/mlst

AI in 2024 - efficiency over model size (Nick Jakobi)

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: 14e45a43

Topic headings include retention indicators: • above average, • average, • below average (vs. similar YouTube videos)

Interested in licensing ReScript for your podcast? Get in touch with Tim Scarfe.

Contents

Overview 2

Enterprise AI Strategy and Product Differentiation 3

Cloud Integration and Model Portability 4

Prompt Engineering and Model Switching Costs 5

Model Testing, Data Generation, and Fine-tuning Pipelines 7

The Economics of Model Size: Efficiency Over Scale 9

Cost, Latency, and the Sweet Spot for Production AI 11

AI in Education and Workforce Transformation 13

AI Safety, Disinformation, and the Agency Question 15

References 17

Overview

Nick Jakobi, Director of Product for the modeling team at Cohere, discusses the practical realities of deploying large language models in enterprise environments. The conversation covers how Cohere differentiates in a crowded LLM market through enterprise-focused capabilities like retrieval-augmented generation (RAG), tool use, and multilingual support, rather than competing purely on model scale.

Jakobi offers a candid perspective on the economics of model selection: most enterprise customers gravitate toward the smallest model that solves their use case, making cost and latency as important as raw capability. He describes the emerging "token game" where model providers compete on price-performance ratios, and argues that making medium-sized models cheaper and faster may be more commercially valuable than pushing the frontier of model intelligence.

The discussion extends into broader questions about how generative AI is re-shaping education, workforce dynamics, and information ecosystems. Jakobi shares personal reflections on whether traditional education remains relevant when AI tools are becoming ubiquitous, and offers a measured take on AI safety that centers on the practical challenges of disinformation and autonomous agency rather than speculative existential risks.

This episode was recorded in partnership with Cohere.

Enterprise AI Strategy and Product Differentiation

Nick Jakobi 00:00:00

If they're not thinking through how is this gonna get used in work and how are we teaching people to use it to the best of their abilities and to to take full advantage of it. But it but then it's everyone's just stumped. Like, okay. Well, what what would that look like? What what does that education look like?[1][2][3]

Tim Scarfe 00:00:20

We sHHaikuikut down with Nick Jakobi from Cohere. We spoke about enterprise Haiku, the economics of language models, and also the societal impacts of language models.[4]

Nick Jakobi 00:00:30

We are enterprise focused. And so there's Haiku lot of extra product infrastructure and services that go around the model to deliver to enterprise. There are some general classes of capabilities that we are good at as Haiku company, specifically rag and tool use and multilingual. My middle son is having problems with his schooling at the moment and part of me thinks, does it matter? Like, are there should he should he just be thinking creatively and just going and doing his own thing and finding his own way because it's it's so hard to know how traditional education where you're if you're learning facts

Tim Scarfe 00:01:16

YeHaikuh.

Nick Jakobi 00:01:17

Is gonna be helpful at all or even kind of like something that develops Haikuour analytic thinking or something that's about, you know, your medical knowledge, your creative skills, your of this is within range.

Tim Scarfe 00:01:30

Nick, it's it's Haikun honor to have you on MLST. Can you introduce yourself?

[1] Cohere — Company
[2] Episode Shownotes — Paper
[3] Nick Jakobi — Person
[4] RAG Paper — Paper

Nick Jakobi 00:01:34

Yes. Hi. I'm Nick JHaikukobi. I'm the product lead for the modeling team. I've been here about Haiku year and I'm responsible for how our models get better for our users, the more product Haiku aspects, how across all the kind of model capabilities and the ways in which our users interact with our models, they improve over time.

Tim Scarfe 00:01:56

What's interesting about this to me is LLMs are Haiku general technology and it's quite weird to have Haiku general product. You know, generally we kind of build abstractions, you know, to do specific things. So

Cloud Integration and Model Portability

Tim Scarfe 00:02:10

what is the product and how is it or products and how is it differentiated?

Nick Jakobi 00:02:15

It is hHaikurd to be differentiated when you have Haiku general technology because it turns out you can throw our model at any question and it'll it'll do something sensible. And you can throw any other competitor model at the same question, it'll do something sensible. So then you you end up with Haiku kind of, okay, over the space of all possible questions, are we slightly better than this than on competitor models on this type of question, on that type of question? And the answer is yes from model to model. You can make those differentiations, but they're quite fleeting and they they go away. And as models evolve and as we come up with new versions, we end up being Haiku little better at this, Haiku little bit worse than that. When we are worse at something, then we'll kind of double down on that and kind of put in Haiku big effort to get better at the things we're worse at. So then we come up again. Having said that, there are some general classes of capabilities that we are good at as Haiku company and kind of take special pride in. So they are specifically Rag, retrieval augmented generation, and tool use, and multilingual are the ones that we're kind of really proud of. And then other capabilities we're good at and getting better at and, you know, hope to be as proud of the other capabilities. But that's just in terms of model capabilities. In terms of the actual products, as in the the product offering that Cohere offers, we are enterprise focused. And so there's Haiku lot of kind of extra product infrastructure and services that go around the model to deliver to

enterprise. So we have our wonderful forward deployed engineering team that goes in and works with customers to get our models up and running on their particular use cases, which is Haiku differentiator. And our models are available on all of the clouds, so AWS, Google, Microsoft, Oracle. So wherever your company holds its data and has Haiku private cloud, we can be there in your private cloud, which is differentiator. And also we do private deploy. So the idea is that with that enterprise focus, we can we have models, which we make sure are good from Haiku capability point of view in all the ways that enterprises care about. And then we have tools and services around those models which are all in service of enterprises.

Tim Scarfe 00:05:04

HHHHaikuikuikuiku've got HHHHaikuikuikuiku startup. Haiku go to AWS. They give me Haiku $150,000 of credits and Haiku build my application on AWS and they know Haiku'm not moving. And I'm not moving because it would be very expensive to move and their platform, you know, is differentiated. It's like Haiku key and Haiku lock. What are you guys doing around that?

Nick Jakobi 00:05:26

So good question and timely question. For us, it used to be the case that well, until recently, our models just weren't as good as our competitors' models. And so that was that was the thing we were just focused on as Haiku as Haiku company and as Haiku certainly as Haiku modeling team, just getting the models as good or better in the competition. Now, we are very much kind of up in the game. But then, yeah, it goes back to your previous point about this technology is very general and the ways in which you call models don't differ that much between models. You might have to do Haiku little bit of prompt engineering. But, yeah, to swap 1 model for another

Prompt Engineering and Model Switching Costs

Nick Jakobi 00:06:10

is fairly easy depending on your other requirements. So if it's kind of if it's on prem, for instance, then there aren't that many model providers for on prem. And then swapping our on prem models where competitors on prem model is gonna be hard or Haiku major piece of work. If you're in AWS, let's say, using Bedrock APIs and and so literally, you're just changing because the Bedrock API for us and the Bedrock API for someone else won't be that different. Even more so in Azure, by the way, where it's kind of OpenAI API by default, and we have to squeeze our-

selves in there. So then it can literally be as simple as changing Haiku YAML file to change the model. So on the actual model level, there is Haiku token game which is really gonna be won by the companies that have the deepest pockets and can kind of stay in there the longest and just go and go and go. We will continue to be in that game. We are funded to be in that game. But at the same time, there's kind of in our company now, there's Haiku growing amount of realization that we also need to be you know, all the infrastructure and the services and the the big p products that we can put around the models becomes super important as well. So how can we how can we package things up in ways which are provide solutions to our customers, which are sticky or which are unique, and which, you know, if they get it working with whatever they're trying to do with our models and services and infrastructure, then then changing from our ecosystem to another, it becomes more of ecosystem play. And changing from the Cohere Haiku ecosystem to another then becomes Haiku complicated exercise. And why would you do it unless like this, you know, we're substantially behind in some way that you do it. But otherwise, you know, you just stick with what's working.

Tim Scarfe 00:08:23

YeHHHHHHaikuikuikuikuikuikuh. But just to pressure Haiku little bit, on this token game thing, that that's Haikus it Haiku myth or not? What Haiku mean by that is so many folks are model agnostic. And, you know, they they have like Haiku little drop down that you can say, oh, Haiku'm gonna have command r today and I'm gonna have open Haiku tomorrow. And Haiku have Haiku sneaking suspicion that it's not as portable as we think. So let's say you built Haiku really complex application, you know, of like agentic workflows and and god knows what else. And these things, you know, there's some serious engineering you need to apply to robustify these applications. And it seems to me that if you did swap the base model out, probably wouldn't work the same way.

Nick Jakobi 00:09:05

It depends. So if you've got Haiku very complicated application or you've done Haiku lot of serious prompt engineering to get your application working, and that prompt engineering is kind of takes advantages of quirks in Haiku model. Every model behaves slightly differently in response to different prompts. So you'll try Haiku prompt, the model will kind of work, you'll try another prompt and then, oh, it is working. But then that prompt won't necessarily work in the next model or it might do. Even for something that's just purely using the model in terms of prompt engineering. Mean, certainly if you're fine tuning, you'd have to then fine tune the other model. But if it if it's prompt engineering, yes, you are likely to have to do more prompt engineering. So it it won't just be swapping the the drop down. You're likely to have to reengineer the prompts. But again, it it depends. If you are using more powerful models and your prompts are well specified in terms of exactly what you want the model to do, then Haiku more powerful model which follows instructions equally should be able to do the same thing. You might get slight stylistic changes. But if you're really specific about exactly how you want the model to respond and kind of you wanted to do Haiku complicated JSON with lots of different fields. And we have length specifications for each of the field and exactly what content needs to be in each these type of things. Haiku mean, we run across quite Haiku few examples of that. Commonly, that's how models are used in SAS software to add kind of generative AI sugar, as I call it, to SAS software. Then you need to get very specific about what you want the model to do. And then if if

Model Testing, Data Generation, and Fine-tuning Pipelines

Nick Jakobi 00:11:00

another model is powerful enough, it should also be able to do that.

Tim Scarfe 00:11:03

We Haikure building all of these kind of APHaikus, controllers, prompt injection, you know, applications on the top. And we think of them as being quite decoupled from the language model. And maybe in the future when the models become really general, then we can think of them as being decoupled. But that's not really a reality yet. So we need to do lots of engineering on the top. When we upgrade the model, we need to kind of test our entire, you know, life cycle and test all of our prompts and test everything is working correctly.

Nick Jakobi 00:11:31

Well, so I think I think of it as the model which is kind of being instruct trained to be generally good at instructions, and everybody does some of that. Then you have kind of API calls. So like rag, tool use, these type of things where effectively what's happening is that the API is using specific prompts under the hood. And we've done a whole load of intricate data collection and fine tuning the model to be really, really good at those particular prompts. And so that's 1 kind layer higher level, I would say, than just the generic model behavior. So if our APIs we train our model to be able to do certain things in our APIs. Most models have similar APIs or similar ways of doing the same things. But it's interesting to think whether there will be divergence there. It's not that hard to kind of copy what other people are doing. Mean it is hard obviously. But there's nothing that prevents other companies from copying what you're doing. So if they see your API does something, they're gonna think, oh, that's a good idea. We'll do something similar. And then they have to do the hard job of going and assembling the data and assembling the evaluations and training the model. Of that is trivial. But there's nothing that stops everyone continuing to compete on these different capabilities.

Tim Scarfe 00:13:12

Yeah. Really interesting. So I mean, you've seen behind the curtains. I mean, you you know how these things are trained and how they work. I'm I'm interested in your mental reference frame, you know, how you understand the behavior of of large language models.

Nick Jakobi 00:13:25

So on the on the product team, of course, we get lots of issues reported to us. So, you know, a client tried such and such and the model didn't work. And, you know, a large part of our job is to be able to kind of understand that, aggregate it with all the other things that are coming in, and then pick out the kind of failure mode or capability that needs improving, and then work with the modeling teams to generate the data and the evaluation to improve the model in that, to solve that failure mode or improve in that capability. And then getting the level right on that is a real art. If it's just 1 example, clearly you don't I mean, what you can do and we do do is like we get a prompt and it doesn't it's doing something completely weird. We will just kind of show the model what the correct thing to do in response to that prompt is. But it might not learn anything general from that. So if you can spot patterns of things going wrong and general behaviors that are going to be important for everything, you can you can create special datasets for that.

Tim Scarfe 00:14:42

Who knows if we spend another 10 x or a 100 x, maybe we'll we'll move the needle. But, you know, my my kind of naive take on it at least is that it feels like it's

The Economics of Model Size: Efficiency Over Scale

Tim Scarfe 00:14:50

saturating, and there might be economic pressures behind that as well and we're trying to build smaller models anyway. Mean, at some point, haven't got enough money to keep, you know, doing inferencing at at scale. But do you think it will get to a point where it's just good enough like the iPhone 15?

Nick Jakobi 00:15:07

I think we're already at a point where it's good enough. The way I think about it is, you know, you have these small amazingly powerful, you know, 7,000,000,000 parameter models. They're good for some things. They can be fine tuned to do some quite powerful things. More importantly, they're very cheap and they're very fast and kind of low low low latency. And it like the narrower the use case, the more you can you can train the model. So some of those some of the code completion models are even smaller than that and faster because they just do this 1 thing which is kind of like take half your coding bar and then suggest what else it could be. And latency becomes the most important thing there. So those those models, although they're tiny, they are they are perfectly formed for that use case. Then as you go up, you know, our 35 b model and our 103 b, so this is r and r plus. We find that some of our biggest customers prefer command r to command r plus because it works. And so why would you use r plus? It's just bigger and slower and more expensive. And really, I think kind of the algorithm that customers run is they have a use case, they have a problem. Look at all these models out here, some bigger, some more confusing. Clearly, need a model that can solve the use case for them. So it needs to be it needs to be high enough IQ, but there's no point in paying for more than you need. And it needs to be fast enough to solve the use case. And if the model's too big, it'll be too slow. So kind of go up the IQ until you get to a point where, okay, that 1 is good enough. And then there's no point in going more. So it's interesting to speculate. You know, we have our Command R. We have Command R Plus, which is higher IQ, slower, more expensive. Then there are much bigger models out there, you know, GPT 4, Claude Opus, all of these things. They're higher IQ again. They're like 3, 5 times the price of r plus. They're

much slower. They're too they're too slow for many use cases. You know, you can't wait seconds for this thing to come back. But there there are definitely use cases where you need something of that kind of power around. What are the use cases that are unlocked by a GPT 5 type model? I don't know. Interesting to speculate. But the question is, like, you have all these models kind of lined up, so you know, from smaller to bigger, and then you've got all of the customers running the algorithm I said, which is like, the smallest model that works for you, the smallest cheapest model that works for you. Then it's gonna be really interesting in about 5 years time or maybe even 1 year's time to plot like spend over model size, I suspect we will see the most spend or the most usage, let's say, because the bigger models are more expensive so that of course needs to be taken in account. But the most usage around the kind of command are g which is kind of GPT-3.5 Claude Haikyu type of level. I expect that's the sweet spot. There might there's also a whole load of use cases which will be in the kind of Claude Sonnet command r plus niche there. And then and then I think it's it starts to I mean, GPT-4 and Claude Opus get a lot of usage but mainly for training smaller models and creating training data for smaller models and that type of thing. I mean, obviously, they're used for very mission critical cognitively demanding tasks where you need accuracy and lack of hallucination and all those things. But they are expensive and they are slow.

Tim Scarfe 00:19:14

Yeah. Yeah. Eloquently put. And at some point, the the economics of it come into play. And I think we're really gonna see that playing out.

Nick Jakobi 00:19:21

Yeah. And I think there will be 2 2 kind of thrusts. The whole industry has been driven up until recently by how high IQ can your model be. Like how how cognitively powerful can you make your model? And the leaders in the space are the ones with the the with the biggest most powerful models. But it but I think we're really starting to see as people start to put generative AI into production, we are starting to see cost and latency and these things becoming more important. So then if you can get that kind of GPT 3.5

Cost, Latency, and the Sweet Spot for Production AI

Nick Jakobi 00:20:00

command r kind of level, if you can get that cheaper with lower latency, with higher throughput, like that is just as hard and demanding a research direction and just as technically hard to make that happen. But it may end up actually being the most commercially profitable direction to go in.

Tim Scarfe 00:20:25

Yeah. 100%.

Nick Jakobi 00:20:27

On the absolute abilities, large percentage of use cases and questions and types of questions you can ask these models, The medium sized models and the very large models will give equally good answers. And so you start having to do to ask these very complicated, involve, cognitively demanding type questions to really pull them apart. And then on something like chatbot arena, most people won't be asking those questions. They won't know how to ask those questions. Exactly. And then so then you're just getting into a whole load of personal preference that I prefer that 1, I prefer that 1, which is ultimately just noise. Yeah. And then, yeah. So I'm I'm sure we're getting into that because that is only as useful as the questions that get put into chatbot arena. And it's like telling apart the best mathematicians in the country. Like if you give them a simple kind of 16 year old test to do, they'll all get a 100%. So you have to come up with these incredibly complicated intricate questions to kind of tease apart their abilities. And and because chatbot arena, most of the questions just come from the general public, you won't see that.[5]

Tim Scarfe 00:21:40

Yeah. So language models are just trained on on data out there on the Internet. So presumably, its capabilities are too limited on just what's out there unless you believe that they have some emergent reasoning which transcends the training data.

Nick Jakobi 00:21:52

I think so. It depends. Not every company has access to that kind of data. So, you know, OpenAI, ChatGPT, amazing consumer success. They have a large amount

[5] LMSYS Chatbot Arena — Tool

of prompt data that came in because of that, which def has definitely given them an an advantage. And that they've, you know, they've got armies of annotators sitting behind. So with all this kind of real world prompt data coming in, and it gets annotated and adding to training. And that gave them a kind of a good early start, I would say. And then, you know, GPT 4 is then used to create training data. And so a lot of these open source models that you see doing really well have been trained on GPT 4 data. So all the early llamas and all the, you know, alpacas and all of this. We do get data at Cohere. We get data from people interacting with our website and entering prompts and getting answers and telling us whether they're good or not. We use that data. But it's actually very small compared to something like ChatGPT. And then our models get used on AWS and, Google Cloud and Azure. And it's part of our USP, which is that your data stays your data. It stays in your cloud. So we don't actually get to see that data. So all of the kind of annotation and data generation we do, or the very large part of it, we're actually generating the data from scratch ourselves. So prompts, completion, preference pairs, whatever we're doing, we're generating from scratch 1 way or another a lot of the time. And that certainly allows us to avoid some of the biases because we can really direct what we're generating. But at the same time, it's still humans doing it and we still get human biases in. And then a lot of the biases to models just creep in from from the pre training. And the pre training is is not really anybody interacting with the model. It's just entering all the information in the world that already preexists as much as possible. And and of course, that all contains biases.

Tim Scarfe 00:24:10

Yeah. Interesting. We've we've got things like Microsoft Office and we've got all of this software and usually it's pretty fully baked. I mean, you know, Google were famous I suppose for having things in perpetual beta, you know, back about 20 years ago. But it feels like language models were thrust on the general public. And I mean, maybe it's just a new norm, but you know, sometimes they get things wrong. And and now we're kind of saying to to people, don't you know, always always check the answers. Obviously, you've done rag and and stuff like that. But it kind of feels like they're being developed in a in a very open way.

Nick Jakobi 00:24:44

Yeah. I mean, you know, this is the biggest experiment that's ever been run. Yeah. I think I'm right in saying that GPT

AI in Education and Workforce Transformation

Nick Jakobi 00:24:50

usage is the fastest growing technology or the fastest adopted technology of any technology ever by by a factor of like 10 or or more. Because even mobile phones, although everybody wanted 1, there's kind of a a real bottleneck around producing mobile phones and selling them out. But with generative AI and not just ChatGPT, but suddenly it could get switched on in, you know, every single Microsoft application around the planet and Google Docs all around the planet and Bing and Google Search and all of a sudden, AI boom just becomes accessible to everybody. I would say within the space of within the space of 6 months but probably less, you know, from from hardly anybody to the majority of kind of computer users around and and mobile phone users around the planet.

Tim Scarfe 00:25:47

Yeah. Yeah. Interesting. So on AI safety, now that there are loads of, you know, huge concerns. I mean, I've got some written down here. So things like enfeeblement taking jobs, malicious use, concentration of power, bioterrorism, you know, like changing our relationship with technology, cultural risks in organization, the erosion of our agency, things like bias and fairness, automated warfare, persuasion, and also some of the more esoteric, urgential risks. So people talk about existential AI and stuff like that. And this is a really difficult line to walk because there's, you know, some people say, Oh, this is paternalism, and some folks are utilitarians and consequentialists, and they're kind of thinking, Well, I can tell the future, and I know that if you use these things, then bad things are going to happen. Therefore, I don't want you to do it. Versus some people who say, Well, we can't predict the future, and should be free. We should use the models however we want to use them. How do you think about AI safety?

Nick Jakobi 00:26:47

I mean, I think so generally, AI and generative AI are clearly changing the world. It's interesting to think how they will change the world. It's interesting how they will change the world of work especially. We can already see it. Know, every single SaaS software, everyone uses SaaS software, every company uses SaaS. Every everyone uses tools, personal productivity tools, you know, Google Docs, Microsoft Office, calendars, all of these sorts of things. All of these things are just going to get better and more powerful And each person is just going to be able to do more is how I I see it. And so if you've got a job now, you're probably pretty good because you'll just be able to do more. And you won't necessarily get paid more, but you'll be able to do more. I do worry that it means that the, you know, the number of people needed to do any particular thing becomes less. The number of peoples who want to do things remains the same. And so I do worry even kind of how education works now and the type of ways in which we're teaching people. It's so hard to know where to go and what's happened. I spoke at this conference of there's all these these business school professors and there were all these all these professors who ran business schools. So there was INSEAD and UCL and LSE and all of these places were were there. And it was really a kind of conference to just get together and just like, what are what are we doing about generative AI? How is it going to change everything? What are we going to do? And half the conference was about how do we stop people from cheating, how do we stop students from using it, how do we as professors use it in the right way and not in the wrong way and all of these kind of things. But then the discussion quickly went to the fact that might be able to hold back the tide maybe 6 months. But then of course, everyone's going to be using it. Certainly as business schools, the business school professors need to think through like if they're not thinking through how is this going to get used in work and how are we teaching people to use it to the best of their abilities and to take full advantage of it. But then it's Everyone's just stumped. Like, okay, well, what would that look like? What does that education look like? I have 3 sons. My middle son is having problems with his schooling at the moment. And I go backwards and forwards, part of me thinks, this is terrible. Like, you know, he's he's not as advancing as fast as 1 would hope. And part of me thinks, does it matter? Like, are there should he should he just be thinking creatively and just going and doing his own thing and finding his own way? Because it's it's so hard to know how traditional education way of especially if you're learning facts Yeah. Is gonna be helpful at all. Or even kind of like something that develops your analytic thinking or something that, you know, your medical knowledge, your creative skills, your all of this is within range.

Tim Scarfe 00:30:06

Yeah. It's so true. But the culture in a business, the operating system in a

AI Safety, Disinformation, and the Agency Question

Tim Scarfe 00:30:10

business is quite mimetic. You know, the way we share knowledge is, you know, through quite a physical process. And that's still a bottleneck. So even though the language models generate things for us, you still need to understand it. You still need to kind of, you know, filtrate that and diffuse it to everyone else in the business. And also, there are many kind of, planes that are currently outside the domain of generative AI. So, you know, we have a real emotional presence right now. There's there's there's something beyond the domain of which is captured with AI. So I think there's still a huge need for humans in the workforce. But it's so difficult to predict where it's

Nick Jakobi 00:30:46

going go. Yeah. Absolutely. Not going be anything sudden, as I say. I think just the productivity tools will get better and better. So you will need less people to do the same thing, I think. And then and then some some domains will just completely just be automated. So I think radiography is 1 of them. For instance, looking at looking at x rays and working out whether there's a tumor there or not.

Tim Scarfe 00:31:13

Well, on on that though, Hinton famously said in 5 years, there'll be no radiologists. And apparently, there are more now than than ever before. Yann LeCun put a tweet out the other day and he said, Well, where's the misinformation problem? Where's all of this harm people have been talking about it? I can't see anything. And that's quite a difficult thing to react to, isn't it? Because it happens so slowly and so perniciously over time. So this might be something that we can't see now, but we look back on in 50 years time where we think, oh my god, you know, this was a a huge thing.

Nick Jakobi 00:31:42

Yeah. And I think it's I think it's There's always been fake news. There's always been people lying on the Internet ever since the Internet. Now, it's about the quantity and the quality of that fake news and the fact that the actual people in the system, the accounts you're interacting with aren't necessarily people at all.

And so now it's possible to, you know, spin up a 100,000 people all with a certain political view then interact with everybody else and, and convince them. There's new and interesting ways in which AI can be used to pervert the course natural course of things. Yeah. Exactly. Yeah. But I but I also agree. It's it's very nontrivial to understand if it's going on and it and it by its very nature, it will appear if it's successful, it will appear like it's not going on. Yeah. If if you can spot it, it's not working.

Tim Scarfe 00:32:40

Yeah. It's so true. I mean, for me, the the big bright line, you know, I I like ripping into some of the x risk people. But for me, the the bright line is agency. If these models really do have agency, which means, you know, they they can actually set their own goals and they can act and understand and make the world like their preferences and be completely autonomous, that throws an entirely different complexion on the matter. Because without agency, they're just tools. Ultimately, we're still the bottleneck. If we ever transcend the agency gap, I think that throws an entirely different complexion on the matter and I'm glad people are thinking about it philosophically. But anyway, Nick, we're out of time. This has been a fascinating conversation. Thank you so much for coming on. Really appreciate it. Welcome. Amazing. Well, there you go. That was wonderful.

Nick Jakobi 00:33:25

Thank you.

References

[1] **Cohere**
https://cohere.com/
Company

[2] **Episode Shownotes**
https://www.dropbox.com/scl/fi/sjwqllqh1ughuu8ismonk/NickJakobi2.pdf?rlkey=d20wydhhala2hmh4yaoo55wyy&st=unso2w6i&dl=0
Paper

[3] **Nick Jakobi**
https://www.linkedin.com/in/nickjakobi/
Person

[4] **RAG Paper**
https://arxiv.org/abs/2005.11401
Paper

[5] **Cohere Command R**
https://cohere.com/command
Product

[6] **LMSYS Chatbot Arena**
https://chat.lmsys.org/
Tool