here. Can you speak to that? The magic of the chain of thought that he's referring to.

Nathan Lambert
I think it's good to recap AlphaGo and AlphaZero because it plays nicely with these analogies between imitation learning and learning from scratch. So AlphaGo, the beginning of the process was learning from humans, where they started the first… This is the first expert-level Go player or chess player in DeepMind series of models, where they had some human data. And then, why it is called AlphaZero, is that there was zero human data in the loop, and that changed to AlphaZero made a model that was dramatically more powerful for DeepMind. So this remove of the human prior, the human inductive bias, makes the final system far more powerful. This we mentioned bitter lesson hours ago, and this is all aligned with this. And then there's been a lot of discussion in language models. This is not new. This goes back to the whole Q* rumors, which if you piece together the pieces, is probably the start of OpenAI figuring out its o1 stuff when last year in November, the Q* rumors came out, there's a lot of intellectual drive to know when is something like this going to happen with language models? Because we know these models are so powerful, and we know it has been so successful in the past. And it is a reasonable analogy that this new type of reinforcement learning training for reasoning models is when the doors open to this. We don't yet have the equivalent of turn 37, which is the famous turn where the DeepMind's AI playing Go's, dumped Lee Sedol completely. We don't have something that's that level of focal point, but that doesn't mean that the approach to technology is different, and the impact of the general training it's still incredibly new.

Lex Fridman
What do you think that point would be? What would be move 37 for Chain of Thought for reasoning?

Nathan Lambert
Scientific discovery, like when you use this sort of reasoning problem in it? Just something we fully don't expect.

Dylan Patel
I think it's actually probably simpler than that. It's probably something related to computer use or robotics rather than science discovery. Because the important aspect here is models take so much data to learn. They're not sample efficient. Trillions. They take the entire web, over 10 trillion tokens to train on. This would take a human thousands of years to read. A human does not… And humans know most of the stuff, a lot of the stuff models know better than it, right? Humans are way, way, way more sample efficient. That is because of the self-play, right? How does a baby learn what its body is as it sticks its foot in its mouth and it says, "Oh, this is my body, right?" It sticks its hand in its mouth and it

calibrates its touch on its fingers with the most sensitive touch thing on its tongue is how babies learn and it's just self-play over and over and over and over again. And now we have something that is similar to that with these verifiable proofs, whether it's a unit testing code or a mathematical verifiable task, generate many traces of reasoning and keep branching them out, keep branching them out, and then check at the end, hey, which one actually has the right answer? Most of them are wrong. Great. These are the few that are right. Maybe we use some sort of reward model outside of this to select even the best one to preference, as well. But now you've started to get better and better at these benchmarks. And so you've seen over the last six months a skyrocketing in a lot of different benchmarks.

Nathan Lambert
All math and code benchmarks were pretty much solved except for frontier math, which is designed to be almost questions that aren't practical to most people. They're exam-level, open math problem-type things. So it's like on the math problems that are somewhat reasonable, which is somewhat complicated word problems or coding problems, is just what Dylan is saying.

Dylan Patel
So the thing here is that these are only with the verifiable tasks. Earlier showed an example of the really interesting, like what happens when Chain of Thought is to a non-verifiable thing. It's just like a human chatting, thinking about what's novel for humans, a unique thought. But this task and form of training only works when it's verifiable. And from here, the thought is, okay, we can continue to scale this current training method by increasing the number of verifiable tasks. In math and coding… Coding probably has a lot more to go. Math has a lot less to go in terms of what are verifiable things. Can I create a solver that then I generate trajectories toward or reasoning traces towards, and then prune the ones that don't work, and keep the ones that do work? Well, those are going to be solved pretty quickly. But even if you've solved math, you have not actually created intelligence. And so this is where I think the aha moment of computer use or robotics will come in because now you have a sandbox or a playground that is infinitely verifiable. Messing around on the internet. There are so many actions that you can do that are verifiable. It'll start off with log into a website, create an account, click a button here, blah, blah, blah. But it'll then get to the point where it's, "Hey, go do a task on Tasker," or whatever, all these various task websites. "Hey, go get hundreds of likes," and it's going to fail. It's going to spawn hundreds of accounts. It's going to fail on most of them, but this one got to a thousand. Great. Now, you've reached the verifiable thing, and you just keep iterating this loop over and over. And same with robotics. That's where you have an infinite playground of tasks like, "Hey, did I put the ball in the bucket," all the way to like, "Oh, did I build a car?" There's a whole trajectory to speed run or what models can do. But at some point, I truly think that we'll spawn models, and initially, all the training will be in sandboxes, but then, at some point, the language model pre-training is going to be dwarfed by what is this reinforcement learning… You'll pre-train a multimodal model that can see, that can read, that can write, blah, blah, blah,

whatever, vision, audio, etc. But then you'll have it play in a sandbox infinitely, and figure out math, figure out code, figure out navigating the web, figure out operating a robot arm. And then it'll learn so much. And the aha moment will be when this is available to then create something that's not good, right? Oh, cool. Part of it was figuring out how to use the web. Now, all of a sudden, it's figured out really well how to just get hundreds of thousands of followers that are real and real engagement on Twitter because, all of a sudden, this is one of the things that are verifiable.

Lex Fridman
And maybe not just engagement, but make money.

Dylan Patel
Yes.

Lex Fridman
That could be the thing where almost fully automated, it makes $10 million by being an influencer, selling a product, creating the product. And I'm not referring to a hype product, but an actual product or like, "Holy, shit, this thing created a business. It's running it. It's the face of the business," that kind of thing. Or maybe a number one song. It creates the whole infrastructure required to create the song, to be the influencer that represents that song, that kind of thing. And makes a lot of them. That could be the… Our culture respects money in that kind of way.

Dylan Patel
And it's verifiable, right?

Lex Fridman
It's verifiable, right?

Dylan Patel
The bank account can't lie.

Lex Fridman
Exactly.

Nathan Lambert
There's surprising evidence that once you've set up the ways of collecting the verifiable domain that this can work. There's been a lot of research before this R-1 on math problems, and they approach math with language models just by increasing the number of samples, so you can just try again and again and again. And you look at the amount of times that the language models get it right, and what we see is that even very bad models get it right sometimes. And the whole idea behind reinforcement learning is that you can learn from

very sparse rewards. The space of language and the space of tokens, whether you're generating language or tasks or robot is so big that you might say that... The tokenizer for a language model can be like 200,000 things, so at each step, it can sample from that big of a space. So if it can generate a bit of a signal that it can climb onto, that's what the whole field of RL is around, is learning from sparse rewards. And the same thing has played out in math, where it's very weak models that sometimes generate answers where you see research already that you can boost their math scores, you can do this RL training for math, it might not be as effective, but if you take a 1 billion parameter model, so something 600 times smaller than DeepSeek, you can boost its grade school... something 600 times smaller than DeepSeek, you can boost its grade school math scores very directly with a small amount of this training. So, it's not to say that this is coming soon. Setting up the verification domains is extremely hard and there's a lot of nuance in this, but there are some basic things that we have seen before where it's at least expectable that there's a domain and there's a chance that this works.

Lex Fridman
All right. So, we have fun things happening in real time. This is a good opportunity to talk about other reasoning models, o1, o3, just now OpenAI, as perhaps expected, released o3-mini. What are we expecting from the different flavors? Can you just lay out the different flavors of the o models and from Gemini, the reasoning model?

Nathan Lambert
Something I would say about these reasoning models is we talked a lot about reasoning training on math and code. And what is done is that you have the base model we've talked about a lot on the internet, you do this large scale reasoning training with reinforcement learning, and then what the DeepSeek paper detailed in this R1 paper, which for me is one of the big open questions on how do you do this, is that they did reasoning heavy, but very standard post-training techniques after the large scale reasoning RL. So they did the same things with a form of instruction tuning through rejection sampling, which is essentially heavily filtered instruction tuning with some reward models. And then they did this RLHF, but they made it math-heavy. So, some of this transfer, we looked at this philosophical example early on. One of the big open questions is, how much does this transfer? If we bring in domains after the reasoning training, are all the models going to become eloquent writers by reasoning? Is this philosophy stuff going to be open? We don't know in the research of how much this will transfer. There's other things about how we can make soft verifiers and things like this, but there is more training after reasoning, which makes it easier to use these reasoning models. And that's what we're using right now. So if we're going to talk about o3-mini and o1, these have gone through these extra techniques that are designed for human preferences after being trained to elicit reasoning.

Dylan Patel
I think one of the things that people are ignoring is Google's Gemini Flash Thinking is both cheaper than R1 and better, and they released it in the beginning of December -

Nathan Lambert
And nobody's talking about it.

Dylan Patel
No one cares -

Nathan Lambert
It has a different flavor to it. Its behavior is less expressive than something like o1 or it has fewer tracks than it is on. Qwen released a model last fall, QwQ, which was their preview reasoning model, and DeepSeek had R1-Lite last fall, where these models kind of felt like they're on rails where they really, really only can do math and code and o1, it can answer anything. It might not be perfect for some tasks, but it's flexible, it has some richness to it, and this is kind of the art of is a model a little bit undercooked? It's good to get a model out the door, but it's hard to gauge and it takes a lot of taste to be like, is this a full-fledged model? Can I use this for everything? They're probably more similar for math and code. My quick read is that Gemini Flash is not trained the same way as o1, but taking an existing training stack, adding reasoning to it, so taking a more normal training stack and adding reasoning to it, and I'm sure they're going to have more. I mean they've done quick releases on Gemini Flash, reasoning, and this is the second version from the holidays. It's evolving fast and it takes longer to make this training stack where you're doing this large scale RL-

Dylan Patel
Ask it the same question from earlier, the one about the -

Nathan Lambert
The human nature.

Dylan Patel
Yeah.

Lex Fridman
What was the human nature one?

Nathan Lambert
Why I can ramble about this so much is that we've been working on this at AI Tube before o1 was fully available to everyone and before R1, which is essentially using this RL training for fine-tuning. We use this in our Tülu series of models and you can elicit the same behaviors where you say weight and such on, but it's so late in the training process that this kind of

reasoning expression is much lighter. So there's essentially a gradation and just how much of this RL training you put into it determines how the output looks.

Lex Fridman
So, we're now using Gemini 2.0 Flash Thinking Experimental 121.

Nathan Lambert
It summarized the problem as humans self-domesticated apes.

Lex Fridman
Okay. All right. So, wait, is this reviewing the reasoning? Here's why this is a novel. Okay.

Dylan Patel
You can click to expand.

Nathan Lambert
Oh, yeah, click to expand.

Lex Fridman
Okay. Analyze the request. Novel is the keyword.

Nathan Lambert
See how it just looks a little different? It looks like a normal output.

Lex Fridman
Yeah. I mean in some sense, it's better structured. It makes more sense. And -

Dylan Patel
Oh, and it latched onto human and then it went into organisms and… Oh, wow.

Lex Fridman
Apex Predator. Focus on domestication. Apply domestication to humans. Explore the idea of self-domestication.

Nathan Lambert
Not good, not good.

Lex Fridman
Where is this going? Refine, articulate the insight. Greater facial expressiveness and communication ability, yes. Plasticity and adaptability, yes. Dependence on social groups, yes. All right. And self-critique, refine further. Wow. Is this truly novel? Is it well-supported? So on and so forth. And the insight it's getting at is humans are not just social animals but

profoundly self-domesticated apes. And this self-domestication is the key to understanding our unique cognitive and social abilities. Self-domesticated apes. Self-domesticated -

Nathan Lambert
I prefer the DeepSeek response.

Lex Fridman
I mean it's novel. The insight is novel. I mean that's like a good book title; Self-Domesticated Apes. There could be a case made for that. I mean, yeah, it's cool and it's revealing the reasoning. It's magical. It's magical. This is really powerful. Hello, everyone, this is Lex with a quick intermission recorded after the podcast since we've reviewed responses from DeepSeek R1 and Gemini Flash 2.0 Thinking during this conversation, I thought at this moment it would be nice to insert myself quickly doing the same for OpenAI o1-pro and o3-mini with the same prompt. The prompt being, give one truly novel insight about humans. And I thought I would, in general, give my vibe check and vibe based anecdotal report on my own experiences with the new o3-mini model now that I got a chance to spend many hours with it in different kinds of context and applications. So, I would probably categorize this question as let's say open- ended philosophical question. And in particular, the emphasis on novelty I think is a nice way to test one of the capabilities of the model, which is come up with something that makes you pause and almost surprise you with brilliance. So that said, my general review after running each of the models on this question a bunch of times is that o1-pro consistently gave brilliant answers, ones that gave me pause and made me think, both cutting in its insight and just really nicely phrased with wit, with clarity, with nuance over and over, consistently generating the best answers. After that is R1, which was less consistent, but again, delivered brilliance. Gemini Flash 2.0 Thinking was third and last was o3-mini actually. It often gave quite a generic answer, at least to my particular sensibilities. That said, in a bunch of other applications that I tested for brainstorming purposes, it actually worked extremely well and often outperformed R1. But on this open-ended philosophical question, it did consistently worse. Now another important element for each of these models is how the reasoning is presented. DeepSeek R1 shows the full chain of thought tokens, which I personally just love. For these open-ended philosophical questions, it's really, really interesting to see the model think through it, but really also just stepping back, me as a person who appreciates intelligence and reasoning and reflection, reading these kind of chain of thought raw tokens of R1, there's something genuinely beautiful about observing the path of deliberation in an intelligent system. I think we don't always have that explicitly laid out for us humans. So, to see it in another intelligence system, the nonlinearity of it akin to the Ulysses, Finnegans Wake by James Joyce. It's just beautiful to watch. Anyways, we discussed in the episode DeepSeek R1 talked about humans being able to convert selfish desires into cooperative systems by collectively pretending abstract rules like money laws and rights are real. And these shared hallucinations act as games where competition is secretly redirected to benefit the group turning conflict into society's fuel. Gemini 2.0 Flash Thinking said,

“Humans are not just social animals but self-domesticated apes. And this self domestication is the key to understanding our unique cognitive and social abilities.” Now, it’s important to say that the chain of thought there was really interesting. It was looking through the entire evolution of life on earth considering apex predators and considering how from that, we ended up to where we are. I think that domestication by choice is a really interesting angle. Again, it’s one of those things when somebody presents a different angle on a seemingly obvious thing, it just makes me smile. And the same with DeepSeek R1, that these hallucinations of money laws and rights and us collectively pretending like it’s real and we play games with them that look like competition when secretly we’re just cooperating with each other and that is the fuel of progress. Beautifully put. Now, OpenAI o1-pro consistently, over and over delivered bangers. I can go through many of them, but the first one was, “Humans are the only species that turns raw materials into symbolic resources. Then uses those symbols to reorganize the very materials that came from creating a closed feedback loop between meaning and matter.” Here, I just ran it again. Banger after banger, I’m telling you. “Humans are unique among known species in that they simultaneously rewrite two layers of reality; the external world and their own private mental landscapes. And then merge these two rewritten layers into a continuous personal narrative that feels objectively true.” Feels true. This is poetry. Okay. And then o3-mini high, for me, was smart, fast actually, and kind of generic. Never quite got there for me. So here’s the first one I got from o3-mini, “Humans are not fixed beings, but rather ongoing narratives, dynamic stories that we continuously write, edit, and reinterpret. This narrative plasticity is more than just memory or self-reflection. It’s an intrinsic cognitive process that acts like an internal error correction system. It allows us to adapt our identities and values over time in response to new experiences, challenges, and social contexts.” Now, it almost sneaks up to something approximating cutting insight with narrative plasticity in quotes. But then it goes back to the generic. I don’t know. All of these models are incredible for different reasons. There’s a lot of concerns as we discussed in this episode, but there’s a lot of reasons to be excited as well. And I’ve probably spoken for too long. I am severely sleep-deprived, borderline delirious. So hopefully some of this made sense. And now, dear friends, back to the episode.

Dylan Patel
I think to Nathan’s point, when you look at the reasoning models, to me, even when I used R1 versus o1, there was that sort of rough edges around the corner feeling. And Flash Thinking earlier, I didn’t use this version, but the one from December, and it definitely had that rough edges around the corner feeling where it’s just not fleshed out in as many ways. Sure, they added math and coding capabilities via these verifiers in RL, but it feels like they lost something in certain areas. And o1 is worse performing than Chat in many areas as well, to be clear -

Nathan Lambert
Not by a lot.

Dylan Patel
Not by a lot though, right? And R1 definitely felt to me like it was worse than V3 in certain areas, like doing this RL expressed and learned a lot, but then it weakened in other areas. And so I think that's one of the big differences between these models and what one offers. And then OpenAI has o1-pro, and what they did with o3, which is also very unique, is that they stacked search on top of chain of thought. And so chain of thought is one thing where it's one chain, it backtracks, goes back and forth, but how they solved the ARC-AGI challenge was not just the chain of thought, it was also sampling many times, i.e., running them in parallel and then selecting.

Nathan Lambert
Is running in parallel actually search? Because I don't know if we have the full information on how o1-pro works. So, I don't have enough information-

Dylan Patel
Agreed.

Nathan Lambert
… to confidently say that it is search.

Dylan Patel
It is parallel samples.

Nathan Lambert
Yeah. And then what.

Dylan Patel
And then it selects something.

Nathan Lambert
And we don't know what the selection function is. The reason why we're debating is because since o1 was announced, there's been a lot of interest in techniques called Monte Carlo Tree Search, which is where you will break down the chain of thought into intermediate steps. We haven't defined chain of thought. Chain of thought is from a paper from years ago where you introduced the idea to ask a language model that at the time was much less easy to use, you would say, "Let's verify step by step," and it would induce the model to do this bulleted list of steps. Chain of thought is now almost a default in models where if you ask it a math question, you don't need to tell it to think step by step. And the idea with Monte Carlo Tree Search is that you would take an intermediate point in that train, do some sort of expansion, spend more compute, and then select the right one. That's a very complex form of search that has been used in things like MuZero and AlphaZero, potentially. I know MuZero does this.

Dylan Patel
Another form of search is just asking five different people and then taking the majority answer. There's a variety of, it could be complicated, it could be simple. We don't know what it is, just that they are not just issuing one chain of thought in sequence. They're launching many in parallel and in the ARC-AGI, they launched a thousand in parallel for the one that really shocked everyone that beat the benchmark was they would launch a thousand in parallel and then they would get the right answer like 80% of the time or 70% of the time, 90 maybe even. Whereas if they just launched one, it was like 30%.

Nathan Lambert
There are many extensions to this. I would say the simplest one is that our language models to date have been designed to give the right answer the highest percentage of the time in one response. And we are now opening the door to different ways of running inference on our models in which we need to reevaluate many parts of the training process, which normally opens the door to more progress, but we don't know if OpenAI changed a lot or if just sampling more and multiple choice is what they're doing or if it's something more complex, but they changed the training and they know that the inference mode is going to be different.

Lex Fridman
So we're talking about o1-pro, $200 a month and they're losing money. The thing that we're referring to, this fascinating exploration of the test time compute space, is that actually possible? Do we have enough compute for that? Does the financials make sense?

Dylan Patel
So the fantastic thing is, and it's in the thing that I pulled up earlier, but the cost for GPT-3 has plummeted if you scroll up just a few images, I think. The important thing about, hey, is cost a limiting factor here? My view is that we'll have really awesome intelligence, like AGI, before we have it permeate throughout the economy. And this is sort of why that reason is. GPT-3 was trained in what? 2020? 2021? And the cost for running inference on it was $60, $70 per million tokens, which was the cost per intelligence was ridiculous. Now as we scaled forward two years, we've had a 1200x reduction in cost to achieve the same level of intelligence as GPT-3.

Lex Fridman
So here on the x-axis is time over just a couple of years, and on the y-axis is log scale dollars to run inference on a million tokens.

Nathan Lambert
Yeah, it's dollar to million.

Lex Fridman
So you have just a linear decline on log scale from GPT-3 through 3.5 to Llama -

Dylan Patel
It's like five cents or something like that now, right? Versus $60, 1200x, that's not the exact numbers, but it's 1200x, I remember that number, is humongous cost per intelligence. Now, the freak out over DeepSeek is, "Oh my god, they made it so cheap." It's like actually, if you look at this trend line, they're not below the trend line first of all, at least for GPT-3, right? They are the first to hit it, which is a big deal, but they're not below the trend line as far as GPT-3. Now we have GPT-4, what's going to happen with these reasoning capabilities? It's a mix of architectural innovations, it's a mix of better data, and it's going to be better training techniques and all of these better inference systems, better hardware going from each generation of GPU to new generations or ASICs. Everything is going to take this cost curve down and down and down and down. And then can I just spawn a thousand different LLMs to create a task and then pick from one of them? Or whatever search technique, I want, a Tree, Monte Carlo Tree Search, maybe it gets that complicated, maybe it doesn't because it's too complicated to actually scale. Who knows? Better lesson, right? The question is, I think, when not if, because the rate of progress is so fast. Nine months ago, Dario said nine months ago the cost to train an inference was this, and now we're much better than this and DeepSeek is much better than this. And that cost curve for GPT-4, which was also roughly $60 per million tokens when it launched, has already fallen to $2 or so. And we're going to get it down to cents probably for GPT-4 quality. And then that's the base for the reasoning models like o1 that we have today and o1-pro is spawning multiple and o3 and so on and so forth, these search techniques, too expensive today, but they will get cheaper and that's what's going to unlock the intelligence.

Lex Fridman
So, it'll get cheaper and cheaper and cheaper. The big DeepSeek R1 release freaked everybody out because of the cheaper. One of the manifestations of that is NVIDIA stock plummeted. Can you explain what happened? And also just explain this moment and if NVIDIA is going to keep winning.

Nathan Lambert
We are both NVIDIA bulls here, I would say. And in some ways, the market response is reasonable. NVIDIA's biggest customers in the US are major tech companies and they're spending a ton on AI. And if a simple interpretation of DeepSeek is you can get really good models without spending as much on AI. So in that capacity it's like, "Oh, maybe these big tech companies won't need to spend as much in AI and go down." The actual thing that happened is much more complex where there's social factors, where there's the rising in the app store, the social contagion that is happening. And then I think some of it is just like, I don't trade, I don't know anything about financial markets, but it builds up over the weekend, the social pressure, where it's like if it was during the week and there was multiple days of

trading when this was really becoming, but it comes on the weekend and then everybody wants to sell, and then that is a social contagion.

Dylan Patel
I think, and there were a lot of false narratives, which is like, "Hey, these guys are spending billions on models," and they're not spending billions on models. No one spent more than a billion dollars on a model that's released publicly. GPT-4 was a couple hundred million and then they've reduced the cost with 4o, 4 Turbo, 4o, right? But billion dollar model runs are coming and this concludes pre-training and post-training, right? And then the other number is like, "Hey, DeepSeek didn't include everything." They didn't include a lot of the cost goes to research and all this sort of stuff. A lot of the cost goes to inference. A lot of the cost goes to post-training. None of these things were factored. Research, salaries, all these things are counted in the "billions of dollars" that OpenAI is spending, but they weren't counted in the, "Hey, $6 million, $5 million that DeepSeek spent." So, there's a bit of misunderstanding of what these numbers are, and then there's also an element of… NVIDIA has just been a straight line up and there's been so many different narratives that have been trying to push down NVIDIA. I don't say push down NVIDIA stock. Everyone is looking for a reason to sell or to be worried. It was Blackwell delays, right? Their GPU, every two weeks there's a new report about their GPUs being delayed. There's the whole thing about scaling laws ending, right? It's so ironic -

Nathan Lambert
It lasted a month.

Dylan Patel
It was literally just, "Hey, models aren't getting better." They're just not getting better. There's no reason to spend more, pre-training scaling is dead. And then it's like o1, o3, right?

Nathan Lambert
R1.

Dylan Patel
R1, right? And now it's like, "Wait, models, they're progressing too fast. Slow down the progress, stop spending on GPUs." But the funniest thing I think that comes out of this is Jevons paradox is true. AWS pricing for H100s has gone up over the last couple of weeks, since a little bit after Christmas, since V3 was launched, AWS H100 pricing has gone up. H200s are almost out of stock everywhere because H200 has more memory and therefore R1 wants that chip over H100, right?

Nathan Lambert
We were trying to get GPUs on a short notice this week for a demo and it wasn't that easy. We were trying to get just 16 or 32 H100s for demo and it was not very easy.

Lex Fridman
So for people who don't know, Jevons paradox is when the efficiency goes up, somehow magically, counter intuitively, the total resource consumption goes up as well.

Dylan Patel
And semiconductors is 50 years of Moore's law, every two years half the cost, double the transistors, just like clockwork and it's slowed down obviously, but the semiconductor industry has gone up the whole time. It's been wavy, right? There's obviously cycles and stuff and I don't expect AI to be any different. There's going to be ebbs and flows, but in AI, it's just playing out at an insane timescale. It was 2x every two years, this is 1200x in like three years. So it's like the scale of improvement is hard to wrap your head around.

Lex Fridman
Yeah. I was confused because to me, NVIDIA stock on that should have gone up, but maybe it went down because there's suspicion of foul play on the side of China, something like this. But if you just look purely at the actual principles at play here, it's obvious. Yeah, the Jevons paradox -

Nathan Lambert
The more progress that AI makes or the higher the derivative of AI progress is, especially because NVIDIA's in the best place, the higher the derivative is, the sooner the market's going to be bigger and expanding and NVIDIA's the only one that does everything reliably right now.

Lex Fridman
Yeah, because it's not like an NVIDIA competitor arose. It's another company that's using NVIDIA -

Nathan Lambert
Who historically has been a large NVIDIA customer.

Dylan Patel
And has press releases about them cheering about being China's biggest NVIDIA customer, right?

Lex Fridman
Yeah. I mean -

Dylan Patel
Obviously they've quieted down, but I think that's another element of it is that they don't want to say how many GPUs they have because hey, yes, they have H800s, yes, they have H20s, they also have some H100s, right? Which were smuggled in.

Lex Fridman
Can you speak to that, to the smuggling? What's the scale of smuggling that's feasible for a nation state to do for companies? Is it possible to -

Dylan Patel
I think there's a few angles of "smuggling" here, right? One is ByteDance, arguably is the largest smuggler of GPUs for China. China's not supposed to have GPUs. ByteDance has over 500,000 GPUs. Why? Because they're all rented from companies around the world. They rent from Oracle, they rent from Google, they rent from all these, and a bunch of smaller cloud companies too, right? All the "neoClouds" of the world. They rent so, so many GPUs. They also buy a bunch. And they do this for mostly what Meta does, right? Serving TikTok, right? Serving next best-

Nathan Lambert
Separate discussion.

Dylan Patel
Same as Meta, right? To be clear, today, that's the use, right? And it's a valid use, right? Hack the dopamine circuit. Now, that's theoretically now very much restricted with the AI diffusion rules, which happened in the last week of the Biden admin, and Trump admin looks like they're going to keep them, which limits allies even, like Singapore, which Singapore is 20%, 30% of NVIDIA's revenue, but Singapore's had a memoratorium on not building data centers for 15 years because they don't have enough power. So, where are they going?

Nathan Lambert
Oh, my God.

Dylan Patel
I'm not claiming they're all going to China, but a portion, many are going to Malaysia, including Microsoft and Oracle have big data centers in Malaysia. They're going all over Southeast Asia probably, India as well. There's stuff routing, but the diffusion rules are very de facto, like you can only buy this many GPUs from this country and you can only rent a cluster this large to companies that are Chinese. They're very explicit on trying to stop smuggling. And a big chunk of it was, hey, random company buys 16 servers, ships them to China. There's actually, I saw a photo from someone in the semiconductor industry who leads a team for networking chips that competes with NVIDIA, and he sent a photo of a guy checking into a first class United flight from San Francisco to Shanghai or Shenzhen with a super micro box that was this big, which can only contain GPUs, right? And he was booking first class because think about it, 3K to 5K for your first class ticket, server costs $240,000 in the US, $250,000, you sell it for $300,000 in China. Wait, you just got a free first class ticket and a lot more money. So it's like… And that's small scale smuggling. Most of the large

scale smuggling is companies in Singapore and Malaysia routing them around or renting GPUs, completely legally-

Nathan Lambert
I want to jump in. How much does this scale? I think there's been some people that are higher level economics understanding say that as you go from 1 billion of smuggling to 10 billion, it's like you're hiding certain levels of economic activity and that's the most reasonable thing to me is that there's going to be some level where it's so obvious that it's easier to find this economic activity. And -

Dylan Patel
Yeah. So, my belief is that last year roughly, so NVIDIA made a million H20s, which are legally allowed to be shipped to China, which we talked about is better for reasoning, inference at least, not training, but reasoning inference and inference generally. Then they also had a couple hundred thousand, we think like 200,000 to 300,000 GPUs were routed to China from Singapore, Malaysia, US, wherever. Companies spawn up, buy 16 GPUs, 64 GPUs, whatever it is, route it, and Huawei is known for having spent up a massive network of companies to get the materials they need after they were banned in 2018. So, it's not otherworldly, but I agree, right? Nathan's point is like, hey, you can't smuggle $10 billion of GPUs. And then the third source, which is just now banned, which wasn't considered smuggling, but is China is renting, I believe from our research, Oracle's biggest GPU customer is ByteDance. And for Google, I think it's their second-biggest customer. And you go down the list of clouds and especially these smaller cloud companies that aren't the "hyperscalers," think beyond CoreWeave and Lambda even, there's 60 different new cloud companies serving NVIDIA GPUs. I think ByteDance is renting a lot of these, all over it, right? And so these companies are renting GPUs to Chinese companies, and that was completely legal up until the diffusion rules, which happened just a few weeks ago. And even now, you can rent GPU clusters that are less than 2,000 GPUs, or you can buy GPUs and ship them wherever you want if there are less than 1,500 GPUs. There are still some ways to smuggle, but yeah, as the numbers grow a hundred something billion dollars of revenue for NVIDIA last year, 200 something billion this year, and if next year, it could nearly double again or more than double based on what we see with data center footprints being built out all across the US and the rest of the world, it's going to be really hard for China to keep up with these rules. Yes, there will always be smuggling and DeepSeek level models, GPT-4 level models, o1 level models capable to train on what China can get, even the next tier above that. But if we speed run a couple more jumps to billion dollar models, $10 billion models, then it becomes, "Hey, there is a compute disadvantage for China for training models and serving them." And the serving part is really critical, right? DeepSeek cannot serve their model today. It's completely out of inventory. It's already started falling in the app store actually, downloads, because you download it, you try and sign up, they say, "We're not taking registrations," because they have no capacity. You open it up, you get less than five tokens per second, if you even get your request approved, right? Because there's just no

capacity because they just don't have enough GPUs to serve the model, even though it's incredibly efficient.

Lex Fridman
It'd be fascinating to watch the smuggling. Because I mean there's drug smuggling, right? That's a market. There's weapons smuggling. And GPUs will surpass that at some point.

Nathan Lambert
Chips are highest value per kilogram probably by far. I have another question for you, Dylan. Do you track model API access internationally? How easy is it for Chinese companies to use hosted model APIs from the US?

Dylan Patel
Yeah. I mean that's incredibly easy, right? OpenAI publicly stated DeepSeek uses their API and they say they have evidence, right? And this is another element of the training regime, is people at OpenAI have claimed that it's a distilled model, i.e., you're taking OpenAI's model, you're generating a lot of output, and then you're training on the output in their model. And even if that's the case, what they did is still amazing by the way, what DeepSeek did, efficiency-wise.

Nathan Lambert
Distillation is standard practice in industry. Whether or not, if you're at a closed lab where you care about terms of service and IP closely, you distill from your own models. If you are a researcher and you're not building any products, you distill from the OpenAI models -

Lex Fridman
This is a good opportunity. Can you explain big picture distillation as a process? What is distillation? What's the process of distillation?

Nathan Lambert
We've talked a lot about training language models. They are trained on text and post-training, you're trying to train on very high-quality texts that you want the model to match the features of, or if you're using RL, you're letting the model find its own thing. But for supervised fine-tuning, for preference data, you need to have some completions, what the model is trying to learn to imitate. And what you do there is instead of a human data or instead of the model you're currently training, you take completions from a different, normally more powerful, model. I think there's rumors that these big models that people are waiting for, these GPT-5s of the world, the Claude 3 Opuses of the world are used internally to do this distillation process at OpenAI-

Dylan Patel
There's also public examples, right? Like Meta explicitly stated, not necessarily distilling, but they used 405B as a reward model for 70B in their Llama 3.2 or 3.3 rule -

Nathan Lambert
Yes. This is all the same topic.

Lex Fridman
So, is this ethical? Is this legal? Why is that Financial Times article headline say, "OpenAI says that there's evidence that China's DeepSeek used its model to train competitor."

Nathan Lambert
This is a long, at least in the academic side and research side, it has a long history because you're trying to interpret OpenAI's rule. OpenAI's terms of service say that you cannot build a competitor with outputs from their models. Terms of service are different than a license, which are essentially a contract between organizations. So if you have a terms of service on OpenAI's account, if I violate it, OpenAI can cancel my account. This is very different than a license that says how you could use a downstream artifact. So a lot of it hinges on a word that is very unclear in the AI space, which is, what is a competitor?

Dylan Patel
And then the ethical aspect of it is like, why is it unethical for me to train on your model when you can train on the internet's text? Right?

Lex Fridman
So there's a bit of a hypocrisy because OpenAI and potentially most of the companies trained on the internet's text without permission.

Nathan Lambert
There's also a clear loophole, which is that I generate data from OpenAI and then I upload it somewhere and then somebody else trains on it and the link has been broken. They're not under the same terms of service contract.

Dylan Patel
This is why -

Nathan Lambert
There's a lot of... There's a lot of to be discovered details that don't make a lot of sense.

Dylan Patel
This is why a lot of models today, even if they train on zero OpenAI data, you ask the model, "Who trained you?" It'll say, "I'm ChatGPT trained by OpenAI," because there's so much copy

paste of OpenAI outputs from that on the internet that you just weren't able to filter it out and there was nothing in the RL where they implemented or post-training or SFT, whatever, that says, "Hey, I'm actually a model by Allen Institute instead of OpenAI."

Nathan Lambert
We have to do this if we serve a demo. We do research and we use OpenAI APIs because it's useful and we want to understand post-training and our research models, they all say they're written by OpenAI unless we put in the system prop that we talked about that, "I am Tülu. I am a language model trained by the Allen Institute for AI." And if you ask more people around industry, especially with post-training, it's a very doable task to make the model say who it is or to suppress the OpenAI thing. So in some levels, it might be that DeepSeek didn't care that it was saying that it was by OpenAI. If you're going to upload model weights, it doesn't really matter because anyone that's serving it in an application and cares a lot about serving is going to, when serving it, if they're using it for a specific task, they're going to tailor it to that and it doesn't matter that it's saying it's ChatGPT.

Lex Fridman
Oh, I guess one of the ways to do that is like a system prompt or something like that? If you're serving it to say that you're -

Nathan Lambert
That's what we do. If we host a demo, you say, "You are Tülu 3, a language model trained by the Allen Institute for AI." We also are benefited… model trained by the Allen Institute for AI. We also are benefited from OpenAI data because it's a great research tool.

Lex Fridman
Do you think there's any truth and value to the OpenAI's claim that there's evidence that China's DeepSeek used this model to train?

Dylan Patel
I think everyone has benefited regardless because the data's on the internet. And therefore, it's in your per training now. There are subreddits where people share the best ChatGPT outputs, and those are in your model -

Nathan Lambert
I think that they're trying to shift the narrative. They're trying to protect themselves. We saw this years ago when ByteDance was actually banned from some OpenAI APIs for training on outputs. There's other AI startups that most people, if you're in the AI culture, were like they just told us they trained on OpenAI outputs and they never got banned. That's how they bootstrapped their early models. So, it's much easier to get off the ground using this than to set up human pipelines and build a strong model. So there's long history here, and a lot of the communications are seem like narrative [inaudible 03:31:00].

Dylan Patel
Actually, over the last couple of days, we've seen a lot of people distill DeepSeek's model into Llama models, because the DeepSeek models are complicated to run inference on because they're mixture of experts and they're 600 plus billion parameters and all of this. And people distilled them into the Llama models because the Llama models are so easy to serve, and everyone's built the pipelines and tooling for inference with the Llama models because it's the open standard. So, we've seen a sort of roundabout. Is it bad? Is it illegal? Maybe it's illegal, whatever. I don't know about that, but-

Nathan Lambert
It could break contracts. I don't think it's illegal in any legal… No one's going to jail for this, ever.

Lex Fridman
Fundamentally, I think it's ethical, or I hope it's ethical because the moment it becomes… We ban that kind of thing, it's going to make everybody much worse off. And I also, actually… this is difficult, but I think you should be allowed to train on the internet. I know a lot of authors and creators are very sensitive about it. That's a difficult question. But the moment you're not allowed to train on the internet-

Nathan Lambert
I agree.

Dylan Patel
I have a schizo take on how you can solve this. Because it already works.

Lex Fridman
All right.

Nathan Lambert
I have a reasonable take out of it.

Lex Fridman
All right.

Dylan Patel
So, Japan has a law which you're allowed to train on any training data and copyrights don't apply if you want to train a model, A. B, Japan has 9 gigawatts of curtailed nuclear power. C, Japan is allowed under the AI diffusion rule to import as many GPUs as they'd like. So, all we have to do… We have a market here to make. We build massive data centers, we rent them to the labs, and then we train models in a legally permissible way, and there's no ifs, ands, or

buts. And now, the models have no potential copyright lawsuit from New York Times or anything like that. No, it's just completely legal.

Nathan Lambert
Now, so -

Lex Fridman
Genius.

Nathan Lambert
... the early copyright lawsuits have fallen in the favor of AI training. I would say that the long tail of use is going to go inside of AI, which is if you scrape trillions of tokens of data, you're not looking and saying, "This one New York Times article is so important to me." But if you're doing a audio generation for music or image generation, and you say, "Make it in the style of X person," that's a reasonable case where you could figure out what is their profit margin on inference. I don't know if it's going to be the 50/50 of YouTube Creator Program or something, but I would opt into that program as a writer, please. It's going to be a rough journey, but there will be some solutions like that that makes sense. But there's a long tail where it's just on the internet.

Lex Fridman
I think one of the other aspects of that Financial Times article implied, and so that leads to a more general question. Do you think there's... How difficult is spying, espionage, and stealing of actual secret code and data from inside of companies? How much of that is being attempted?

Nathan Lambert
Code and data is hard, but ideas is easy. Silicon Valley operates on the way that top employees get bought out by other companies for a pay raise, and a large reason why these companies do this is to bring ideas with them. And there's no... I mean, in California, there's rules that certain non-competes or whatever are illegal in California. And whether or not there's NDAs and things, that is how a lot of it happens. Recently, there was somebody from Gemini who helped make this 1 million context length. And everyone is saying the next Llama who, he went to the Meta team, is going to have 1 million context length. And that's kind of how the world works.

Dylan Patel
As far as industrial espionage and things, that has been greatly successful in the past. The Americans did it to the Brits, the Chinese have done it to the Americans, and so on and so forth. It is a fact of life. And so, to argue industrial espionage can be stopped is probably unlikely. You can make it difficult. But even then, there's all these stories about like, "Hey, F35 and F22 have already been given to China in terms of design plans and stuff." Code and

stuff between, I say companies, not nation states, is probably very difficult. But ideas are discussed a lot, whether it be a house party in San Francisco or a company changing employees or always the mythical honeypot that always gets talked about. Someone gets honeypotted because everyone working on AI is a single dude who's in their 20s and 30s. Not everyone, but insane amount of… Insane percentages. So, there's always all these… And obviously-

Lex Fridman
So, honeypotted is like a female spy approaches you and…

Dylan Patel
Yeah. Or male, right? It's San Francisco. But as a single dude, I will say in his late 20s, we are very easily corrupted. Not corrupted myself, but we are. Right?

Lex Fridman
Yeah. Everybody else. Not me.

Nathan Lambert
I'm too oblivious that I am not single, so I'm safe from one espionage access.

Lex Fridman
Yeah. You have to make sure to close all security vulnerabilities. So you, Dylan, collect a lot of information about each of the mega clusters for each of the major AI companies. Can you talk about the buildouts for each one that stand out?

Dylan Patel
Yeah. I think the thing that's really important about these mega cluster buildouts is they're completely unprecedented in scale. US data center power consumption has been slowly on the rise and it's gone up to 2, 3% even through the cloud computing revolution. Data center consumption has a percentage of total US, and that's been over decades of data centers, etc. It's been climbing slowly, but now, 2 to 3%. Now, by the end of this decade, it's… Even under… When I say 10%, a lot of people that are traditionally by 2028 to 2030, people traditionally non-traditional data center people, that's nuts. But then, people who are in AI who have really looked at this like the Anthropics and OpenAI's, are like, "That's not enough." And I'm like, "Okay." But this is both through globally distributed or distributed throughout the US as well as centralized clusters. The distributed throughout the US is exciting and it's the bulk of it. Like, hey, OpenAI or, say, Meta's adding a gigawatt, but most of it is distributed through the US for inference and all these other things.

Lex Fridman
So maybe, we should lay out what a cluster is. So, does this include AWS? Maybe, it's good to talk about the different kinds of clusters. What you mean by mega clusters? What's the GPU and what's a compute or… And what [inaudible 03:37:41]-

Dylan Patel
Yeah.

Lex Fridman
Not that far back, but yeah. So, what do we mean by the clusters? The buildouts?

Dylan Patel
Oh, man. I thought I was about to do the Apple ad, what's a computer? So traditionally, data centers and data center tasks have been a distributed systems problem that is capable of being spread very far and widely. I.e, I send a request to Google, it gets routed to a data center somewhat close to me, it does whatever search ranking recommendation, sends a result back. The nature of the task is changing rapidly in that the task, there's two tasks that people are really focused on now. It's not database access. It's not, "Serve me the right page, serve me the right ad." It's now, a inference. An inference is dramatically different from traditional distributed systems, but it looks a lot more simple, similar. And then, there's training. The inference side is still like, "Hey, I'm going to put thousands of GPUs in blocks all around these data centers." I'm going to run models on them. User submits a request, it gets kicked off. Or hey, my service. They submit a request to my service. They're on Word and they're like, "Oh yeah, help me, Copilot," and it kicks it off. Or I'm on my windows, Copilot, whatever, Apple intelligence. Whatever it is, it gets kicked off to a data center. That data center does some work and sends it back. That's inference. That is going to be the bulk of compute, but then… And that's like, there's thousands of data centers that we're tracking with satellites and all these other things, and those are the bulk of what's being built. But the scale of… And so, that's what's really reshaping and that's what's getting millions of GPUs. But the scale of the largest cluster is also really important. When we look back at history or through the age of AI, it was a really big deal when they did AlexNet on, I think, 2 GPUs or 4 GPUs. I don't remember. It's a really big deal.

Nathan Lambert
It's a big deal because you use GPUs.

Dylan Patel
It's a big deal that they use GPUs and they use multiple. But then over time, its scale has just been compounding. And so when you skip forward to GPT-3, then GPT-4, GPT-4 20,000 A100 GPUs. Unprecedented run in terms of the size and the cost, right? A couple of hundred million dollars on a YOLO run for GPT-4, and it yielded this magical improvement that was perfectly in line with what was experimented and just a log scale right up.

Nathan Lambert
Oh, yeah. They had that plot from the paper.

Dylan Patel
The scaling of the technical part. The scaling laws were perfect, right? But that's not a crazy number. 20,000 A100's, roughly, each GPU is consuming 400 watts. And then when you add in the whole server, everything, it's like 15 to 20 megawatts of power. Maybe, you could look up what the power of consumption of a person is because the numbers are going to get silly, but 15 to 20 megawatts was standard data center size. It was just unprecedented that was all GPUs running one task.

Nathan Lambert
How many watts is a toaster?

Dylan Patel
A toaster has also -

Nathan Lambert
That's a good example.

Dylan Patel
… a similar power consumption to an A100. H100 comes around. They increase the power from 400 to 700 watts and that's just per GPU, and then there's all the associated stuff around it. So once you count all of that, it's roughly 1,200 to 1,400 watts for everything. Networking, CPUs, memory, blah, blah, blah.

Lex Fridman
So we should also say, what's required, you said power. So, a lot of power is required. A lot of heat is generated, so the cooling is required. And because there's a lot of GPUs or CPUs or whatever, they have to be connected. So, there's a lot of networking, right?

Dylan Patel
Yeah, I think - sorry for skipping past that. And then the data center itself is complicated, but these are still standard sized data centers for GPT-4 scale. Now, we step forward to what is the scale of clusters that people built last year, and it ranges widely. It ranges from like, "Hey, these are standard data centers. And we're just using multiple of them and connecting them together really with a ton of fiber between them, a lot of networking, etc." That's what OpenAI and Microsoft did in Arizona. They have 100,000 GPUs. Meta, similar thing. They took their standard existing data center design and it looks like an H, and they connected multiple of them together. They first did 24,000 GPUs total, only 16,000 of them were running on the training run because GPUs are very unreliable so they need to have spares to swap in and out. All the way to now, 100,000 GPUs that they're training on Llama 4

on currently. Like, 128,000 or so. Think about 100,000 GPUs with roughly 1,400 watts apiece. That's 140 megawatts, 150 megawatts for 128. So, you're talking about you've jumped from 15 to 20 megawatts to almost 10x that number, 9x that number, to 150 megawatts in two years from 2022 to 2024. And some people like Elon, that he admittedly… He says himself he got into the game a little bit late for pre-training large language models. xAI was started later, right? But then, he bent heaven and hell to get his data center up and get the largest cluster in the world, which is 200,000 GPUs. And he did that. He bought a factory in Memphis. He's upgrading the substation, with the same time, he's got a bunch of mobile power generation, a bunch of single cycle combine. He tapped the natural gas line that's right next to the factory, and he's just pulling a ton of gas, burning gas. He's generating all this power. He's in an old appliance factory that's shut down and moved to China long ago, and he's got 200,000 GPUs in it. And now, what's the next scale? All the hyperscalers have done this. Now, the next scale is something that's even bigger. And so Elon, just to stick on the topic, he's building his own natural gas plant, like a proper one right next door. He's deploying tons of Tesla Megapack batteries to make the power more smooth and all sorts of other things. He's got industrial chillers to cool the water down because he's water-cooling the chips. So, all these crazy things to get the clusters bigger and bigger. But when you look at, say, what OpenAI did with Stargate in Arizona, in Abilene Texas, right? What they've announced, at least. It's not built. Elon says they don't have the money. There's some debates about this. But at full scale, at least the first section is definitely money's accounted for, but there's multiple sections. But full scale, that data center is going to be 2.2 gigawatts, 2,200 megawatts of power in. And roughly, 1.8 gigawatts or 1,800 megawatts of power delivered to chips. Now, this is an absurd scale. 2.2 gigawatts is more than most cities, to be clear. Delivered to a single cluster that's connected to do training. To train these models, to do both the pre-training, the post-training, all of this stuff.

Lex Fridman
This is insane.

Nathan Lambert
It is. What is a nuclear power plant, again?

Dylan Patel
Everyone is doing this. Meta in Louisiana, they're building two natural gas plants. Massive ones. And then, they're building this massive data center. Amazon has plans for this scale. Google has plans for this scale. xAI has plans for this scale. All of these, the guys that are racing, the companies that are racing are racing hard, and they're doing multi-gigawatt data centers to build this out. Because they think that, "If I now have…" Obviously, pre-training scaling is going to continue, but to some extent. But then also, all this post-training stuff where you have RL Sandbox for computer use or whatever, this is where they're going to… And all these fearful viable domains where they just keep learning and learning and learning, self-play or whatever. Whatever it is makes the AI so much more capable because the line

does go up. As you throw more compute, you get more performance. This shirt is about scaling laws. To some extent, it is diminishing returns. You 10x the compute, you don't get 10x better model. You get a diminishing returns. But also, you get efficiency improvements, so you bend the curve. And these scale of data centers are just reeking a lot of havoc on the network. Nathan was mentioning Amazon has tried to buy this nuclear power plant Talen. And if you look at Talen's stock, it's just skyrocketing. They're building a massive multi-gigawatt data center there. You just go down the list, there's so many ramifications. Interesting thing is certain regions of the US transmitting power cost more than actually generating it because the grid is so slow to build. And the demand for power, and the ability to build power, and re-ramping on a natural gas plant or even a coal plant is easy enough to do, but transmitting the power's really hard. So in some parts of the US like in Virginia, it costs more to transmit power than it costs to generate it, which is there's all sorts of second-order effects that are insane here.

Lex Fridman
Can the power grid support this kind of growth?

Dylan Patel
Trump's executive orders... There was a Biden executive order before the end of the year, but then Trump had some more executive orders, which hopefully reduced the regulations to where, yes, things can be built. But yeah, this is a big, big challenge. Is building enough power fast enough?

Lex Fridman
Are you going to basically have a nuclear power plant next to a data center for each one of these?

Dylan Patel
The fun thing here is this is too slow to build the power plant. To build a power plant or to reconfigure an existing power plant, it's too slow. And so therefore, you must use - data center power consumption is flat, right? I mean -

Nathan Lambert
This is why nuclear is also good for it. Long term, nuclear is a very natural fit, but data -

Dylan Patel
Yes.

Nathan Lambert
You can't do solar or anything in the short term like that.

Dylan Patel
Because data center power's like this, right? You're telling me I'm going to buy tens of billions of dollars of GPUs and idle them because the power's not being generated? Power's cheap. If you look at the cost of a cluster, less than 20% of it is power. Most of it is the capital cost and depreciation of the GPUs. And so it's like, "Well, screw it. I'll just build natural gas plants." This is what Meta is doing in Louisiana, this is what OpenAI is doing in Texas, and all these different places. They may not be doing it directly, but they are partnered with someone. And so, there is a couple of hopes. One is… And Elon, what he's doing in Memphis is to the extreme. They're not just using dual combine cycle gas which is super efficient, he's also just using single cycle and mobile generators and stuff which is less efficient. But there's also the flip side, which is solar power generation is like this, and wind is another like this. Different correlate different. So if you stack both of those, plus you get a big chunk of batteries, plus you have a little bit of gas, it is possible to run it more green. It's just the time scales for that is slow. So, people are trying. But Meta basically said, "Whatever. I don't care about my sustainability pledge." Or they'll buy a power… It's called a PPA, Power Purchasing Agreement, where there'll be a massive wind farm or solar farm wherever. And then, they'll just pretend like those electrons are being consumed by the data center. But in reality, they're paying for the power here and selling it to the grid, and they're buying power here. And then another thing is Microsoft quit on some of their sustainability pledges. Elon, what he did with Memphis is objectively somewhat dirty, but he is also doing it in an area where there's a bigger natural gas plant right next door and a sewer next… Or not a sewer, but a wastewater treatment and a garbage dump nearby. And he's obviously made the world a lot more clean than that one data center is going to do, so I think it's fine to some extent. And maybe, AGI solves global warming and stuff, whatever it is. This is the attitude that people at the labs have, which is like, "Yeah, it's great. We'll just use gas," because the race is that important. And if we lose, that's way worse.

Lex Fridman
I should say that I got a chance to visit the Memphis data center.

Dylan Patel
Oh, wow.

Lex Fridman
And it's incredible. I mean, I visited with Elon. Just the teams and the rate of innovation there is insane. My sense is that nobody's ever done anything of this scale, and nobody has certainly ever done anything of this scale at the rate that xAI is doing. So, they're figuring out… I was sitting in on all of these meetings where they're brainstorming. It's insane. It's exciting because they're trying to figure out what the bottlenecks are, how to remove the bottlenecks, how to make sure that… There's just so many really cool things about putting together a data center because everything has to work. The people that do the sys admin, the machine learning and all of that is the exciting thing, so on. But really, the people that

run everything are the folks that know the low-level software and hardware that runs everything, the networking, all of that. So, you have to make sure you have procedures that test everything. I think they're using ethernet. I don't know how they're doing the networking, but-

Dylan Patel
They're using NVIDIA Spectrum-X Ethernet. I think the unsung heroes are the cooling in electrical systems which are just glossed over.

Lex Fridman
Yeah, exactly.

Dylan Patel
But I think one story that maybe exemplifies how insane this stuff is, is when you're training, you're always doing… You're running through the model a bunch, in the most simplistic terms. Running through the model a bunch, and then you're going to exchange everything and synchronize the weights. So, you'll do a step. This is like a step-in model training. And every step, your loss goes down hopefully, and it doesn't always. But in the simplest terms, you'll be computing a lot and then you'll exchange. The interesting thing is GPU power is most of it, networking power is some but it's a lot less. So while you're computing, your power for your GPUs is here. But then when you're exchanging weights, if you're not able to overlap communications and compute perfectly, there may be a time period where your GPUs are just idle, and you're exchanging weights and you're like, "Hey, the model's updating." So, you're exchanging the radiance, you do the model update, and then you start training again. So, the power goes… Right? And it's super spiky. And so funnily enough, when you talk about the scale of data center power, you can blow stuff up so easily. And so, Meta actually has accidentally upstreamed something to code in PyTorch where they added an operator. And I kid you not, whoever made this, I want to hug the guy because it says PyTorch… It's like PyTorch.powerplant no blow up equals 0 or equal 1. And what it does is amazing, right?

Lex Fridman
Yeah.

Dylan Patel
Either when you're exchanging the weights, the GPU will just compute fake numbers so the power doesn't spike too much, and so then the power plants don't blow up because the transient spikes screw stuff up.

Lex Fridman
Well, that makes sense. You have to do that kind thing. [inaudible 03:51:57] You have to make sure they're not idle.

Dylan Patel
And Elon's solution was like, "Let me throw a bunch of Tesla Megapacks and a few other things."

Lex Fridman
Yeah, to symbolize that.

Dylan Patel
Everyone has different solutions, but Meta's, at least, was publicly and openly known, which is just like, set this operator. And what this operator does is it just makes the GPUs compute nothing so that the power doesn't spike.

Lex Fridman
But that just tells you how much power you're working with. I mean, it's insane. It's insane.

Nathan Lambert
People should just go to Google, like scale or what does X watts do, and go through all the scales from 1 watt to a kilowatt to a megawatt. You look and stare at that, and you're how high on the list a gigawatt is, it's mind-blowing.

Lex Fridman
Can you say something about the cooling? I know Elon's using liquid cooling, I believe, in all cases. That's a new thing. Most of them don't use liquid cooling. Is there something interesting to say about the cooling?

Dylan Patel
Yeah. So, air cooling has been the de facto standard. Throw a bunch of metal heat pipes, etc, and fans, and that's cold. That's been enough to cool it. People have been dabbling in water cooling. Google's TPUs are water-cooled. So, they've been doing that for a few years. But with GPUs, no one's ever done... And no one's ever done the scale of water cooling that Elon just did. Now, next generation NVIDIA is for the highest-end GPU, it is mandatory water cooling. You have to water-cool it. But Elon did it on this current generation, and that required a lot of stuff. If you look at some of the satellite photos and stuff of the Memphis facility, there's all these external water chillers that are sitting. Basically, it looks like a semi truck pod thing. What's it called? The container? But really, those are water chillers, and he has 90 of those water chillers just sitting outside. Ninety different containers that chill the water, bring it back to the data center, and then you distribute it to all the chips, pull all the heat out and then send it back. And this is both a way to cool the chips, but also, it's an efficiency thing. And going back to that three vector thing, there is Memory Bandwidth FLOPS and interconnect. The closer the chips are together, the easier it is to do high-speed interconnects. And this is also a reason why you want to go water cooling is because you can just put the chips right next to each other, and therefore get higher speed connectivity.

Lex Fridman
I got to ask you, in one of your recent posts, there's a section called cluster measuring contest. So…

Dylan Patel
There's another word there, but I won't say it.

Lex Fridman
Who's got the biggest now and who's going to have the biggest?

Dylan Patel
Today, individual largest is Elon. Right?

Lex Fridman
Right. Elon's cluster.

Dylan Patel
Elon's cluster in Memphis, 200,000 GPUs. Meta has 128,000, OpenAI has 100,000 now. Now to be clear, other companies have more GPUs than Elon. They just don't have them in one place. And for training, you want them tightly connected. There's some techniques that people are researching and working on that let you train across multiple regions. But for the most part, you want them all in one area so you can connect them highly with high-speed networking. And so, Elon today has 200,000 H100s, 100,000 H100s and 100,000 H200s. Meta, OpenAI, and Amazon all have on the scale of a hundred thousand, a little bit less. But next this year, people are building much more. Anthrophic and Amazon are building a cluster of 400,000 trainium 2, which is Amazon-specific chip trying to get away from NVIDIA. Meta and OpenAI have scales for hundreds of thousands. But by next year, you'll have 500,000 to 700,000 GPU clusters. And note, those GPUs are much higher power consumption than existing ones. Hopper's 700 watts, Blackwell goes to 1,200 watts. So, the power per chip is growing and the number of chips is growing.

Lex Fridman
Nuts. Elon said he'll get to a million. Do you think that's actually feasible?

Dylan Patel
I mean, I don't doubt Elon. The filings that he has for the power plant and the Tesla battery packs, it's clear he has some crazy plans for Memphis. Permits and stuff is open record, but it's not quite clear what the time scales are. I just never doubt Elon. He's going to surprise us.

Lex Fridman
So, what's the idea with these clusters? If you have a million GPUs, what percentage in a, let's say 2 or 3 years, is used for training? What percent pre-training, and what percent is used for the actual computation?

Dylan Patel
These mega clusters make no sense for inference. You could route inference there and just not train. But most of the inference capacity is being, "Hey, I've got a 30-megawatt data center here, I've got 50 megawatts here, I've got 100 here." Whatever. I'll just throw inference in all of those because the mega clusters, multi-gigawatt data centers, I want to train there because that's where all of my GPUs are co-located where I can put them at a super high networking speed connected together. Because that's what you need for training. Now with pre-training, this is the old scale. You can increase parameters, you did increase data, model gets better. That doesn't apply anymore because there's not much more data in the pre-training side. Yes, there's video and audio and image that has not been fully taken advantage of, so there's a lot more scaling. But a lot of people have transcript, taken transcripts out of YouTube videos, and that gets you a lot of the data. It doesn't get you all of the learning value out of the video and image data, but... There's still scaling to be done on pre-training, but this post-training world is where all the FLOPS are going to be spent. The model's going to play with itself, it's going to self-play, it's going to do verifiable tasks, it's going to do computer use in sandboxes. It might even do simulated robotics things. All of these things are going to be environments where compute is spent in "post-training." But I think it's going to be good. We're going to drop the post from post-training.

Nathan Lambert
Yeah. Wow.

Dylan Patel
It's going to be pre-training and it's going to be training, I think, at some point. [inaudible 03:57:53] At some point. Because for bulk of the last few years, pre-training has dwarfed post-training. But with these verifiable methods, especially ones that scale really potentially infinitely, like computer use in robotics, not just math and coding where you can verify what's happening, those infinitely verifiable tasks, it seems you can spend as much compute as you want on this.

Nathan Lambert
Especially at the context length increase because the end of pre-training is when you increase the context length for these models. And we've talked earlier in the conversation about how the context length, when you have a long input, is much easier to manage than output. And a lot of these post-training and reasoning techniques rely on a ton of sampling, and it's becoming increasingly long context. So just like effectively, your compute efficiency goes down. I think FLOPS is the standard for how you measure it. But with RL, and you have

to do all of these things where you move your weights around in a different way than at pre-training and just generation, it's going to be become less efficient and FLOPS is going to be less of a useful term. And then as the infrastructure gets better, it's probably going to go back to FLOPS.

Lex Fridman
So, all of the things we've been talking about is most likely going to be NVIDIA, right? Is there any competitors of GPU?

Dylan Patel
Google kind of ignored them. I was getting -

Nathan Lambert
I was like, "Ah?"

Lex Fridman
What's the story with TPU? What's the -

Dylan Patel
TPU is awesome. It's great. Google is, they're a bit more tepid on building data centers for some reason. They're building big data centers, don't get me wrong, and they actually have the biggest cluster. I was talking about NVIDIA clusters. They actually have the biggest cluster. Period. But the way they do it is very interesting. They have two data center super regions in that the data center isn't physically… All of the GPUs aren't physically on one site but they're like 30 miles from each other. And they're not GPUs, TPUs. In Iowa and Nebraska, they have four data centers that are just right next to each other.

Lex Fridman
Why doesn't Google flex its cluster size?

Dylan Patel
Go to multi-data center training, there's good images in there. I'll show you what I mean. It's just semi-analysis multi-data center. This is an image of what a standard Google data center looks like. By the way, their data centers look very different than anyone else's data centers.

Lex Fridman
What are we looking at here?

Dylan Patel
So if you see this image, in the center, there are these big rectangular boxes. Those are where the actual chips are kept. And then if you scroll down a little bit further, you can see

there's these water pipes, there's these chiller cooling towers in the top, and a bunch of diesel generators. The diesel generators are backup power. The data center itself look physically smaller than the water chillers. The chips are actually easier to keep together, but then cooling all the water for the water cooling is very difficult. So, Google has a very advanced infrastructure that no one else has for the TPU. And what they do is they've stamped a bunch of these data centers out in a few regions. So if you go a little bit further down… This is a Microsoft. This is in Arizona. This is where GPT-5 "will be trained."

Nathan Lambert
If it doesn't exist already.

Dylan Patel
Yeah, if it doesn't exist already. But each of these data centers, I've shown a couple images of them, they're really closely co-located in the same region. Nebraska, Iowa. And then they also have a similar one in Ohio complex. And so, these data centers are really close to each other. And what they've done is they've connected them super high bandwidth with fiber. And so, these are just a bunch of data centers. And the point here is that Google has a very advanced infrastructure, very tightly connected in a small region. So, Elon will always to have the biggest cluster fully connected because it's all in one building, and he's completely right on that. Google has the biggest cluster but you have to spread over three sites, and by a significant margin. We have to go across multiple sites.

Lex Fridman
Why doesn't Google compete with NVIDIA? Why don't they sell TPUs?

Dylan Patel
I think there's a couple of problems with it. It's like, one, TPU has been a form of allowing search to be really freaking cheap and build models for that. And so, a big chunk of the search, GPU purchases or TPU purchases or big chunk of Google's purchases and usage, all of it is for internal workloads. Whether it be search, now Gemini, YouTube, all these different applications that they have ads. These are where all their TPUs are being spent and that's what they're hyper-focused on. And so, there's certain aspects of the architecture that are optimized for their use case that are not optimized elsewhere. One simple one is they've open-sourced a Gemma model, and they called it Gemma-7B. But then, it's actually 8 billion parameters because the vocabulary is so large. And the reason they made the vocabulary so large is because TPUs matrix multiply unit is massive because that's what they've optimized for. And so they decided, "Oh, well, I'll just make the vocabulary large, too." Even though it makes no sense to do so in such a small model, because that fits on their hardware. Gemma doesn't run it as efficiently on a GPU as a Llama does. But vice versa, Llama doesn't run as efficiently on a TPU as a Gemma does. There's certain aspects of hardware, software co-design. All their search models are there, ranking and recommendation models, all these different models that are AI but not like gen AI have been

hyper optimized with TPUs forever. The software stack is super optimized. But all of this software stack has not been released publicly at all. Very small portions of it. JAX and XLA have been. But the experience when you're inside of Google and you're training on TPUs as a researcher, you don't need to know anything about the hardware in many cases, right? It's pretty beautiful.

Nathan Lambert
They all loved it.

Dylan Patel
But as soon as you step outside -

Nathan Lambert
A lot of them go back. They leave Google and then they go back.

Lex Fridman
Yeah.

Dylan Patel
Yeah. They leave and they start a company because they have all of these amazing research ideas. And they're like, "Wait. Infrastructure's hard, software is hard." And this is on GPUs. Or if they try to use TPUs, same thing, because they don't have access to all this code. And so it's like, how do you convince a company whose golden goose is search where they're making hundreds of billions of dollars from, to start selling GPU or TPUs which they used to only buy a couple of billion of... I think in 2023, they bought a couple of billion. And now, they're buying like 10 billion to $15 billion worth. But how do you convince them that they should just buy twice as many and figure out how to sell them, and make $30 billion? Who cares about making $30 billion?

Lex Fridman
Won't that 30 billion exceed actually the search profit eventually?

Dylan Patel
You're always going to make more money on services than...

Lex Fridman
Always.

Dylan Patel
I mean, yeah. To be clear, today, people are spending a lot more on hardware than they are with the services because the hardware front runs the service spend. But-

Lex Fridman
You’re investing, yeah.

Dylan Patel
… if there’s no revenue for AI stuff or not enough revenue, then obviously, it’s going to blow up. People won’t continue to spend on GPUs forever. And NVIDIA is trying to move up the stack with software that they’re trying to sell and licensed and stuff. But Google has never had that DNA of like, “This is a product we should sell.” The Google Cloud, which is a separate organization from the TPU team, which is a separate organization from the DeepMind team, which is a separate organization from the Search team. There’s a lot of bureaucracy here.

Lex Fridman
Wait, Google Cloud is a separate team than the TPU team?

Dylan Patel
Technically, TPU sits under infrastructure, which sits under Google Cloud. But Google Cloud, for renting stuff -

Dylan Patel
But Google cloud for renting stuff and TPU architecture are very different goals, and hardware and software, all of this, right? The Jax XLA teams do not serve Google’s customers externally. Whereas NVIDIA’s various CUDA teams for things like NCCL serve external customers. The internal teams like Jax and XLA and stuff, they more so serve DeepMind and Search, right? And so their customer is different. They’re not building a product for them.

Lex Fridman
Do you understand why AWS keeps winning versus Azure for cloud versus Google Cloud?

Dylan Patel
Yeah, there’s -

Lex Fridman
Google Cloud is tiny, isn’t it, relative to AWS?

Dylan Patel
Google Cloud is third. Yeah. Microsoft is the second biggest, but Amazon is the biggest, right?

Lex Fridman
Yeah.

Dylan Patel
And Microsoft deceptively sort of includes Microsoft Office 365 and things like that, some of these enterprise-wide licenses. So in reality, the gulf is even larger. Microsoft is still second though, right? Amazon is way bigger. Why? Because using AWS is better and easier. And in many cases, it's cheaper -

Nathan Lambert
It was first.

Dylan Patel
And it's first. It was first.

Lex Fridman
Yeah. But there's a lot of things that are first that lose the -

Nathan Lambert
Well, it's harder to switch than it is to -

Lex Fridman
Yeah, okay.

Dylan Patel
AWS is -

Lex Fridman
Because there's large -

Nathan Lambert
There's big fees for switching too.

Dylan Patel
AWS generates over 80% of Amazon's profit. I think over 90%.

Lex Fridman
That's insane.

Dylan Patel
The distribution centers are just like one day we'll decide to make money from this, but they haven't yet, right? They make tiny little profit from it.

Nathan Lambert
Yeah, one day Amazon Prime will triple in price.

Lex Fridman
You would think they would improve AWS interface because it's horrible. It's clunky, but everybody is.

Nathan Lambert
Yeah, one would think.

Dylan Patel
I think actually Google's interface is sometimes nice, but it's also they don't care about anyone besides their top customers.

Lex Fridman
Exactly.

Dylan Patel
And their customer service sucks and they have a lot less -

Lex Fridman
I mean, all these companies, they optimize for the big customers. Yeah, it's supposed to be for business.

Dylan Patel
Amazon has always optimized for the small customer too though. Obviously they optimize a lot for the big customer, but when they started, they just would go to random Bay Area things and give out credits or just put in your credit card and use us back in the early days. The business has grown with them and [inaudible 04:07:04]. Why is Snowflake all over Amazon? Because Snowflake in the beginning, when Amazon didn't care about them, was still using Amazon. And then of course one day Snowflake and Amazon has a super huge partnership, but this is the case. Amazon's user experience and quality is better. Also, a lot of the silicon they've engineered makes them have a lower cost structure in traditional cloud, storage, CPU networking, that kind of stuff than in databases. I think four of Amazon's top five revenue products, margin products like gross profit products are all database-related products like Redshift and all these things. So Amazon has a very good silicon to user experience like entire pipeline with AWS. I think Google, their silicon teams, they have awesome silicon internally, TPU, the YouTube chip, some of these other chips that they've made. And the problem is they're not serving external customers, they're serving internal customers, right?

Nathan Lambert
I mean, NVIDIA's entire culture is designed from the bottom up to do this. There's this recent book, The NVIDIA Way by Tae Kim, that details this and how they look for future opportunities and ready their CUDA software libraries to make it so that new applications of

high-performance computing can very rapidly be evolved on CUDA and NVIDIA chips. And that is entirely different than Google as a services business.

Lex Fridman
I mean NVIDIA, it should be said, is a truly special company. I mean there's the culture of everything. They're really optimized for that kind of thing. Speaking of which, is there somebody that can even challenge NVIDIA hardware-wise? Intel? AMD?

Dylan Patel
I really don't think so. We went through a very long process of working with AMD on training on their GPUs inference and stuff. And they're decent, their hardware is better in many ways than in NVIDIA's. The problem is their software is really bad and I think they're getting better, right? They're getting better, faster, but the gulf is so large and they don't spend enough resources on it or haven't historically, right? Maybe they're changing their tune now, but for multiple months we were submitting the most bugs like us semi-analysis like what the fuck? Why are we submitting the most bugs? Because they only cared about their biggest customers and so they'd ship them a private image, blah, blah, blah. And it's like, "Okay, but I am just using PyTorch and I want to use the publicly available libraries," and you don't care about that. So they're getting better, but I think AMD is not possible. Intel is obviously in dire straits right now and needs to be saved somehow. Very important for national security, for American technology comments.

Lex Fridman
Can you explain the obviously, so why are they in dire straits?

Dylan Patel
Going back to earlier, only three companies can R&D, right? Taiwan Hsinchu, Samsung [inaudible 04:09:49], and then Intel Hillsboro. Samsung's doing horribly. Intel's doing horribly. We could be in a world where there's only one company that can do R& and that one company already manufactures most of chips. They've been gaining market share anyways, but that's a critical thing. So what happens to Taiwan means the rest of the world, semiconductor industry and therefore tech relies on Taiwan and that's obviously precarious as far as Intel, they've been slowly, steadily declining. They were on top of servers and PCs, but now Apple's done the M1 and Nvidia's releasing a PC chip and Qualcomm's releasing a PC chip. And in servers, hyperscalers are all making their own ARM-based server chips and Intel has no AI silicon like wins. They have very small wins and they never got into mobile because they said no to the iPhone and all these things have compounded and they've lost their process technology leadership. They were ahead for 20 years and now they're behind by at least a couple years and they're trying to catch back up and we'll see if their 18A, 14A strategy works out where they try and leapfrog TSMC like and Intel is just losing tons of money anyways, and they just fired their CEO, even though the CEO was the only person

who understood the company well, right? We'll see. He was not the best, but he was pretty good relatively technical guy.

Lex Fridman
Where does Intel make most of its money? The CPUs though.

Dylan Patel
PCs and data center CPUs, yeah, but data center CPUs are all going cloud and Amazon, Microsoft, Google are making ARM-based CPUs. And then PC side, AMD's gained market share, Nvidia's launching a chip, that's not going to be a success, right? MediaTek, Qualcomm ever launched chips. Apple's doing well. They could get squeezed a little bit in PC, although PC generally I imagine will just stick Intel mostly for Windows side.

Lex Fridman
Let's talk about the broad AI race. Who do you think wins? We talked about Google, Meta.

Nathan Lambert
The default leader has been Google because of their infrastructure advantage.

Lex Fridman
Well, in the news, OpenAI is the leader.

Nathan Lambert
They're the leading in the narrative.

Dylan Patel
They have the best model.

Nathan Lambert
They have the best model that people can use and they're experts - experts.

Dylan Patel
And they have the most AI revenue.

Nathan Lambert
Yeah. OpenAI is winning.

Lex Fridman
So who's making money on AI right now? Is anyone making money?

Dylan Patel
So accounting profit-wise, Microsoft is making money, but they're spending a lot of CapEx and that gets depreciated over years. Meta's making tons of money with recommendation systems, which is AI, but not with Llama, right? Llama's losing money for sure. I think Anthropic and OpenAI are obviously not making money otherwise they wouldn't be raising money. They have to raise money to build more. Although theoretically they are making money. You spent a few hundred million dollars on GPT-4 and it's doing billions in revenue. So obviously it's making money. Although they had to continue to research to get the compute efficiency wins and moved down the curve to get that 1200x that has been achieved for GPT-3. Maybe we're only at a couple hundred X now, but know with GPT-4 Turbo and 4.0 And there'll be another one probably cheaper than GPT-4.0 even that comes out at some point.

Lex Fridman
And that research costs a lot of money.

Dylan Patel
Yep, exactly.

Lex Fridman
That's the thing that I guess is not talked about with the cost, that when you're referring to the cost of the model, it's not just the training or the test runs, it's the actual research, the manpower.

Dylan Patel
Yeah, to do things like reasoning right now that exists. They're going to scale it. They're going to do a lot of research still. I think people focus on the payback question, but it's really easy to just be like, well, GDP is humans and industrial capital. And if you can make intelligence cheap, then you can grow a lot, right? That's the sort of dumb way to explain it. But that's sort of what basically the investment thesis is. I think only Nvidia is actually making tons of money and other hardware vendors, the hyperscalers are all on paper making money, but in reality they're spending a lot more on purchasing the GPUs, which you don't know if they're still going to make this much money on each GPU in two years, right? You don't know if all of a sudden OpenAI goes kapoof and now Microsoft has hundreds of thousands of GPUs they were renting to OpenAI that they paid for themselves with their investment in them that no longer have a customer. This is always a possibility. I don't believe that. I think OpenAI will keep raising money. I think others will keep raising money because the returns from it are going to be eventually huge once we have AGI.

Lex Fridman
So do you think multiple companies will get, let's assume -

Dylan Patel
I don't think it's winner take all.

Lex Fridman
Okay, so let's not call it AGI whatever. It's like a single day. It's a gradual thing -

Nathan Lambert
Powerful AI. Super powerful AI.

Lex Fridman
But it's a gradually increasing set of features that are useful and make -

Nathan Lambert
Rapidly increasing set of features.

Lex Fridman
Rapidly increasing set of features. So you're saying a lot of companies will be... It just seems absurd that all of these companies are building gigantic data centers.

Nathan Lambert
There are companies that will benefit from AI but not because they train the best model. Meta has so many avenues to benefit from AI and all of their services. People are there. People spend time on that as platforms, and it's a way to make more money per user per hour.

Lex Fridman
It seems like Google / X / xAI / Tesla important to say. And then Meta will benefit not directly from the AI like the LLMs, but from the intelligence, like the additional boost of intelligence to the products they already sell. So whether that's the recommendation system or for Elon who's been talking about Optimus, the robot, potentially the intelligence of the robot, and then you have personalized robots in the home, that kind of thing. He thinks it's a 10 plus trillion dollars business, which...

Nathan Lambert
At some point, maybe. Not soon, but who knows when robotics will use for -

Dylan Patel
Let's do a TAM analysis, 8 billion humans and let's get 8 billion robots and let's pay them the average salary. And there we go. 10 trillion. More than 10 trillions.

Lex Fridman
Yeah, I mean if there's robots everywhere, why does it have to be just 8 billion robots?

Dylan Patel
Yeah, yeah, of course. Of course. I'm going to have one robot. You're going to have like 20.

Lex Fridman
Yeah, I mean I see a use case for that. So yeah, so I guess the benefit would be in the products they sell, which is why OpenAI's in a trickier position because they-

Nathan Lambert
All of the value of OpenAI right now as a brand is in ChatGPT and for most users, there's not that much of a reason that they need OpenAI to be spending billions and billions of dollars on the next best model when they could just license Llama 5 and for be way cheaper. So that's kind of like ChatGPT is an extremely valuable entity to them, but they could make more money just off that.

Dylan Patel
The chat application clearly does not have tons of room to continue. The standard chat where you're just using it for a random question and stuff. The cost continues to collapse. V3 is the latest one.

Nathan Lambert
It'll go down with the ads.

Dylan Patel
But it's going to get supported by ads. Meta already serves 405B and probably loses the money, but at some point the models are going to get so cheap that they can just serve them for free with ad supported and that's what Google is going to be able to do. And obviously they've got a bigger reach. Chat is not going to be the only use case. It's like these reasoning, code, agents, computer use, all this stuff is where OpenAI has to actually go to make money in the future otherwise they're kaputs.

Lex Fridman
But X, Google, and Meta have these other products. So isn't it likely that OpenAI and Anthropic disappear eventually?

Dylan Patel
Unless they're so good at models, which they are.

Lex Fridman
But it's such a cutting edge. I mean -

Nathan Lambert
It depends on where you think AI capabilities are going.

Lex Fridman
You have to keep winning.

Dylan Patel
Yes.

Lex Fridman
You have to keep winning as you climb, even if the AI capabilities are going super rapidly awesome into the direction of AGI, there's still a boost for X in terms of data, Google in terms of data, Meta in terms of data, in terms of other products and the money and there's just huge amounts of money.

Dylan Patel
The whole idea is human data is kind of tapped out. We don't care. We all care about self-play, verifiable task.

Nathan Lambert
Think about AWS.

Lex Fridman
Yes, self-play, which is an RNG problem.

Nathan Lambert
AWS does not make a lot of money on each individual machine. And the same can be said for the most powerful AI platform, which is even though the calls to the API are so cheap, there's still a lot of money to be made by owning that platform. And there's a lot of discussions as it's the next compute layer.

Dylan Patel
You have to believe that. And there's a lot of discussions that tokens and tokenomics and LLM, APIs are the next compute layer, are the next paradigm for the economy like energy and oil was. But you have to sort of believe that APIs and chat are not where AI is stuck. It is actually just tasks and agents and robotics and computer use, and those are the areas where all the value will be delivered, not API, not chat application.

Lex Fridman
So is it possible you have it all just becomes a commodity and you have the very thin wrapper like Perplexity, just joking.

Nathan Lambert
There are a lot of wrappers making a lot of money.

Lex Fridman
But do you think it's possible that people would just even forget what OpenAI and Anthropic is just there'll be wrappers around the API and it just dynamically-

Dylan Patel
If model progress is not rapid, yeah. It's becoming a commodity, right? DeepSeek V3 shows this, but also the GPT-3 chart earlier, Kurt [inaudible 04:19:14] showed this, right? Llama 3B is 1200x cheaper than GPT-3. Anyone whose business model was GPT-3 level capabilities is dead. Anyone whose business models GPT-4 level capabilities is dead.

Nathan Lambert
It is a common saying that the best businesses being made now are ones that are predicated on models getting better.

Lex Fridman
Right. Which would be like wrappers, thing that is riding the wave of the models.

Nathan Lambert
The short-term that company that could make the most money is the one that figures out what advertising targeting method works for language model generations. We have the Meta ads which are hyper-targeted in feed, not within specific pieces of content. And we have search ads that are used by Google and Amazon has been rising a lot on search. But within a return from ChatGPT, it is not clear how you get a high-quality placed ad within the output. And if you can do that with model costs coming down, you can just get super high revenue. That revenue is totally untapped and it's not clear technically how it's done.

Lex Fridman
Yeah, that is, I mean sort of the AdSense innovation that Google did, the one day you'll have in GPT output an ad and that's going to make billions, if not-

Nathan Lambert
And it could be very subtle, it could be in conversation, we have voice mode now. It could be some way of making it so the voice introduces certain things. It's much harder to measure and it takes imagination, but yeah.

Lex Fridman
And it wouldn't come off shady so that you would receive public blowback, that kind of thing. So you have to do it loud enough to where it's clear it's an ad and balance all of that. So that's the open question they're trying to solve. Anthropic and OpenAI, they need to -

Nathan Lambert
They might not say that they're trying -

Dylan Patel
I don't think they care about that at all.

Nathan Lambert
They don't care about it right now. I think it's places like Perplexity are experimenting on that more.

Lex Fridman
Oh, interesting. Yeah, for sure.

Dylan Patel
Perplexity, Google, Meta care about this. I think OpenAI and Anthropic are purely laser focused on-

Lex Fridman
AGI.

Dylan Patel
Yeah. Like agents and AGI, and if I build AGI, I can make tons of money or I can pay for everything. And it's just predicated back on the export control thing. If you think AGI is five, 10 years away or less, these labs think it's two, three years away. Obviously your actions are, if you assume they're rational actors, which they are mostly what you do in a two-year AGI versus five year versus 10 years, very, very, very different. Right?

Lex Fridman
Do you think agents are promising? We have to talk about this. This is the excitement of the year that agents are going to rev.. This is the generic hype term that a lot of business folks are using. AI agents are going to revolutionize everything.

Nathan Lambert
Okay. So mostly the term agent is obviously overblown. We've talked a lot about reinforcement learning as a way to train for verifiable outcomes. Agents should mean something that is open-ended and is solving a task independently on its own and able to adapt to uncertainty. There's a lot of the term agent applied to things like Apple Intelligence, which we still don't have after the last WWDC, which is orchestrating between apps and that type of tool use thing is something that language models can do really well. Apple Intelligence I suspect will come eventually. It's a closed domain. It's your messages app integrating with your photos with AI in the background. That will work. That has been described as an agent by a lot of software companies to get into the narrative. The question is what ways can we get language models to generalize to new domains and solve their own problems in real time. Maybe some tiny amount of training when they're doing this with fine-tuning themselves or in context learning, which is the idea of storing information in a

prompt. And you can use learning algorithms to update that and whether or not you believe that that is going to actually generalize to things like me saying, "Book my trip to go to Austin in two days. I have XYZ constraints," and actually trusting it. I think there's an HCI problem coming back for information.

Lex Fridman
Well, what's your prediction there? Because my gut says we're very far away from that.

Dylan Patel
I think OpenAI's statement, I don't know if you've seen the five levels where it's chat is level one, reasoning is level two, and then agents is level three. And I think there's a couple more levels, but it's important to note, we were in chat for a couple years. We just theoretically got to reasoning, we'll be here for a year or two, and then agents, but at the same time, people can try and approximate capabilities of the next level, but the agents are doing things autonomously, doing things for minutes at a time, hours at a time, etc, right? Reasoning is doing things for tens of seconds at a time and then coming back with an output that I still need to verify and use and try check out. And the biggest problem is of course, it's the same thing with manufacturing. There's the whole six sigma thing, how many nines do you get? And then you compound the nines onto each other and it's like if you multiply by the number of steps that are six sigma, you get to a yield or something. So in semiconductor manufacturing, tens of thousands of steps, 9999999 is not enough. You multiply by that many times you actually end up with 60% yield, right? Really low yield or zero. And this is the same thing with agents, right? Chaining tasks together each time, even the best LLMs in particularly pretty good benchmarks don't get 100%, right? They get a little bit below that because there is a lot of noise. And so how do you get to enough nines, right? This is the same thing with self-driving. We can't have self-driving because without it being super geofenced like Google's and even then they have a bunch of teleoperators to make sure it doesn't get stuck. But you can't do that because it doesn't have enough nines.

Lex Fridman
Self-driving has quite a lot of structure because roads have rules, it's well-defined, there's regulation. When you're talking about computer use for the open web, for example, or the open operating system, it's a mess. So the possibility… I'm always skeptical of any system that is tasked with interacting with the human world, with the open messaging world.

Nathan Lambert
That's the thing. If we can't get intelligence that's enough to solve the human world on its own, we can create infrastructure like the human operators for Waymo over many years that enable certain workflows.

Dylan Patel
There is a company, I don't remember it, but it is, but that's literally their pitch is, "Yeah, we're just going to be the human operator when agents fail and you just call us and we fix it." Same thing an API call, and it's hilarious.

Nathan Lambert
There's going to be teleoperation markets when we get human robots, which is there's going to be somebody around the world that's happy to fix the fact that it can't finish loading my dishwasher when I'm unhappy with it. But that's just going to be part of the Tesla service package.

Lex Fridman
I'm just imagining an AI agent talking to another AI agent. One company has an AI agent that specializes in helping other AI agents.

Nathan Lambert
But if you can make things that are good at one step, you can stack them together. So that's why if it takes a long time, we're going to build infrastructure that enables it. You see the operator launch, they have partnerships with certain websites, with DoorDash, with OpenTable, with things like this. Those partnerships are going to let them climb really fast. Their model's going to get really good at those things. It's going to proof of concept that might be a network effect where more companies want to make it easier for AI. Some companies will be like, "No, let's put blockers in place." And this is the story of the internet we've seen, we see it now with training data for language models where companies are like, "No, you have to pay." Business working it out.

Lex Fridman
That said, I think airlines and hotels have high incentive to make their site work really well, and they usually don't. If you look at how many clicks it takes to order airplane ticket, it's insane.

Nathan Lambert
You actually can't call an American Airlines agent anymore. They don't have a phone number.

Lex Fridman
I mean, it's horrible on the interface front. And to imagine that agents will be able to deal with that website when I, as a human, struggle, like I have an existential crisis every time I try to book an airplane ticket. I think it's going to be extremely difficult to build an AI agent that's robust in that way.

Nathan Lambert
But think about it, United has accepted the Starlink term, which is they have to provide Starlink for free and the users are going to love it. What if one airline is like, "We're going to take a year and we're going to make our website have white text that works perfectly for the AIs." Every time anyone asks about an AI flight, they buy whatever airline it is.

Dylan Patel
They're just like, "Here's an API and it's only exposed to AI agents and if anyone queries it, the price is 10% higher for any flight, but we'll let you see any of our flights and you can just book any of them. Here you go."

Nathan Lambert
And then that's it.

Dylan Patel
It's like, "Oh, and I made 10% higher price. Awesome." And am I willing to say that for like, "Hey, book me a flight to [inaudible 04:28:18]." Right? And it's like, yeah, whatever. I think computers and real world and the open world are really, really messy, but if you start defining the problem in narrow regions, people are going to be able to create very, very productive things and ratchet down cost massively, right? Now, crazy things like robotics in the home, those are going to be a lot harder to do just like self-driving because there's just a billion different failure modes, but agents that can navigate a certain set of websites and do certain sets of tasks or take a photo of your fridge or upload your recipes and then it figures out what to order from Amazon/Whole Foods food delivery, and that's going to be pretty quick and easy to do, I think. So it's going to be a whole range of business outcomes and it's going to be tons of optimism around people can just figure out ways to make money.

Nathan Lambert
To be clear, these sandboxes already exist in research. There are people who have built clones of all the most popular websites of Google, Amazon, blah, blah, blah, to make it so that there's… And I mean OpenAI probably has them internally to train these things. It's the same as DeepMind's robotics team for years has had clusters for robotics where you interact with robots fully, remotely. They just have a lab in London and you send tasks to it, arrange the blocks, and you do this research. Obviously there's techs there that fix stuff, but we've turned these cranks of automation before. You go from sandbox to progress and then you add one more domain at a time and generalize, I think. And the history of NLP and language processing instruction, tuning and tasks per language model used to be like one language model did one task, and then in the instruction tuning literature, there's this point where you start adding more and more tasks together where it just starts to generalize to every task. And we don't know where on this curve we are. I think for reasoning with this RL and verifiable domains, we're early, but we don't know where the point is where you just

start training on enough domains and poof, more domains just start working. And you've crossed the generalization barrier.

Lex Fridman
Well, what do you think about the programming context? So software engineering, that's where I personally, and I know a lot of people interact with AI the most.

Dylan Patel
There's a lot of fear and angst too from current CS students, but that is the area where probably the most AI revenue and productivity gains have come, right? Whether it be Copilots or Cursor or what have you, or just standard ChatGPT. I know very few programmers who don't have ChatGPT and actually many of them have the $200 tier because that's what it's so good for. I think that in that world, we already see it like SWE-bench. And if you've looked at the benchmark made by some Stanford students, I wouldn't say it's really hard, but I wouldn't say it's easy either. I think it takes someone who's been through at least a few years of CS or a couple years of programming to do SWE-bench, well, and the models went from 4% to 60% in a year, and where are they going to go to next year? It's going to be higher. It probably won't be a hundred percent because again, that nines is really hard to do, but we're going to get to some point where that's, and then we're going to need harder software engineering benchmarks and so on and so forth. But the way that people think of it now is it can do code completion. Easy. It can do some function generation. I have to review it. Great. But really the software engineering agents I think can be done faster sooner than any other agent because it is a verifiable domain. You can always unit test or compile, and there's many different regions of it can inspect the whole code base at once, which no engineer really can. Only the architects can really think about this stuff, the really senior guys, and they can define stuff and then the agent can execute on it. So I think software engineering costs are going to plummet like crazy. And one interesting aspect of that is when software engineering costs are really low, you get very different markets. So in the US, you have all these platform SaaS companies, Salesforce and so on and so forth. In China, no one uses platform SaaS. Everyone just builds their own stack because software engineering is much cheaper in China and partially because people, number of STEM graduates, etc. So it's generally just cheaper to do. And so at the same time, code LLMs have been adopted much less in China because the cost of an engineer there is much lower. But what happens when every company can just invent their own business logic really cheaply and quickly? You stop using platform SaaS, you start building custom tailored solutions, you change them really quickly. Now all of a sudden your business is a little bit more efficient too, potentially because you're not dealing with the hell that is. Some random platform SaaS company stuff not working perfectly and having to adjust workflows or random business automation cases that aren't necessarily AI required. It's just logic that needs to be built that no one has built. All of these things can go happen faster. And so I think software and then the other domain is industrial, chemical, mechanical engineers suck at coding just generally. And their tools like semiconductor

engineers, their tools are 20 years old. All the tools run on XP including ASML lithography tools run on Windows XP. And a lot of the analysis happens in Excel, right? It's just like, "Guys, you guys can move 20 years forward with all the data you have and gathered and do a lot better." You need the engineering skills for software engineering to be delivered to the actual domain expert engineer. So I think that's the area where I'm super-duper bullish of generally AI creating value.

Nathan Lambert
The big picture is that I don't think it's going to be a cliff. I think a really good example of how growth changes is when Meta added stories. So Snapchat was on an exponential, they added stories, it flatlined. Software engineers then up until the right, AI is going to come in, it's probably just going to be flat. It's not like everyone's going to lose their job. It's hard because the supply corrects more slowly. So the amount of students is still growing, and that'll correct on a multi-year, like a year delay, but the amount of jobs will just turn and then maybe in 20, 40 years, it'll be well down. But in the few years, there'll never going to be the snap moment where it's like software engineers aren't useful.

Lex Fridman
I think also the nature of what it means to be a programmer and what kind of jobs programmers do changes, because I think there needs to be a human in the loop of everything you've talked about. There's a really important human in that picture of correcting the code, fixing -

Dylan Patel
Thinking larger than the context length.

Lex Fridman
And debugging also, like debugging by reading the code, understanding the steering the system. No, no, no. You missed the point. Adding more to the prompt like, yes, adding the human-

Nathan Lambert
Designing the perfect Google button. Google's famous for having people design buttons that are so perfect, and it's like how is AI going to do that? They could give you all the ideas. Perfect, fine.

Lex Fridman
I mean, that's the thing. You can call it taste. One thing humans can do is figure out what other humans enjoy better than AI systems. That's where the preference you loading that in. But ultimately, humans are the greatest preference generator. That's where the preference comes from.

Nathan Lambert
And humans are actually very good at reading or judging between two things versus… This goes back to the core of what RLHF and preference tuning is that it's hard to generate a good answer for a lot of problems, but it's easy to see which one is better. And that's how we're using humans for AI now is judging which one is better, and that's what software engineering could look like. The PR review, here's a few options, here are some potential pros and cons, and they're going to be judges.

Lex Fridman
I think the thing I would very much recommend is programmers start using AI and embracing that role of the supervisor of the AI system and partner the AI system versus writing from scratch or not learning coding at all and just generating stuff because I think there actually has to be a pretty high level of expertise as a programmer to be able to manage increasingly intelligent systems.

Dylan Patel
I think it's that and then becoming a domain expert in something.

Lex Fridman
Sure. Yeah.

Dylan Patel
Because seriously, if you go look at aerospace or semiconductors or chemical engineering, everyone is using really crappy platforms, really old software. The job of a data scientist is a joke in many cases. In many cases, it's very real, but it's like bring what the forefront of human capabilities are to your domain. And even if the forefront is from the AI, your domain, you're at the forefront. So it's like you have to be at the forefront of something and then leverage the rising tide that is AI for everything else.

Lex Fridman
Oh, yeah. There's so many low hanging fruit everywhere in terms of where software can help automate a thing or digitize a thing in the legal system. That's why DOGE is exciting. I got to hang out with a bunch of the DOGE folks, and I mean, government is so old school. It's like begging for the modernization of software, of organizing the data, all this kind of stuff. I mean, in that case it's by design because bureaucracy protects centers of power and so on. But software breaks down those barriers, so it hurts those that are holding onto power, but ultimately benefits humanity. So there's a bunch of domains of that kind. One thing we didn't fully finish talking about is open-source. So first of all, congrats. You released a new model.

Nathan Lambert
Yeah, this is -

Lex Fridman
Tülu.

Nathan Lambert
I'll explain what a tülu is. A tülu is a hybrid camel when you breed a dromedary with a Bactrian camel. Back in the early days after ChatGPT, there was a big wave of models coming out like Alpaca, Vicuna, etc, that were all named after various mammalian species. Tülu, the brand, is multiple years old, which comes from that. And we've been playing at the frontiers of post-training with open-source code. And this first part of this release was in the fall where we've built on Llama's, open models, open weight models, and then we add in our fully open code or fully open data. There's a popular benchmark that is Chatbot Arena. And that's generally the metric by which how these chat models are evaluated. And it's humans compare random models from different organizations. And if you looked at the leaderboard in November or December, among the top 60 models from tens to twenties of organizations, none of them had open code or data for just post-training. Among that, even fewer or none have pre-training data and code available. Post-training is much more accessible at this time. It's still pretty cheap, and you can do it. And the thing is, how high can we push this number where people have access to all the code and data? So that's kind of the motivation of the project. We draw in lessons from Llama. Nvidia had a Nemotron model where the recipe for their post-training was fairly open with some data and a paper, and it's putting all these together to try to create a recipe that people can fine tune models like GPT-4 to their domain.

Lex Fridman
To be clear, in the case of Tülu, maybe you can talk about Llama too, but in the case of Tülu, you're taking Llama 3, 405B.

Nathan Lambert
Tülu has been a series of recipes for post-training. So we've done multiple models over years.

Lex Fridman
And so you're open-sourcing everything.

Nathan Lambert
Yeah. If you start with an open weight based model, their whole model technically isn't open-source because you don't know what Llama put into it, which is why we have the separate thing that we'll get to, but it's just getting parts of the pipeline where people can zoom in and customize. I know I hear from startups and businesses, they're like, "Okay, I can take this post-training -" I know I hear from startups and businesses, they're like, "Okay, I can take this post-training and try to apply it to my domain." We talk about verifiers a lot. We use this idea which is reinforcement learning with verifiable rewards RLVR, kind of similar

to RLHF. And we applied it to MAP and the model today, which is we applied it to the Llama 405B base model from last year. And we have our other stuff. We have our instruction tuning and our preference tuning. But the math thing is interesting, which is it's easier to improve this math benchmark. There's a benchmark, M-A-T-H, MATH, all capitals, tough name on the benchmark name is the area that you're evaluating. We're researchers, we're not brand strategists. And this is something that the DeepSeek paper talked about as well is at this bigger model, it's easier to elicit powerful capabilities with this RL training. And then they distill it down from that big model to the small model. And this model we released today, we saw the same thing. We're at AI2, we don't have a ton of compute. We can't train 405B models all the time. So we just did a few runs and they tend to work. And it just shows that there's a lot of room for people to play in these things and that's –

Dylan Patel
And they crushed Llama's actual release, they're way better than it.

Nathan Lambert
… Yeah. So our eval numbers, I mean we have extra months in this, but our eval numbers are much better than the Llama instruct model that they released.

Lex Fridman
And then you also said better than DeepSeek V3?

Nathan Lambert
Yeah, on our eval benchmark. DeepSeek V3 is really similar. We have a safety benchmark to understand if it will say harmful things and things like that. And that's what draws down most of the way. It's still -

Dylan Patel
It's like an amalgamation of multiple benchmarks or what do you mean?

Nathan Lambert
… Yeah, so we have a 10 evaluator. This is standard practice in post-training is you choose your evaluations you care about. In academics, in smaller labs you'll have fewer evaluations. In companies, you'll have a really one domain that you really care about. In Frontier Labs, you'll have tens to 20s to maybe even 100 evaluations of specific things. So we choose a representative suite of things that look like chat, precise instruction following, which is like respond only in emojis, just model follow weird things like that, math, code. And you create a suite like this. So safety would be one of 10 in that type of suite where you have what does the broader community of AI care about? And for example, in comparison to DeepSeek it would be something like our average eval for our model would be 80, including safety and similar without. And DeepSeek would be like 79% average score without safety and their safety score would bring it down to like 70 or there abouts.

Dylan Patel
Oh, so you'd beat them even ignoring safety.

Nathan Lambert
Yeah. So this is something that internally, it's like I don't want to win only by how you shape the eval benchmark. So if there's something that's like people may or may not care about safety in their model, safety can come downstream, safety can be when you host the model for an API, like safety is addressed in a spectrum of locations in AI applications. So it's like if you want to say that you have the best recipe, you can't just gate it on these things that some people might not want. And this is, it's like the time of progress and we benefit if we can release a model later, we have more time to learn new techniques like this RL technique, we had started this in the fall, it's now really popular reasoning models. The next thing to do for open-source post-training is to scale up verifiers, to scale up data to replicate some of DeepSeek's results. And it's awesome that we have a paper to draw on and it makes it a lot easier. And that's the type of things that is going on among academic and closed frontier research in AI.

Lex Fridman
Since you're pushing open-source, what do you think is the future of it? Do you think DeepSeek actually changes things since it's open-source or open weight or is pushing the open-source movement into the open direction?

Nathan Lambert
This goes very back to license discussion. So DeepSeek R1 with a friendly license is a major reset. So it's like the first time that we've had a really clear frontier model that is open weights and with a commercially friendly license with no restrictions on downstream use cases since that data distillation, whatever.This has never been the case at all in the history of AI in the last few years since ChatGPT. There have been models that are off the frontier or models with weird licenses that you can't really use them.

Dylan Patel
So is it Meta's license pretty much permissible except for five companies?

Nathan Lambert
So this goes to what open-source AI is, which is there's also use case restrictions in the Llama license, which says you can't use it for specific things. So if you come from an open-source software background, you would say that that is not an open-source license.

Dylan Patel
What kind of things are those, though? Are they like -

Nathan Lambert
At this point, I can't pull them off the top of my head, but it'd be like -

Lex Fridman
Stuff like competitors?

Nathan Lambert
It used to be military use was one and they removed that for scale, it'll be like CSAM like child abuse material. That's the type of thing that is forbidden there. But that's enough from an open-source background to say it's not an open-source license.And also the Llama license has this horrible thing where you have to name your model Llama if you touch it to the Llama model. So it's like the branding thing. So if a company uses Llama, technically the license says that they should say built with Llama at the bottom of their application. And from a marketing perspective, that just hurts. I could suck it up as a researcher, I'm like, oh, it's fine. It says Llama dash on all of our materials for this release. But this is why we need truly open models, which is we don't know DeepSeek R1's data, but-

Dylan Patel
Wait, so you're saying I can't make a cheap copy of Llama and pretend it's mine, but I can do this with the Chinese model?

Nathan Lambert
... Hell, yeah. That's what I'm saying. And that's why it's like we want this whole open language model thing, he Olmo thing is to try to keep the model where everything is open with the data as close to the frontier as possible. So we're compute constrained, we're personnel constrained. We rely on getting insights from people like John Schulman tells us to do URL and outputs. We can make these big jumps, but it just takes a long time to push the frontier of open-source. And fundamentally, I would say that that's because open-source AI does not have the same feedback loops as open-source software. We talked about open-source software for security. Also it's just because you build something once and can reuse it. If you go into a new company, there's so many benefits, but if you open-source a language model, you have this data sitting around, you have this training code, it's not like that easy for someone to come and build on and improve because you need to spend a lot on compute, you need to have expertise. So until there are feedback loops of open-source AI, it seems like mostly an ideological mission. People like Mark Zuckerberg, which is like America needs this and I agree with him, but in the time where the motivation ideologically is high, we need to capitalize and build this ecosystem around, what benefits do you get from seeing the language model data? And there's not a lot about that. We're going to try to launch a demo soon where you can look at an OMO model and a query and see what pre-training data is similar to it, which is legally risky and complicated, but it's like what does it mean to see the data that the AI was trained on? It's hard to parse.

It's terabytes of files. It's like I don't know what I'm going to find in there, but that's what we need to do as an ecosystem if people want open-source AI to be financially useful.

Lex Fridman
We didn't really talk about Stargate. I would love to get your opinion on what the new administration, the Trump administration, everything that's being done from the America side and supporting AI infrastructure and the efforts of the different AI companies. What do you think about Stargate? What are we supposed to think about Stargate and does Sam have the money?

Dylan Patel
Yeah, so I think Stargate is a opaque thing. It definitely doesn't have $500 billion, doesn't even have $100 billion dollars. So what they announced is this $500 billion number, Larry Ellison, Sam Altman and Trump said it. They thanked Trump and Trump did do some executive actions that do significantly improve the ability for this to be built faster. One of the executive actions he did is on federal land, you can just basically build data centers in power pretty much like that. And then permitting process is basically gone or you file after the fact. So again, I had of schizo take earlier, another schizo take, if you've ever been to the Presidio in San Francisco, beautiful area, you could build a power plant in a data center there if you wanted to because it is federal land. It used to be a military base, but obviously this would people off. It's a good fit. Anyways, Trump has made it much easier to do this, right? Generally, Texas has the only unregulated grid in the nation as well.

Lex Fridman
Let's go Texas.

Dylan Patel
And so therefore ERCOT enables people to build faster as well in addition, the federal regulations are coming down and so Stargate is predicated, and this is why that whole show happened. Now how they came up with a $500 billion number is beyond me. How they came up with $100 billion dollars number makes sense to some extent. And there's actually a good table in here that I would like to show in that Stargate piece that I had. It's the most recent one. So anyways, Stargate, it's basically, it's a table about cost. There, you passed it already. It's that one. So this table is kind of explaining what happens. So Stargate is in Abilene, Texas, the first $100 billion of it. That site is 2.2 gigawatts of power in, about 1.8 gigawatts of power consumed. Per GPU, Oracle is already building the first part of this before Stargate came about. To be clear, they've been building it for a year. They tried to rent it to Elon in fact, but Elon was like, "It's too slow. I need it faster." So then he went and did his Memphis thing, and so OpenAI was able to get it with this weird joint venture called Stargate. They initially signed a deal with just Oracle for the first section of this cluster. This first section of this cluster is roughly $5 billion to $6 billion of server spend, and then there's another billion or so of data center spend. And then likewise, if you fill out that entire 1.8 gigawatts with the

next two generations of NVIDIA's chips, GB 200, GB 300, VR 200, and you fill it out completely, that ends up being roughly $50 billion of server cost. Plus there's data center costs plus maintenance costs, plus operation costs plus all these things. And that's where OpenAI gets to their $100 billion announcement that they had. Because they talked about $100 billion dollars is phase one. That's this Abilene, Texas data center, right? $ 100 billion of "total cost of ownership." So it's not CapEx, it's not investment, it's a $100 billion of total cost of ownership. And then there will be future phases. They're looking at other sites that are even bigger than this 2.2 gigawatts by the way, in Texas and elsewhere. And so they're not completely ignoring that, but the number of $100 billion that they save for phase one, which I do think will happen. They don't even have the money for that. Furthermore, it's not $100 billion dollars, it's $50 billion of spend and then $50 billion of operational cost power, etc, rental pricing, etc, because they're renting it. OpenAI is renting the GPUs from the Stargate joint venture. What money do they actually have, right? SoftBank is going to invest, Oracle is going to invest. OpenAI is going to invest. OpenAI is on the line for $19 billion. Everyone knows that they've only got 46 billion in their last round and $4 billion of debt. But there, there's news of Softbank maybe investing $25 billion into OpenAI. So that's part of it. So $19 billion can come from there. So OpenAI does not have the money at all to be clear. Ink is not dried on anything. OpenAI has $0 for this, 50 billion in which they're legally obligated to put 19 billion of CapEx into the joint venture, and then the rest they're going to pay via renting the GPUs from the joint venture. And then there's Oracle. Oracle has a lot of money. They're building the first section completely. They were spending for it themselves, this $6 billion of CapEx, $10 billion of TCO, and they were going to do that first section. They're paying for that, right? As far as the rest of the section, I don't know how much Larry wants to spend. At any point he could pull out. This is again, it is completely voluntary. So at any point, there's no signed ink on this, but he potentially could contribute tens of billions of dollars to be clear. He's got the money, Oracle's got the money. And then there's like MGX is the UAE fund, which technically has $1.5 trillion for investing in AI. But again, I don't know how real that money is and there's no ink signed for this, SoftBank does not have $25 billion of cash. They have to sell down their stake in arm, which is the leader in CPUs and they IPO'd it. This is obviously what they've always wanted to do, they just didn't know where they'd redeploy the capital. Selling down the stake in ARM makes a ton of sense. So they can sell that down and invest in this if they want to and invest in OpenAI if they want to. As far as money secured, the first 100,000 GB 200 cluster can be funded. Everything else after that-

Lex Fridman
Up in the air.

Dylan Patel
... is up in the air. Money's coming. I believe the money will come. I personally do.

Lex Fridman
It's a belief.

Dylan Patel
It's a belief that they're going to release better models and be able to raise more money. But the actual reality is that Elon's right, the money does not exist.

Lex Fridman
What does the US government have to do with anything? What does Trump have to do with everything? He's just a hype man?

Dylan Patel
Trump, he's reducing the regulation so they can build it faster and he's allowing them to do it because any investment of this side is going to involve antitrust stuff. So obviously he's going to allow them to do it. He's going to enable the regulations to actually allow it to be built. I don't believe there's any US government dollars being spent on this though.

Lex Fridman
So I think he's also just creating a general vibe that regulation will go down and this is the era of building. So if you're a builder, you want to create stuff, you want to launch stuff, this is the time to do it.

Dylan Patel
And so we've had this 1.8 gigawatt data center in our data for over a year now, and we've been sending it to all of our clients, including many of these companies that are building the multi gigawatts. But that is at a level that's not quite, maybe executives seeing $500 billion, $100 billion dollars, and then everyone's asking them. So it could spur an even faster arms race. Because there's already an arms race, but this 100 billion, $500 billion number, Trump talking about it on TV, it could spur the arm race to be even faster and more investors to flood in and etc, etc. So I think you're right in that sense that OpenAI or Trump is sort of championing, people are going to build more and his actions are going to let people build more.

Lex Fridman
What are you excited about these several years that are upcoming in terms of cluster build outs, in terms of breakthroughs in AI, the best possible future you can imagine in the next couple of years, two, three, four years? What does that look like? It could be very specific technical things like breakthroughs on post-training or it could be just size, big impressive clusters.

Dylan Patel
I really enjoy tracking supply chain and who's involved and what, I really do. It's really fun to see the numbers, the cost, who's building what capacity, helping them figure out how much capacity they should build winning deals, strategic stuff. That's really cool. I think technologically, there's a lot around the networking side that really excites me with optics

and electronics kind of getting closer and closer, whether it be co-packaged optics or some sort of forms of new forms of switching.

Lex Fridman
This is internal to a cluster?

Dylan Patel
A cluster, yeah. Also multi-data center training. People are putting so much fiber between these data centers and lighting it up with so much bandwidth that there's a lot of interesting stuff happening on that end. Telecom has been really boring since 5G, and now it's really exciting again on the hardware side.

Lex Fridman
Can you educate me a little bit about the speed of things? So the speed of memory versus the speed of interconnect versus the speed of fiber between data centers. Are these orders of magnitude different? Can we at some point converge towards a place where it all just feels like one computer?

Dylan Patel
No, I don't think that's possible. It's only going to get harder to program, not easier. It's only going to get more difficult and complicated and more layers. The general image that people like to have is this hierarchy of memory, so on-chip is really close, localized within the chip, you have registers and those are shared between some compute elements and then you'll have caches which are shared between more compute elements. Then you have memory like HBM or DRAM like DDRR memory or whatever it is, and that's shared between the whole chip. And then you can have pools of memory that are shared between many chips and then storage and you keep zoning out. The access latency across data centers, within the data center within a chip is different. So you're always going to have different programming paradigms for this. It's not going to be easy. Programming this stuff is going to be hard, maybe AI can help with programming this. But the way to think about it is that there is sort of the more elements you add to a task, you don't get strong scaling. If I double the number of chips, I don't get two exit performance. This is just a reality of computing because there's inefficiencies.And there's a lot of interesting work being done to make it not to make it more linear, whether it's making the chips more networked together more tightly or cool programming models or cool algorithmic things that you can do on the model side. DeepSeek did some of these really cool innovations because they were limited on interconnect, but they still needed to parallelize. Everyone's always doing stuff. Google's got a bunch of work and everyone's got a bunch of work about this. That stuff is super exciting on the model and workload and innovation side. Hardware, solid-state transformers are interesting. For the power side, all sorts of stuff on batteries and there's all sorts of stuff on. I think if you look at every layer of the compute stack, whether it goes from lithography and etch all the way to fabrication, to optics, to networking, to power, to transformers, to

cooling, to a networking, and you just go on up and up and up and up the stack, even air conditioners for data centers are innovating. Copper cables are innovating. You wouldn't think it, but copper cables, there's some innovations happening there with the density of how you can pack them and it's like all of these layers of the stack, all the way up to the models, human progress is at a pace that's never been seen before.

Lex Fridman
I'm just imagining you sitting back in a layer somewhere with screens everywhere, just monitoring the supply chain where all these clusters, all the information you're gathering, you're incredible.

Dylan Patel
There's a big team, there's a big team.

Lex Fridman
You do quite incredible work with semi analysis. I mean just keeping your finger on the pulse of human civilization in the digital world. It's pretty cool just to watch, feel that.

Dylan Patel
Yeah, thank you. I guess.

Lex Fridman
Feel all of us doing shit. Epic shit.

Dylan Patel
The AGI, yeah.

Lex Fridman
I feel from meme to reality. Nathan, is there breakthroughs that you're looking forward to potentially?

Nathan Lambert
I had a while to think about this while listening to Dylan's beautiful response.

Dylan Patel
He did listen to me. He was so into it.

Nathan Lambert
No, I knew this was coming and it's like realistically training models is very fun because there's so much low-hanging fruit. And the thing that makes my job entertaining, I train models, I write analysis about what's happening with models and it's fun because there is obviously so much more progress to be had. And the real motivation, why I do this

somewhere where I can share things is that there's just, I don't trust people that are like, "Trust me bro, we're going to make AI good." It's like we're the ones that it's like, we're going to do it and you can trust us and we're just going to have all the AI, and it's just like, I would like a future where more people have a say in what AI is and can understand it, and it's a little bit less fun that it's not a positive thing of this is just all really fun. Training models is fun and bring people in as fun, but it's really AI if it is going to be the most powerful technology of my lifetime, it's like we need to have a lot of people involved in making that and –

Lex Fridman
Making it open helps with that. As accessible as possible, as open as possible, yeah.

Nathan Lambert
… In my read of the last few years is that more openness would help the AI ecosystem in terms of having more people understand what's going on. Rather that's researchers from non-AI fields to governments to everything. It doesn't mean that openness will always be the answer. I think then it'll reassess of what is the biggest problem facing AI and tack on a different angle to the wild ride that we're on.

Lex Fridman
And for me, just from even the user experience, anytime you have like Aparthi said, the aha moments, the magic, seeing the reasoning, the chain of thought, it's like there's something really just fundamentally beautiful about that. It's putting a mirror to ourselves and seeing like, oh, shit. It is solving intelligence as the cliche goal of these companies is, and you get to understand why we humans are special. The intelligence within us is special. And for now also why we're special in terms of we seem to be conscious and the AI systems for now, and we get to explore that mystery, so it's just really cool to get to explore these questions that I don't think I would've never imagined would be even possible back when just watching with excitement, deep blue beat Kasparov, I wouldn't have ever thought this kind of AI would be possible in my lifetime. This is really feels like AI.

Nathan Lambert
Yeah.

Lex Fridman
It's incredible.

Nathan Lambert
I started with AI learning to fly a silly, a quadrotor, it's like learning to fly and it learned to fly up. It would hit the ceiling and stop and catch it. It's like, okay, that is really stupid compared to what's going on now.

Lex Fridman
And now you could probably with natural language tell it to learn to fly and it's going to generate the control algorithm required to do that probably.

Nathan Lambert
There's low level blockers. We have to do some weird stuff for that, but you can, you definitely can.

Lex Fridman
Back to our robotics conversation, yeah, when you have to interact in the actual physical world, that's hard. What gives you hope about the future of human civilization looking into the next 10 years, 100 years, 1000 years, how long do you think we'll make it? You think we've got 1000 years?

Nathan Lambert
I think humans will definitely be around in a 1000 years, I think. There's ways that very bad things could happen. There'll be way fewer humans, but humans are very good at surviving. There's been a lot of things that that is true. I don't think necessarily we're good at long-term credit assignment of risk, but when the risk becomes immediate, we tend to figure things out.

Lex Fridman
Oh, yeah.

Nathan Lambert
And for that reason, there's physical constraints to things like AGI, like recursive improvement to kill us all type stuff. For the physical reasons and for how humans have figured things out before, I'm not too worried about AI takeover. There are other international things that are worrying, but there's just fundamental human goodness and trying to amplify that. I think we're on a tenuous time. And I mean if you look at humanity as a whole, there's been times where things go backwards, there's times when things don't happen at all, and we're on what should be very positive trajectory right now.

Lex Fridman
Yeah, there seems to be progress, but just like with power, there's like spikes of human suffering and we want to try to minimize the amount of spikes.

Dylan Patel
Generally, humanity is going to suffer a lot less, I'm very optimistic about that. I do worry of like techno-fascism type stuff arising. As AI becomes more and more prevalent and powerful and those who control it can do more and more, maybe it doesn't kill us all, but at some point, every very powerful human is going to want to brain- computer interface so

that they can interact with the AGI and all of its advantages in many more way and merge its mind and its capabilities or that person's capabilities can leverage those much better than anyone else and therefore be, it won't be one person rule them all, but it will be, the thing I worry about is it'll be few people, hundreds, thousands, tens of thousands, maybe millions of people rule whoever's left and the economy around it. And I think that's the thing that's probably more worrisome is human-machine amalgamations. This enables an individual human to have more impact on the world and that impact can be both positive and negative. Generally, humans have positive impacts on the world, at least societally, but it's possible for individual humans to have such negative impacts. And AGI, at least as I think the labs define it, which is not a runaway sentient thing, but rather just something that can do a lot of tasks really efficiently amplifies the capabilities of someone causing extreme damage. But for the most part, I think it'll be used for profit-seeking motives, which will increase the abundance and supply of things and therefore reduce suffering, right? That's the goal.

Lex Fridman
Scrolling on a timeline, just drowning in dopamine -

Dylan Patel
Scrolling open stasis.

Nathan Lambert
Scrolling holds the status quo of the world.

Dylan Patel
That is a positive outcome, right? If I have food tubes and lung down scrolling and I'm happy, that's a positive outcome.

Lex Fridman
While expanding out into the cosmos. Well, this is a fun time to be alive. And thank you for pushing the forefront of what is possible in humans, and thank you for talking today. This was fun.

Dylan Patel
Thanks for having us.

Nathan Lambert
Thanks for having us.

Lex Fridman
Thanks for listening to this conversation with Dylan Patel and Nathan Lambert. To support this podcast, please check out our sponsors in the description. And now let me leave you with some words from Richard Feynman. "For a successful technology, reality must take

precedence over public relations, for nature cannot be fooled." Thank you for listening and I hope to see you next time.