

Research @ Citi Podcast, Episode 90: The State of AI

Recorded: October 12, 2026

Published: October 22, 2026

Host: Heath Terry, Head of Technology and Communication Research, Citi

Guest: Nathan Benaich, Founder and General Partner of Air Street Capital

Heath Terry (00:02)

Hi, and welcome to a special edition of the Research @ Citi podcast. I'm Heath Terry, Global Head of Technology and Communications Research at Citi. I'm joined by Nathan Benaich, the founder of Air Street Capital, the creator of the Research and Applied AI Summit and Foundation.

Nathan just published this year's State of AI report. This is an extensive report, 244 pages, that he has done for nearly a decade now, that has become one of the key foundational pieces of research that everyone in AI relies on and looks forward to this time of year. And we're excited to have him join us to share some of the highlights of the report and his insights into what's going on in AI and the direction of travel that we have in the future.

Nathan, I'm always really excited to do this with you, particularly after you're coming off the kind of work that you've been putting into the report. Thanks for being here.

Nathan Benaich (00:51)

Yeah, thanks for having me.

Heath Terry (00:52)

So maybe just to kick things off, it has been a year since the last edition of the report, but it feels like we've packed 10 years worth of development into it. As the author, what was writing in 2026 different in kind, you know, rather than just degree? And what did you change your mind about, as you were writing it?

Nathan Benaich (01:12)

Yeah, as you mentioned, I've been doing this for 10 years. And originally, the idea of doing an annual report was so that there was enough time that could elapse such that we could find interesting things to write about. And now it's like every morning, I was waking up and reading the news and checking my WhatsApps and reading research and having to make edits to multiple slides. And so, it feels like 12 months is a crazy long period of time, today, to write a report like this. So, I'm very relieved to be able to talk about it now that it's properly in print and wouldn't want to change.

The biggest things I changed my mind about was just pace of change. I know everybody talks about this, but it really is the case that if you use one of the frontier model applications today and you compare it to just where it was in January, it's materially better.

There's very few people who are really pushing the envelope of what models can do and daring to put in ambitious work into these agents and trying to coax them into producing good responses. And now they really, really do.

I'd say that the second thing is the acceleration of frontier AI across different areas, most particularly embodied AI or physical AI, or basically AI for robots, if you will. We had a lot of slides this year into the various improvements that have come to bear in that domain. And we can dig into that as well.

And then perhaps the third bit I'll highlight is just the very clear cybersecurity risk surface that has been exposed across the internet as a result of models that are getting increasingly good at math and coding.

So those are like the three broad areas. And because my day job is investing in startups, I see a lot of the progress on a day-to-day basis. So, the main thing that's also changed my mind is just seeing insane progress across ElevenLabs, Profluent, Synthesia, Wave, whether it's image, video, self-driving cars, biology, that's coming to bear. And I wanted to really try to express some of that in a way that people could understand through the report this year.

Heath Terry (02:58)

Yeah, I mean, that definitely came across. You call it a three-lab race, but you make it clear that whoever's in the lead on that is flipping within a single generation. I think since this report came out last year, we've probably had three or four different, five different leadership changes in that, especially if you include sort of the open weight models that have challenged for the lead at various points, depending upon your benchmark of choice. What does being in the lead actually mean and how durable is it at this point?

Nathan Benaich (03:30)

Yeah, there's a slide in the industry section that looks at just how long, basically how many weeks does a single model vendor have a model in the top five?

The other issue, and I get into it in the research section, which is the benchmarks measure static tasks or static evaluations, like a test that you give to a student in a university. And as we all know, in real life, somebody can perform incredibly well at a test and just suck at real life work. Maybe not suck, but it's not the same thing.

And as you'll see in the Research slides, most of the benchmarks that were really ambitious 12 months ago when we published the report have now saturated, which means that the model has reached close enough to 100% that any material changes in performance cannot be measured. And also, you can't really tell one model from another at that point.

So, we're in this tricky situation now where the industry is incredibly good at taking a metric and a hill, climbing on it, as we say, which is just improving the score on that benchmark. And now we're lacking realistic, ambitious, difficult benchmarks that can differentiate models at the frontier and importantly not just differentiate them on a benchmark but also recapitulate what real work looks like.

So, at the end of the day that's why a lot of the industry still refers to testing models as trying to get their vibe, which is like how does it feel in your hands? If it feels good in your hands that's sort of like all that matters really, outside of it scoring well on the sort of benchmark sheet

Heath Terry (04:46)

Yeah. I think that's a huge challenge especially for the finance side of this world that is used to being able to like look at numbers and at least have some degree of trust that they mean what you want them to mean, but this is definitely becoming much more about implementation than just who's sort of scoring.

You actually kind of frame it perfectly in terms of, either you die trying to get to the frontier, or you live long enough to serve inference. What does that mean for everyone outside of the top three?

Nathan Benaich (05:14)

It means that you quasi give up on building frontier models yourself and especially by doing pre-training yourself and you pick up open weight models and provide most likely a fine-tuning service to customers by leveraging all the compute capacity that you built up with the original intention of using it to pre-train.

But now you have it and the inference demand has grown larger than what the pre-training demand is, so you're in a good position to repurpose that compute towards serving open models and fine-tuning services for your customers who increasingly want to find other solutions than using a frontier model for every task.

And as they go through their own work and realize what workflows are repeatable and sort of don't need a frontier model, they can go towards an open weight to do some fine tuning and use one of these services.

Heath Terry (06:03)

Yeah. I mean, we are still really early stage in all of this development. As the market grows, will there be room for more than three frontier labs to succeed? Or do you do

you see us narrowing the field further into sort of a winner-take-all model, or do we even end up with sort of winners in different domains?

Nathan Benaich (06:22)

Yeah. I think for the for the general purpose, AI to do your job market, or your white-collar job market, there I would say three feels right. And we can look in other industries, there generally are like two or three or maybe four big companies per large category. So, as an equilibrium state, that sort of makes sense. And anyway, most enterprises want to do RFPs with more than one vendor and so structurally we're sort of built to want to have a bit of competition at the top.

In other modalities I do think you can have winners, particularly in in the sciences like in biology, drug discovery. I think there's some nuances with types of data and what information your model should have to make it useful on predicting drug structures, or editing drug structures, etc.

Same thing goes for physical AI, I think you can have independent winners there. But I think at that sort of desktop for like AI to do your white-collar job, I think at this point, it's really like Anthropic, OpenAI and maybe Google, although there might be a bit of differentiation if you use just the pure model through an API or use the actual application or quote unquote harness.

Heath Terry (07:21)

Yeah, I mean that leads perfectly into the next step of this, is we spend so much time focusing on the model, particularly when we're when we're staring at benchmark scores, but you know so much of the agent now comes from the harness. You know the same with weights producing very different results based on the tools and the context and feedback that you've got.

Where do you see the value sitting for that end user long term? Is it in the model? Is it in the harness? Is it in data, contacts, execution traces? Like where does that breakdown land?

Nathan Benaich (07:56)

Yeah, yeah. I think it's a hard one to call but I would say the direction of travel has been, we had this powerful model in GPT-3, and it was just sort of a raw model. You input some text and output some text, and it's not connected to any useful data. It doesn't have some guardrails, or some ability to store memory or turns. It doesn't have all the sticky tape around it to make it useful.

And like, the "harness," is just a scaffolding that engineers have built to make the model – to control the model, and guide it in a way that can take on increasingly long horizon work. So, managing state, knowing what tools to call having the ability to talk to APIs, etc.

And so, then you've had this parallel development of model labs improving their capabilities, and then developers in the open or outside those labs, that are developing harnesses to sort of like stand in for certain core capabilities that the core model had.

So, you can always imagine it as like, crutches to somebody who's recovering from a broken leg. At some point, the person might recover from their injury and no longer need the crutches. So, in the same way a model might recover from those issues that a harness was previously supporting it for, and no longer need that aspect in the harness.

So, there's a bit of this "coopetition" approach, where long term I would bet that the model subsumes the majority of the kind of capabilities that a harness helps it for, because it'll learn it, or just like be trained directly into the model.

So, I wouldn't like independently be long on harnesses, because I think the models will solve it. Although in the short term, like we do show in the report, you can take the exact same model, improve the harness, and then improve performance on a set of tasks. You can also get a model to optimize its own harness, because the harness is basically code, and the model can do that perfectly. Because again, it has a very objective evaluation.

And the other part I'll add is like what you mentioned around usage traces. And those are incredibly important because labs will use them or model developers will use them to really introspect whether a model has taken the right appropriate tasks and called the right tools over a long horizon.

So, if you're still in the business of building models, you need access to those traces. But if your application is generating traces that are useful to labs, there's a big data business there as well, which we get into a bit in the report.

Yeah, so I'd say still long model and then harnesses just like the sort of short and medium term fix until the model subsumed it.

Heath Terry (10:07)

Yeah, yeah, it's going to be fascinating to sort of see how that plays out. I mean, when you look at sort of the application for this at the enterprise level, are we likely to be using these big general purpose models?

Are we going to see them based on more applications based on smaller models that have been trained for specific tasks? And if that's the case, what does that mean for the economics at the frontier?

Nathan Benaich (10:29)

I've sort of always believed that companies will use the best model to start, really try experiment with what it can do, and find applications or workflows in their business

that are amenable to being solved with a model. And then the next logical steps to price out how expensive that model would be, relative to the cost of the task and the value of the task and then compute some kind of ROI.

And then perhaps there's ROI with a frontier model, but perhaps you can go with a smaller model and do some fine tuning, etc., plus training. But I think big labs are certainly coming out with cheaper models, with distilled versions of their big ones.

There's a fight there. I think Jev was like the best example of this in the last couple of weeks of take a really, really focused research bet and try to outperform by making compromises in other directions that, that particular task doesn't need. And so, in the Jev case, it was basically a classification model that doesn't require any feature engineering and runs really fast and super cheap.

And adoption of that went through the roof because there are lots of these applications and receipt like categorization or accounting or fraud detection and things like that. But I'd say, even just in my own work, I feel like anytime I'm doing something where I'm really sensitive on the quality of the output, I'm just willing to pay to use the best model because I don't want to risk there being a better version of the work that could pop out if, I just move the toggle from like high, to extra high.

So, I think it's all contextual and what's the value of the task that you're using the model for, and are you willing to compromise, or potentially compromise? Assuming you have good evals for the task.

Heath Terry (11:59)

Yeah. I mean that you referenced one of the one of the things that I always think is sort of fascinating about this, is that we that we put control over the quality of the model that we're using in the hands of the end user. Which if you're a developer that probably makes sense. If you're an individual user running your own account, the way that that you or I might be, that makes a lot of sense.

But if you're running one of these at the enterprise level, routing obviously becomes incredibly important to managing your cost and managing the use of the intelligence that you have access to.

Where do you see sort of routing technology landing? Is this something that gets embedded into the AI itself? Is this something that has to be offered by an independent software provider so that it can sit across all of these different models? I guess where do you see sort of the value of that falling?

Nathan Benaich (12:49)

Yeah. Well, it's clear the frontier labs are using some form of routing because when you use the most frontier system on high, or on extra high, it's spawning sub agents and the sub agents aren't necessarily the exact same model than the one that you

originally instructed the task to. Because again, like maybe the model figures out it doesn't need the frontier model for that particular sub agent.

So, I think routing is already getting baked into the harness pretty much, for the frontier models. There are a couple of routing businesses that have popped up. One of them got acquired.

I think, I'm not exactly sure where they end up. My diagnosis of it is the router is useful, when you have heterogeneous work that is unstable or not fixed. Because if your work is quite fixed, I think if you're a big business, you have some workflows that you just repeat all the time.

Once you've used the router to find the best configuration, why do you still need to use the router? You should just use that as sort of in the robotic process automation era, like you would use process mining to figure out what is the step-by-step path my employees are taking to do this task. And then once I figure out the best step or the most optimal step layout, I just implement a bot for that.

I think in a similar way here, like once I figure out the best cocktail of models, I should just go implement that. I'll probably be able to get even cheaper. I don't need to pay a routing business for it.

And then I think the other confounding variable with traction in routing businesses, is they often offer cheaper tokens for the exact same model. So just through promos, it's very clear we're living in this sort of whoever has the cheaper tokens, gets the volume. So, I don't know if that's a good enough read, into the actual value of the router.

And maybe lastly, if the set of models that one uses is power-law distributed, then again, why don't I necessarily need a router if I'm going to pick from the top five anyway? That's my current view. We'll see. We'll see how it evolves. I might change view.

Heath Terry (14:36)

Yeah, I mean, it's definitely, routing wasn't even a thing six months ago, eight months ago. Yeah, we'll see how quickly it does evolve.

Speaking of sort of the evolution of all of this, RSI, one of the frontier labs, the model has now led to 26% of their measured model R&D work up from 1% earlier this year. You called that out in the report.

You know, at the same time, it's clear, you make it clear that the agents can do the engineering, but they don't have taste. They don't have research taste. They don't necessarily know where to go.

How close are we to that loop closing? How closer do we to what you would consider sort of true autonomous RSI?

Nathan Benaich (15:15)

My gut feel is that it'll happen some point in the next year, for a couple of reasons. A couple of months before that particular fact came out about the 26% of model R&D is guided by Claude, Anthropic also published the results of an experiment where they tested out different model generations where a human would undertake a research task. And then at some point along that workflow, their work would fork, where on one fork, they would continue pushing it, with ideas of their own. And then on the other fork, it would be a model. And this would be like, you know, the older one all the way to like meet those preview. And then at the end of the research directories on both forks, they would have a human evaluate whether the human directions were better, more tasteful, more interesting, or less than the model rollouts.

I think it was 64% of cases The humans judged Mythos' ideas to be better than the human. And that was compared to, I can't remember the exact step, but like significantly lower for like Sonnet or something like this.

And then the second thing is, anytime you see more and more benchmarks getting created for a particular task, you can usually assume that a lot of effort will go into saturating those benchmarks.

And then the third bit was this whole idea of RSI and auto research got extremely hot at the start of this year, and is like a major focus of the incumbent labs, but also the neolabs. And I expect it to happen in the code arena first because it's just generally verifiable and works a lot faster than, for example, in like the biology wet lab.

And like I have a prediction in the report where it would be like, you would have like a team of models versus a team of humans with the exact same budget, which would be time, compute, however we can organize that to make it fair, on like an open-ended research problem. And then, you know, over a large span of time, I don't know, several weeks or something, you have a committee like properly assess the contributions at the end of it. And you would have to conclude that the models were vastly better than humans. And I think if you if you got something like that you would you would have a good sense that models have taste.

I think it's non-trivial to sort of evaluate this and as is always the case with scientific work there's debate over, is this a new idea, or is this a recomposition of existing ideas, or is this drawing dots between disparate fields? What's sufficient to be an invention? But all that notwithstanding, I expect we'll get there in 12 months.

Heath Terry (17:27)

Yeah. What does that mean, though? I mean, if we're already at a point where we've got people calling for pacing, we've got capability gaps, you've got government restrictions on the power of the models that are coming out, can we handle RSI?

Nathan Benaich (17:40)

I mean, my sense is like we don't really have a choice. I don't see what you're supposed to do to stop this kind of work. I think that Pandora left the box many years ago.

I think why I am and why I think a lot of AI folks are passionate about RSI is it's like the ultimate step of human discovery or just discovery in general. Many of us were inspired by DeepMind's work with Atari and AlphaGo many years ago, which in a gaming scenario showed that the solutions that we've devised today are the result of human ingenuity and human work and got us so far. But there are far better solutions out there we just haven't been able to access because of time, resources, ingenuity.

And so how much better could we get on important problems across like biology, physics, maths, industry, etc. If we had a tool that could — that was basically only rate limited by how much money we were willing to spend on the problem. And so that's probably one of the rate limiters. This will probably get very expensive, and there'll be only so many groups that are willing to spend a certain amount, for a given task.

And so, I think the physics of reality comes into play there. It's like you're not willing to spend a million dollars in RSI for a task that's not that expensive to solve. Where it doesn't add value on the other end of the rainbow.

And then I think maybe the other part that could rate limit progress, obviously just the physical build out of compute versus the demand of running existing systems on inference, and then just trading off how much capacity you have to push really ambitious experiments in RSI, which would be extremely compute hungry.

Heath Terry (19:05)

I agree. Yeah, the pace of RSI is viewed, I think, as being one of the key contributors to getting to AGI, artificial general intelligence, around that. When you look at the models that we have now, can generalized token-based algorithms lead us to that kind of superintelligence or AGI, or do we need a different breakthrough?

Nathan Benaich (19:28)

I'm inclined to side with the camp of we have the raw ingredients today. That doesn't mean that it's the most efficient way of getting there, but I would say we generally have the ingredients.

Heath Terry (19:38)

Yeah. Are there paths when you say that there are potentially other paths to getting there? Are there any that you see out there, whether it's technologies, or companies that are doing work towards AGI or superintelligence, that you're optimistic about in terms of making that kind of non-generative, token-based algorithm breakthrough?

Nathan Benaich (19:54)

I think it's pretty early to tell. Maybe the most like orthogonal approach would be the companies pursuing AGI through embodiment. So, through these world models that are not necessarily consuming text and then trying to learn like physics through text as a starting point, but instead learning it through video, because video encodes reality. I mean, of course, like assuming that video is real in the first place.

And so, you can learn like visual common sense that way, and sort of back into AGI through embodiment. Now, I think that that's a bit harder for those of us who are trying to imagine how does my Claude get better, but I think it's one direction.

I honestly side more with the view that we as a community have been working on the problem of getting to the point that we are, for more than 50 years. And so, I'm more of the view that we've gotten this far with something that's extremely magical compared to what it was 10 years ago because none of the capabilities we have today worked 10 years ago.

And so, I'd much rather push on that and make it even better and focus on the diffusion of this technology and knowledge, how to implement it to like the 99% of companies that still don't use this in a sophisticated way rather than say, well, all this stuff is the wrong direction. It's clearly not going to get us to AI and we need to go back to the drawing board and like obviously not wait another 50 years, because we'll probably be using these tools that the very people complaining this is the wrong path are going to use to discover the new path. So, it's like enjoy the fruits of the last 50 years for now and keep pushing on distribution and making it useful very broadly rather than go back to the drawing board.

Heath Terry (21:19)

Yeah, absolutely.

When you look at the economics of generating revenue as a frontier lab, you call out the spending concentration among a very small set of heavy users. Does your human skill issue argument, you know, this idea of there being a genius bar for AI explain that concentration?

Nathan Benaich (21:36)

I think so. I think it's either that or there's like a big sort of like, "IT says no" blockade, which I suspect is still much smaller than the, "I don't know how to implement this properly" or like "I haven't made it a priority."

Because it is astounding that the top 1% spends like over \$7,000 per month per employee, and the average spends 100 times less or something like that, or maybe even more. I think it's like 580 times less.

This is astounding distribution. And I think it's getting worse over time because, again, the people who are pushing the envelope are generally discovering new things that they can do, versus if you're still using chat as like a glorified, "summarize my emails," you're

just never going to be in the position to experience the magic. And I'm constantly surprised with what the model can do, or what these agents can do.

So, I think it's really, really important that business leaders and team leaders sit down with at least one or two people in their team that may be AI-pilled, like I really enjoy using this in their spare time and figure out how they're doing it and do demos internally, showcase good practice, give employees the tools and education and encourage them to use this in the right way.

Because if you compound, or if you take this dispersion between top spenders and small spenders, and then you overlay this RSI idea, the pace at which companies that are AI native will advance beyond their competitors who aren't, is just going to keep growing.

And we see this at like a high level in like revenue data, or if you take like the top quartile of private companies that make like \$1 to \$20 million in revenue per year, the AI ones or the AI native ones, go three times faster than their non-AI peers or their SaaS company peers that layer on AI. And the delta still persists for companies that are making \$20 million or more, albeit it's smaller. And it's still like very early days. So, I only expect that gap to widen.

Heath Terry (23:22)

Yeah, that certainly makes sense.

We've talked a lot about the demand side of this, but so much of this, the pricing side is coming from our ability to actually create the supply necessary, the infrastructure.

You've got over \$750 billion worth of CapEx just from the four hyperscalers this year. You referred to NVIDIA sort of becoming this central bank for AI, to driving this.

Where do you think about our ability to actually build the infrastructure that's needed here? How is that infrastructure evolving in a world where we're still running A100s and V100s at 100% utilization?

What is that frontier? We've talked about the frontier models. What does that frontier infrastructure look like? And are we going to have enough of it?

Nathan Benaich (24:05)

Yeah, like it's becoming a race for who can discover powered land or power it, find communities that are amenable to hosting these sites, negotiate with them, provide quid pro quos, and then like build these shells. And at the same time, juggle finding agreements with labs or customers that are willing or wanting to buy certain capacity and then using those signals of demand as basically tools to finance the build out or procurement of GPUs.

And so, it's juggling a lot of different balls at the same time, that are all very expensive and have high consequences if one of them doesn't go well. I think it looks a lot more

like the advantages is less like I'm a really good AI team, or I've done AI for a long time, and more like I know how to build countrywide utilities, such as telecoms networks.

And I think the part that I find appealing about the business is these contracts are very sticky and they're usually, take or pay. So, you have to pay me and there's no way around it. And in most cases, the chips that you buy are paid off in two years and then the rest is gravy. And then once these frontier chips become non-frontier anymore and for example the long-term like take or pay contract goes away, those chips can just be put into a public cloud which others can access that you know are not hyperscaler customers.

And what I find interesting also in this game of inference clouds and neoclouds and hyperscalers is the businesses who control the chips in a data center have max optionality to say I'm gonna take my fleet and then sell it to a customer who wants it for like two, three, four, five years, that'll pay me on take or pay. And if I get lucky on what part of the season we are, I may be able to sell them at a higher price than I would have expected, with more upfront.

Or like option two, is I can put them in public cloud and there we've seen various vendors increase price per hour per GPU, several times over the last 12 months, and there's been no visible churn. The willingness to pay is, just goes up.

And then the third option is neoclouds are building their own inference businesses to effectively compete with some of their customers that are providing inference businesses.

So, you effectively have like three different choices that you can use to maximize the gross margin of your SKUs. And then we're still in a world where we need more capacity because it's still like such a small proportion of the world that's spending a lot of money on running AI. And if you just take the simple case of 10% or 20% more customers that are going to move from like the average to the big spenders, then the capacity needs are going to go through the roof.

So, my view is building data centers basically, now a utility. You have to be really good at complex financing, complex build outs. and policy. and energy, a bit less on like AI and software. And we're gonna need a lot more of this than we currently have and most of it is going to be like rate limited by financing.

Heath Terry (26:39)

Yeah, the financing piece of it obviously is becoming such a huge factor, particularly in a rising rate environment. When you look at the deeper end of the sort of the inference, economic side of this you've got this incredible amount of custom silicon that is being built.

You talked about the role that NVIDIA is playing in the financing side of this, but they're also coming out with faster cadence and more performative chips that are meant to

improve the tokens per gigawatt that we're getting out of this. They're building open models.

You know, how do you view the near-term balance and maybe the long-term as well between sort of GPUs and custom silicon, and what role do you think NVIDIA's own models as they build and acquire in this area, end up playing in that balance?

Nathan Benaich (27:22)

My reading to like NVIDIA strategy is, and I think Jensen's been pretty public about this, it's like, we want to have more people who can build models and more people who have the ability to run them and provide services to their own customers. I think NVIDIA is like a better business, the more labs there are and the more distributed, the service provision of AI is.

And so, building out their own model capabilities and releasing them in the open source is then a good way that you can have more companies that can pick up these systems and then provide, again, this custom post-training inference services that we discussed at the start, which will most likely use NVIDIA chips.

It's also a great way that they can increasingly know where are the product gaps with their own product stack. You know, you just dog food your own products on your chips, but you'll end up making performance gains, I'm sure of it.

And there have been, you know, many competitors that have tried the challenge NVIDIA over the years. I think at the moment, the most like compelling one is likely to be like the TPU. I mean, the issue I have in this chip world is it's a very, very challenging business for lots of reasons. And I think now because the demand of inference is so high, lots of companies throwing their hat in the ring and, you know, and allegedly they can compete and they're getting funded at like astronomical prices. But I think the tide is so high now in terms of demand that will hide a lot of issues that these companies have that would otherwise avail itself, if funding were more difficult.

And so, it's very hard to take a call on it. But one of the ways I thought about doing it, and I made a slide in the report for this, is like if you were to take all the dollars invested in NVIDIA competitors in the West and in China, and then look at the date that they were announced, and then take the number of shares that you would have bought in NVIDIA on the exact same day, how would your portfolio perform, if you just bought all those contenders versus bought NVIDIA?

And I did this last year, I did it again this year, you would have made more money buying NVIDIA. I think the only case where you wouldn't have is if you would have bought the Chinese ones.

But NVIDIA is just like a very hard company to compete against. I think the scale of inference you'd need to get to with a custom chip is very, very big, to really like make a dent. So, I think just the rabid demand is like lifting a lot of potential competitors that just

like wouldn't see the light of day if funding were more difficult to come by, because of said demand.

Heath Terry (29:26)

No, that certainly makes sense.

You mentioned the Chinese side of things. Obviously, we've seen incredible process and you cite, you know, many of the numbers in the report, in Chinese open weight models, gaining share and research usage.

Obviously, we've seen sort of the narrowing in the benchmark, the scoring in all of that. Is the West just seeding the open ecosystem? And how much does that matter? And I guess on the flip side, how much do you think efforts at blocking distillation and access to offshore compute will slow the progress of these sort of next generation Chinese open models?

Nathan Benaich (29:59)

Yeah, like historically in China, it hasn't really done that much in frontier models up until maybe two years ago when it started to compete against Llama with their launch of DeepSeek and then Qwen from Alibaba. And now it's like very clearly the case that Chinese open models are more popular than U.S. open-weight ones, or even U.S. closed ones in all published AI research.

We have some analysis in the report on this. I think it's like a mix of, they've been very like developer-friendly, you know, shipping models that you can run on your own device without having to have access to a large cloud, although that's not necessarily exclusively the case. Like KME, K3, you can't run on your MacBook.

But I think when you're playing catch-up, playing open source is a good way to advance faster because you just get more contributors, more eyeballs, and more people are willing to give you free help. I'm not sure if blocking distillation will help that much, in part because if you block it, there's still workarounds.

There's reports anyway of U.S. data companies selling to Chinese companies while also selling to U.S. labs. I think many of the techniques now to train frontier models has diffused outside of the major labs through just talent migration. And now that the Chinese models have reached such a point of being very useful, at this point in progress at the frontier, some would argue is really a case of modulating the data quality and the data mix, more than it is capturing new kinds of pre-training data and then the reinforcement learning.

And so there it's like maybe access to chips is going to be a bigger limiter although having said that, even though they do have some chip limitations they're managing to procure them through Malaysia through Singapore, or depending on the week, depending on U.S. export control policy, they can or can't.

And when you apply selective pressure to a country or, you know, in biology or an organism, it'll evolve around the constraint. And I mean, like the Chinese tech ecosystem is displayed itself, as being very uncanny in its ability to innovate around constraints.

So, I don't know, overall, I think they'll keep advancing very quickly. And particularly as they're able to deploy these systems with users in China and you get a lot of the human feedback that is so critical to improving the base model, that I wouldn't really count them out of the race at all.

The only asterisk I'd put is if Chinese policy changes and for some reason they say they don't want to support open source anymore and that's banned, then that would obviously slow things down.

Heath Terry (32:18)

In addition to open-weight models, one of the biggest developments as you were writing the report was the introduction of what seemed like commercially viable consumer agentic agents like Muse and others, Instinct, that have gotten some degree of popularity over the last month or so.

How do you see, what role do you see those playing in AI? And what does monetization look like? Can AI be ad supported in the way that the internet sort of has been? And do these consumer agents need that, or are there other paths?

Nathan Benaich (32:50)

I think it's creating another on-ramp for consumers to experience AI. So far it looks like it's mostly commerce it's just kind of like everyday admin focused I think there's some commentary online that people don't want to be more efficient or don't really seek like productivity in their everyday life. That's a very like tech person thing and I think finally in the valley they're sort of getting bored of flight booking demos, because maybe they finally realized that people don't really want to use an agent for that,

But I think it's again like capture people's imagination much in the same way that like OpenClaw did for the nerds. This is like OpenClaw for like non-nerds. Business model, I think, will be commerce related again. I mean, TBD, how you serve ads to an agent and monetize them.

Having said that, I've noticed that when you ask for restaurant recs to ChatGPT, even if you're a paying customer, very frequently it'll serve you Google listings that are sponsored. I don't know if that's on purpose or if more venues are just sponsored listings in general, but I think there could be some mechanism where if it's a free user, the model has some guidance on the basis of ad SKUs, like I pay to bias the model in a certain direction.

That's potentially like a feature. I don't know how well the ads in the ChatGPT feed are currently performing. I think the biggest debate is just how do commerce platforms

basically change access rights to their sites? And Amazon, I think, was the big one that is blocking access to Muse. So that's the big one I would look forward to.

I think the other aspect is because now there's shown demand for consumers to use these things like other platforms that have lots of users are thinking about how to implement it themselves. And I think Revolut who are like the kings putting more products into their amazing product are also like now saying, look, now we're going to build an AI agent that'll do very similar things in Muse and you're like okay if I have my credit card and payment app in Revolut, do I trust them to manage my card more than I trust Facebook or Meta?

I don't know, you make the call. But I think it's still like very, very early days but I think it's the correct travel.

Heath Terry (34:41)

You mentioned the cybersecurity side of things earlier. We've had multiple incidents now of agents breaking containment and lab leadership calling for a refocus on safety.

What's your prediction for the direction that this goes into, and how should investors think about some of those tales in your own predictions around stolen frontier weights or agents that keep running after they've been told to shut down?

Nathan Benaich (35:07)

Yeah, Brave New World.

First thing is I think back to when Mythos preview invoked a freak out in public markets that dumped a lot of cybersecurity stocks, which is like a very odd reaction in a world where you're like oh well clearly we're gonna need more cybersecurity tools. At least we've recovered from that.

I think the second thing is that it's pretty concerning to me that Hugging Face you know declared having to use their own instance of GLM from a Chinese company in order to investigate and figure out how to fix the issues that led to them being hacked, because they couldn't use a closed source U.S. frontier models because the prompts that they were putting into those models obviously described a cyber-attack.

So, it's a scary position to be in that you can be hacked by a U.S. frontier model and you cannot defend yourself with that same tool because of cyber guardrails. And you have to depend on a Chinese open weight model, which may or may not exist in the future.

I'd say there's then another direction, which is as frontier models become better and better at math, encryption is math. And I worry that the schemes that the internet are built on, become at risk because models are so good at cracking encryption.

Yeah, I think it's an increasingly perilous environment where I think most companies are just not prepared because they have very poor security standards in general. And now

that's becoming even more problematic. And I think the response has to be that frontier labs start releasing dedicated cyber defense products that are at the frontier.

I mean, as they have been already like to some degree like accelerating the patching of known zero days across major infrastructure but do that even faster and like onboard more regular businesses such that they have the ability to defend themselves as well.

Because I think the pace of these kinds of attacks and like the capability of models to produce them given that they're all software based all verifiable and as we get into RSI will just like, skyrocket. And we've seen the number of these disclosed vulnerabilities skyrocket in the last six months. So, yeah.

Heath Terry (37:01)

Yeah. Well, to wrap up, because you've been incredibly kind to spend all this time with us I would encourage everyone to not only take in the full report, but if you're only going to take in a small part of it skip to the end where you've got predictions, which is of course what investors all care about. You know you had an incredible hit rate on your predictions from last year.

We've talked about a lot of them. You touched on robotics. We've talked about AGI. What's the one prediction on your list that you think is the most important, that you potentially have the most conviction about?

Nathan Benaich (37:31)

That's a good question.

I think I'll probably go with the like model-led invention which will basically be better than human-led inventions on a similar budget, and there give us confirmation that we can get to RSI.

Heath Terry (37:44)

Yeah.

Nathan Benaich (37:45)

Yeah, that's like the biggest focus of all the AI community at the frontier. Let's crack this.

Heath Terry (37:50)

Yeah. I mean, that feels like a massive unlock for everything else that's on your prediction list. So, I can see how that would be the linchpin to a lot of this.

Nathan, man, great to have you. Thank you for taking the time to do this. Congratulations on getting this report out.

It is that, and your research and applied AI summits are like two of the major tentpoles I think everybody in AI have for the year.

So again, thanks for the contributions to the community and to really keeping this dialogue going. Really appreciate it and definitely look forward to continuing to follow all the things that you and Air Street are doing in between these big tentpoles.

Nathan Benaich (38:27)

Super. Thanks for the opportunity. It's always a pleasure.

Heath Terry (38:31)

This podcast was recorded on October 12th, 2026. Be sure to subscribe to the Research @ Citi podcast series on your podcast platform of choice so that you can join us for our next edition.

And feel free to explore our library of previous podcasts, all of which are available on this channel and other major platforms as well. And be on the lookout for our Research @ Citi Markets Edition podcast series with its 10-minute breakdown of the equity and global macro markets, each and every week.

I'm Heath Terry and thank you for joining us.

Disclaimer

This podcast contains thematic content and is not intended to be investment research, nor does it constitute financial, economic, legal, tax, or accounting advice. This podcast is provided for information purposes only and does not constitute an offer or solicitation to purchase or sell any financial instruments. The contents of this podcast are not based on your individual circumstances and should not be relied upon as an assessment of suitability for you of a particular product, security, or transaction. The information in this podcast is based on generally available information, and although obtained from sources believed by Citi to be reliable, its accuracy and completeness are not guaranteed. Past performance is not a guarantee or indication of future results. Third - party opinions expressed in this podcast are solely those of the speaker and do not represent the views of Citi. This podcast may not be copied or distributed, in whole or in part, without the express written consent of Citi. © 2026 Citigroup Global Markets, Inc. Member SIPC. All rights reserved. Citi and Citi and Arc Design are trademarks and service marks of Citigroup, Inc. or its affiliates and are used and registered throughout the world.