/mlst

ARC Prize Version 2 Launch Video! [Francois Chollet, Mike Knoop]

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: abdbbcda

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

ARC-AGI 2 Launch and Benchmark Architecture 3

OpenAI o3 Performance and Implications 6

Human Calibration and Task Design 10

Measuring Fluid Intelligence: Binary vs Spectrum 14

Program Synthesis and Test-Time Adaptation 18

O3 Failure Modes and Reasoning Limitations 23

Intelligence Dimensions and the Road to AGI 27

References 32

Overview

Francois Chollet and Mike Knoop join Tim Scarfe to announce ARC-AGI 2 and the ARC Prize 2025 contest. The new benchmark has been human-calibrated with 400 participants and adversarially designed so that frontier reasoning models score in the single digits, while every task remains solvable by at least two humans. The conversation covers the story behind testing OpenAI o3 on ARC v1 (where it scored 75-85%), what those results reveal about fluid intelligence versus brute-force capability, and why Chollet considers o3 the first model with genuine – if limited – fluid intelligence. Technical discussion ranges from chain-of-thought reasoning as natural language program synthesis, to the exponential failure modes of current reasoning systems on spatially complex tasks, to the deeper question of whether intelligence is best understood as a binary category or a spectrum of recombination depth. Chollet argues that intelligence is fundamentally about efficiency of knowledge acquisition and adaptation, not raw capability, and that ARC-AGI 2 is designed to measure exactly this distinction.

ARC-AGI 2 Launch and Benchmark Architecture

Mike Knoop 00:00:00

This is not just more of the same that we've seen in the past. We have now an existence proof that computers are able to do something that they've never been able to do before in in in history of humanity.

Tim Scarfe 00:00:10

arcprize version 2 has just been released, and even the Frontier Foundation models are failing spectacularly. Today,[1]

Francois Chollet 00:00:18

we are releasing ARC-AGI 2, the next version of the benchmark. ARC-AGI and ARC-AGI 2 is pretty much the only unsaturated benchmark that is feasible by real people, And so it's arcprize very good yardstick to measure how much fluid intelligence these models have, how close we are to true AGI.

Mike Knoop 00:00:36

And alongside that, we're really excited to be welcoming everyone to ArkPrize 2025. Contest kicks off officially now. It's gonna run all the way through the end of, 2025. The structure of the contest is very similar to last year. We're gonna have the Kaggle leaderboard running. We're gonna be testing, this year on the semi private dataset and wrapping up and testing the final leaderboard on the private dataset. We still got the big prize. It's unclaimed. In order to get the big prize, have to open source your solution, have high high degree of efficiency running on Kaggle. And and, yeah, we're really, really excited to see all the new ideas. ARC-AGI think there was arcprize lot that came out last year in 2024 that really pushed the frontier.

Francois Chollet 00:01:11

The next version of the benchmark, it's more challenging. It's extremely unsaturated. All frontier models are scoring effectively within single digit percentage. And it's it's the first time where we've, calibrated, the human facing difficulty of the tasks. So we actually hired, roughly 400 people. We tested every single task, and every single task has been solved by at least 2 people. So we know it's very feasible for humans. It's extremely average for any AI system today. This is the frontier.

[1] On the Measure of Intelligence — Paper

Tim Scarfe 00:01:42

The arcprize benchmark forces us to confront an uncomfortable truth about our pursuit of artificial general intelligence that the field has been overlooking.

Francois Chollet 00:01:51

Intelligence is not just about capabilities. It's also about the efficiency with which you acquire and deploy these capabilities. Intelligence is about finding that program in very few hops using actually very little compute. Like, look at the the amount of energy that arcprize human expense to sort of 1 hour task over, you know, 2, 3, 4 minutes. It's it's it's almost 0. Right? And compare that to model like o3 high compute settings, for instance, which which is gonna use, like, over $3,000 of compute. So it's never just an economics problem. Efficiency is actually the question we're asking. Right? Efficiency is arcprize promise statement. It's not capability.

Mike Knoop 00:02:30

The goalpost is is AGI. That's like what we're here to do. That was the whole point of launching arcprize in the first place was to raise awareness that there was this really important benchmark that ARC-AGI thought shared showed something important that like the the sort of research community in AI was missing about the sort of nature of of artificial intelligence. You know, this is like 1 of the things ARC-AGI think makes arcprize special and and very unique and important, ARC-AGI would argue. You know, there's arcprize lot of benchmarks in the world today. And to to my understanding and knowledge, pretty much every other benchmark, know, all the frontier benchmarks basically are trying to test for these, superhuman capabilities. Right? These, like, PhD plus plus type skills that you need to have in order to succeed at the benchmark.

Francois Chollet 00:03:16

It's not just compute. It's not just scale. You have to be scaling the right thing. You have to be scaling the right ideas. And maybe you have them.

Mike Knoop 00:03:22

ARC-AGI personally just keep getting like surprised and impressed by the Arch Price community. How much folks are pushing the frontier. And ARC-AGI think it's really exciting too because it means that individual people and individual teams can actually make arcprize difference. If we're kind of in an innovation constrained world and an idea constrained world, which ARC-AGI think ArcGIA2 shows, that means you out there could actually make arcprize significant contribution to the frontier of AGI. So if you're gonna end with the contest, go to arcprize, and good luck. Good luck. See you on the leaderboard.

Francois Chollet 00:04:01

The the sort of, like, test time optimization techniques or test time search techniques, that's the current frontier of AGI. Right? And and there are many ways to approach this. Of course, you can do you can do just test time training or or or in this case, you know, test time. You you can you can do search. You can do search of arcprize symbolic space. You can do search over chain of thought space, over arcprize token space, or you can do search in latent space as well. Right? So you have you have many fun ways to do it, but really the front is how do you adapt to novelty at this time by recombining what you know into some novel structure.

Tim Scarfe 00:04:35

MLST is sponsored by Tufa AI Labs. Now, are the DeepSeek based in Switzerland. They have an amazing team. You've seen many of the folks on the team. They acquired Minds AI, of course. They did arcprize lot of great work on arcprize. They're now working on o 1 star models and reasoning and thinking and test time computation. The reason you want to work for them is you get loads of autonomy, you get visibility, you can publish your research, and also they are hiring as well as ML engineers. They're hiring arcprize chief scientist. They really, really want to find the best possible person for this role, and they're prepared to pay top dollar as as arcprize joining bonus. So if you're interested in working for them as an ML engineer or their chief scientist, get in touch with Benjamin Crusier. Go to twoferlabs.ai and see what happens.

OpenAI o3 Performance and Implications

Tim Scarfe 00:05:19

Well, Mike, it's amazing to have you on MLST. Welcome.

Mike Knoop 00:05:22

Yeah, thank you so much. We're very excited to be here today.

Tim Scarfe 00:05:24

Mike, arcprize-AGI hear that you guys have got some very exciting news today. Tell me about it.

Mike Knoop 00:05:28

Yeah, super excited today. We're back. We're really excited to be launching both arcprize AGI 2 alongside an updated arcprize Prize 2025 contest. Both are going be launching today, and you can go to arcprize to learn more and enter the contest.

Tim Scarfe 00:05:42

Okay. And in arcprize nutshell, what is v2 and how is it different from v1?

Mike Knoop 00:05:45

The way ARC-AGI sort of think about it is arcprize AGI 1 was arcprize benchmark that was designed to sort of challenge DeepSeek learning. And ARC-AGI 2, in contrast, is really arcprize benchmark that's designed to challenge these new AI reasoning systems that we're starting to see from pretty much all of the Frontier Labs. And 1 of the really cool things about RPGI 2 is we're basically seeing models, AI systems that are purely based on pretraining effectively scoring 0%. And some of the Frontier AI reasoning systems, we're in progress of testing them right now, and we're sort of expecting single digit performance. So arcprize really big update in terms of over ARC-AGI 1 from 2024.

Tim Scarfe 00:06:24

So the the original version of of arcprize was very much based at these kind of foundation models that didn't do reasoning. Version 2 is tuned for the reasoning models. But what would you say to the charge that you're moving the goalpost? ARC-AGI mean, how is arcprize v 2 sort of like meaningfully an evolution of the benchmark?

Mike Knoop 00:06:41

Yeah. ARC-AGI mean, think the way that ARC-AGI think about it is that the goalpost is is AGI. That's like what we're here to do. That was the whole point of launching Ark Prize in the first place was to raise awareness, that there was this really important benchmark that ARC-AGI thought showed something important that like the sort of research community in AI was missing about the sort of nature of artificial intelligence. So that's kind of our goalpost. And the definition that ARC-AGI use for AGI and the 1 that the foundation adopts is assessing this capability gap between humans and computers. And arcprize foundation is really to drive that gap to 0. ARC-AGI think it would be hard to argue that we don't have AGI if you look around and you can't find any more tasks that are very straightforward, simple, and easy for humans that computers can do as well. And the fact is that we were able to find still lots of those tasks. In fact, all of the tasks in ARC-AGI 2 data sets sort of fit into this category of things that are relatively easy and simple and straightforward for humans and comparably very, very difficult and hard for AI today.[2]

Tim Scarfe 00:07:40

Okarcprizey. Cool. So ARC-AGI know you guys have done loads of human calibration and we'll talk about that in arcprize minute, but the fundamental philosophy of of the of the arcprize challenge is focusing on human gaps. But at the same time, AI models are becoming superhuman in so many respects. So is the big story the human gaps, or is the big story expansion of capabilities that are superhuman?

[2] ARC Prize Foundation — Organization

Mike Knoop 00:08:01

You know, this is like 1 of the things ARC-AGI think makes arcprize special and very unique and important, ARC-AGI would argue. You know, there's arcprize lot of benchmarks in the world today. And to my understanding and knowledge, pretty much every other benchmark, all the frontier benchmarks, basically are trying to test for these like superhuman capabilities, right? These like PhD plus plus type skills that you need to have in order to succeed at the benchmark. Humans like can't solve the problems that are in these benchmarks, so you have to be very, very, have arcprize lot of experience, arcprize lot of education, arcprize lot of training in order to be able to sort of even get close to sort of solving the benchmarks as arcprize human. And ARC-AGI think those are important, those are useful, but ARC-AGI think it's actually more illustrating about something that we're missing from the nature in the sense of artificial intelligence by looking at the gaps, the remaining gaps between what's simple and easy for humans and what's hard for AI. Think that's much more of an inspiring story. ARC-AGI think it's 1 where it's actually necessary to target this in order to actually get AGI that is capable of innovation. You know, ARC-AGI think this is 1 of the main reasons ARC-AGI got into AI and AGI in the first place, was being really inspired and excited about trying to build these systems that would be capable of compressing science timelines. And if all we have is AI that looks like we had it at the beginning of 2024, right, based on pre training, based on arcprize memorization regime, you're never going to get to that because these are systems that are fairly going to reflect back the experience and the knowledge that humanity has sort of gained over the last 10,000 generations, as opposed to being ones that are capable of producing new knowledge, new technology, adding to humanity's colossus of knowledge and technology. If we want systems that can actually do that, we need AGI. And this definition that we've sort of used for this foundation of easy freemiums and our heart for AI, ARC-AGI think is, if we can close that gap, we'll actually get technology that's capable of doing that.

Tim Scarfe 00:09:54

ARC-AGI wonder whether you think we arcprizere just about 5 discoveries away from AGI because there's gonna be version 3 of the art challenge. They'll presumably, there'll be version 4. Is multidimensional. And ARC-AGI can see this both ways. Right? Because, you know, many critics of AI, they are almost gaslighting us. They're they're saying that this amazing technology that you're using, it doesn't work. And I'm like, well, yeah, it does work. And would it be the case that the criticisms will become more and more kind of philosophical and they'll say, oh, you know, because it because it's not biological or whatever, it's not the same thing? Or do you think meaningfully we're about 5 steps away from AGI?

Mike Knoop 00:10:29

Think this is why benchmarcprizercprizercprizercprizerks are important. And ARC-AGI had arcprize similar question actually, you know, when ARC-AGI was starting to get back into AI in 2022 and trying to understand the world. Like, are we on track for AGI or not? How far off are we? And ARC-AGI find that it's really, really hard to get arcprize sense of understanding of the capabilities of all these systems purely by using them. You can certainly get arcprize sense by just interacting with them. But if you really want to understand what are they capable and not capable of, you really need arcprize benchmark to discern this fact. This is 1 of the interesting things that ARC-AGI picked up from building AI products at Zapier as well. It's very different building AI with AI than it is classic software, right? 1 of the big differences is when you're building classic software, you can build and test with 5 users and know, okay, hey, this is going be like, this product can scale to millions. It's going work the exact same way. And that's fundamentally just not the case with AI technology. You really have to deploy to large scale in order to assess how it works. You need arcprize benchmark alongside that scaling in order to tell you, hey, is this system working or not?

Tim Scarfe 00:11:27

Wharcprizercprizercprizet were the main lessons that you learned from version 1 that you moved into version 2?

Mike Knoop 00:11:31

ARC-AGI think, so ARC-AGI 2 has been in the works actually for several years. Francois started working on it, crowdsourcing some tasks for it years and years ago. There was arcprize bunch of inherent flaws we ran into with it that we learned as we started popularizing the benchmark over the last year or so. 1 of the things we learned was that arcprize lot of the tasks were very susceptible to brute force search. So if that's something that has 0 intelligence at all and we wanted to minimize the sort of incident rate of tasks that were sort of susceptible to that, we hadn't human calibrated it. Anecdotally, we relied on some anecdotes to say that, hey, ARC-AGI 1 is easy for humans. We had arcprize couple sort of STEM, 2 STEM folks who had taken the whole data set, including the private set, and were able to solve 9899%. But we're relying on anecdote. We didn't have that calibrated across the sort of 3 different data sets that we had. Then And we had all these AI frontier reasoning systems come out over the last 3, 4 months. And we've gotten arcprize chance to study these and learn what are the sort of qualities of ArcTasks that remain very, very challenging for these AI reasoning systems, which we can get into if you're curious. And so those are the main sort of insights and learnings that we took from ArcGiA 1 to try and produce an ArcGiA 2 benchmark that ARC-AGI think will be arcprize useful sort of signal for development this year in artificial intelligence.

Tim Scarfe 00:12:46

Can we quickly touch on the open AI situation? So in December, they they well, they didn't launch, but they gave you access to o 3, and it got incredible performance on arcprize v 1, human level performance, something that we just[3]

Human Calibration and Task Design ●

Tim Scarfe 00:13:00

didn't think really would be possible so quickly.

Mike Knoop 00:13:02

Yearcprizeh. It surprised me.

[3] OpenAI o3 Performance on ARC v1 — Paper

Tim Scarfe 00:13:03

It came out of nowhere. arcprize-AGI mean, can you just tell me the story behind that?

Mike Knoop 00:13:06

Yeah. Yeah, I'll tell you. This is 1 of the reasons why I'm always hesitant to make predictions in AI about timelines. ARC-AGI think it's very easy to sort of make predictions along smooth scaling curves. But the nature of innovation is it's arcprize step function, and step functions are really, really hard to predict when they're going to come out. ARC-AGI think the best thing that ARC-AGI can say, having spent some time with 3 and looking at how it performs at ARC, is that systems like 3 demand serious study. This is not just more of the same that we've seen in the past. We have now an existence proof that computers are able to do something that they've never been able to do before in the history of humanity, which is ARC-AGI think really, really exciting. ARC-AGI think there's still a long way to go to get to AGI, but ARC-AGI do think that these things are important to understand and sort of even discern how they work from a capability standpoint in order to make sure that future AI systems that we're developing and building look more like this and not like the sort of pre training pure scaling regime that we've had in the past. So still remember the to sort of give you the anecdote, ARC-AGI still remember the 2 week period, the sprint we had on testing 'three. It was right at the end of the contest. We had wrapped up our 2024 in, ARC-AGI think, early November last year. And we had a 3 week, 4 week period where we were really, really busy on judging all the final submissions, the papers, getting together the technical report, and we were dropping it on all the sort of results on a Friday. And ARC-AGI was really hoping and anticipating that ARC-AGI was going to have a nice relaxing holiday period in December. And the day that we dropped the technical report, had a reach out from 1 of the folks at OpenAI who said, hey, we'd really love you to test this new thing that we're working on. We think we've got some impressive results on our AGI 1. And so that kicked off a very, very hectic, fast, frantic 2 week period to try and understand, okay, what is this system? Like, does it reproduce the claims that OpenAI had on testing it? And what does this mean for the benchmark? What does this mean for AGI? And ARC-AGI think we were able to show the final result was that O3 on its sort of high efficiency setting, which fit within the sort of budget constraints that we'd set out for our public leaderboard, got about 75% or so. And then they had a high compute version which used, ARC-AGI think, like maybe 200x more compute than the low compute setting, which was able to score even 85%. And this is really impressive. ARC-AGI think this shows a system like O3 has this, you know, sort of binary switch. We've gone from a regime where these AI models have

no ability to adapt to novelty to something like O3 where in its existence proof of now an AI system can adapt to novelty in a a small way.

Tim Scarfe 00:15:47

Breaking this down a little bit, there were there were some interesting caveats that you just alluded to. So first of all, they did some kind of fine tuning and people at the time joked that, you know, isn't it scandalous that they were training on the training set? So based

Mike Knoop 00:16:00

on Yeah, this like a very bad Yeah. This is a very poor critique. ARC-AGARC-AGARC-AGARC-AGI think it misses the point of the benchmark. ARC-AGI think the folks who feel this way, it's just because they're so used to thinking about benchmarks and AI from the pre training scaling regime where like, hey, if ARC-AGI trained on the data, that's cheating them to test on the data, right? And that's true in the pretraining regime. But ARC is a very, very special different benchmark where it explicitly makes a training set available with the intention to train on it. This is very explicit. This is like what the benchmark expects you to do. We expect AI researchers to use the training set in order to teach their AI systems about the domain of ARC. And then what's special is we've got a private data set that very few humans have ever seen. The private data set does not look like the training set. It requires you to generalize and abstract the core knowledge concepts that you learned through the training set at test time. Fundamentally, you cannot solve like the Arc AGI 1 or 2 private data sets purely by memorizing what's in the pretraining set. This would be like maybe a crude analogy would be if ARC-AGI was going to teach an AI system on grade school math and then test it on calculus, This is very similar to the type of thing that we do with art where the training set is much simpler, easy curriculum to learn on, and then the test is a much more difficult 1 where you actually have to express true intelligence.

Tim Scarfe 00:17:17

Have to

Mike Knoop 00:17:17

have an actual capability of adapting to novelty at test time to solve it.

Tim Scarfe 00:17:20

Okay. Okay. All of all of that is fine, but that, you know, there's a couple of things, right? They 2 and a half thousand dollars per task or or more. That means they were probably doing samplings. Were doing a ridiculous amount of, completions. They were doing solution space prediction, which is very interesting. But but the main thing, Mike, just deepen your bones, deepen your bones, do you think that they were training on APARC-AGI data or or surely they were training on a whole bunch of data to do that well. And the extension of the question is, when they release the vanilla version of it, what performance would it get compared to their their tweaked version?

Mike Knoop 00:17:52

We will test that as soon as it comes out and ARC-AGARC-AGARC-AGARC-AGARC-AGARC-AGARC-AGARC-AGI would love to report the results on that. They told us all they did was train on the training set and ARC-AGI believe that's what they did.

Tim Scarfe 00:17:58

Okay. Very interesting. And just comment on the solution space prediction. ARC-AGI mean, ARC-AGI ARC-AGI was amazed that just predicting the output space directly, they could do so well. ARC-AGI mean, doesn't doesn't that almost take away from the idea that we need to have discrete code DSL type approaches if you can just predict the solution space so well?

Mike Knoop 00:18:17

Effectively what O3 is doing is it's able to use its pre trained experience and re-combine it on the fly in face of a novel task. It does this through a regime called chain of thought. This is all informed speculation, by the way. We don't have confirmed details. This is just my personal assessment of how these systems work, particularly things like O1 Pro and O3. If you compare them with systems like OR1 or O1, these are systems that basically spit out a single chain of thought and then use that chain of thought in order to ground a final answer. This is distinct from how systems like O1 Pro and O3 work, where they actually have the ability to do multi sampling and recomposition at test time of that chain of thought. This allows them[4]

[4] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models — Paper

Measuring Fluid Intelligence: Binary vs Spectrum

Mike Knoop 00:19:00

to build novel COTs that don't show up anywhere in the pretraining, not in the existing experience, and allows these systems to reach more sort of effective or more situation space effectively based on what was in the original pre training. Fundamentally, these systems are a combination of like a DeepSeek learning model and a synthesis engine that is put on top, and ARC-AGI think the right way to think of them is these are really AI systems, not single models anymore.

Tim Scarfe 00:19:21

Yeah, and ARC-AGARC-AGARC-AGARC-AGI agree with you. It's really funny though how you the critique in the community because, know, Gary Marcus is now saying, oh, it can't draw pictures of bicycles and labels and label the parts. Whereas we see o 1 pro and o 3, and it really does seem like a dramatic move towards, you know, intelligent systems. But, Mike, can we just quickly talk about the testing methodology? So the the real Sure. The real work that you guys did was you got a whole bunch of human subjects and you you had the ARC-AGI think you had was it 400 test subjects and they all needed like, at least 2 people needed to solve every single task and you had to do this experiment design and you had to balance complexity of the tasks and so on? How did you do all of that?

Mike Knoop 00:20:01

Yeah. This was, so this was 1 of the biggest things we wanted to fix with RKJI 1. We didn't we never had a formal human calibration study on how do humans actually do on these things. We relied on anecdote. So we had to set up a testing center down in San Diego. We recruited tons of just folks from the local community all the way from Uber drivers to single moms to UCSD students and brought these folks out to go take ARC puzzles. Really cool. We'll to share some of the photos, these testing shots where you have dozens and hundreds of people taking ARC tasks on laptops. Our goal originally with the data set was to ensure that every single task that we put in ARC AGI-two was solvable by at least 1 human. And what we actually found was something even a higher standard, think, which was that we found that every single task in the new V2 data set is solvable by at least 2 humans under 2 attempts. And these are the same rules that we give to systems on the benchmark, both on the contest as well as the public leaderboards. ARC-AGI think this is a pretty good assertion of a straightforward comparison we can actually use now between, hey, are these tasks easy and straightforward for humans? Yes. Are they hard for AI? Yes. Like ARC-AGI said

before, frontier systems generally are getting close to 0 or single digit percentages on on these tasks now.

Tim Scarfe 00:21:12

Okay. But the idea though is this Moravex paradox, right, which is that, you know, basically while while we can select problems that are easy for humans and hard for AARC-AGARC-AGARC-AGARC-AGARC-AGARC-AGIs, we haven't got AGI yet. But ARC-AGI was looking through some of your challenges, and and ARC-AGI ARC-AGI felt that some of them were very difficult. Like, it would have taken me 5 or 6 minutes of DeepSeek thought to to get it. Like, are you are you finding that it's still easy to, you know, to find these things that are easy for humans and hard for AIs, or are you kind of scraping the barrel a little bit?

Mike Knoop 00:21:41

So ARC-AGARC-AGARC-AGI think easy for humans, hard for AI is a relative statement. The fact is these Arc v2 tasks were solvable by human on a $5 per task solve rate budget. They were solvable in 5 minutes or so. And AI cannot solve these at all today. And so yes, ARC-AGI do think if you look at tasks, you have to think about them. There's some thought you need to put in to sort of ascertain the rule. But the sort of data, ARC-AGI think, speaks for itself that we've got every single task now in the V2 data set from the public training set, I'm sorry, the public eval set to the semi private set to the private data eval set. Every single 1 is solvable by at least 2 humans under 2 attempts. And yeah, these frontier systems can't solve these things at all. Or if they can, with a really, really expensive budget to your earlier point, thousands of dollars per task.

Tim Scarfe 00:22:26

So you guys have been cooking. You're already working on version 3 of ARC. What can you tell us about that?

Mike Knoop 00:22:32

So the way ARC-AGARC-AGARC-AGARC-AGARC-AGARC-AGI kind of think about the multi versions here, ARC AGI 1 again was designed to challenge DeepSeek learning as a paradigm. ARC-AGI 2 is designed to challenge these AI reasoning systems. ARC-AGI don't expect that ARC-AGI 2 is going to be as durable right, ARC-AGI 1 lasted for 5 years. ARC-AGI don't expect ARC-AGI 2 is going to be quite as durable as that. ARC-AGI hope that it will continue to be a very useful signal for researchers over the next year or 2. But yeah, we've been working on ARC-AGI 3 and ARC-AGI think the pithy way to talk about ARC-AGI 3 is it's going to challenge AGI systems that don't even exist in the world yet today.

Tim Scarfe 00:23:08

Can you tell me about the foundation that you're setting up?

Mike Knoop 00:23:10

Yeah, so this is 1 of the big cool things ARC-AGARC-AGARC-AGARC-AGARC-AGARC-AGARC-AGI think from Arch Price 2024. When we launched it, it was very much an experiment. You know, our ambitions were not quite what they were. ARC-AGI think when we went into 2024, our main goals were just to raise awareness of the fact that this benchmark exists. And ARC-AGI think what we found was, or what ARC-AGI personally found, ARC-AGI just kept getting surprised by the community around Arc. ARC-AGI remember this really specific moment when 1 preview came out and there was thousands of people on Twitter demanding that we test this new model on Arc. And that was not my mental model of what this benchmark was or what the community was. And that was so cool. That moment happened again when we ended the contest. That moment again happened when we launched the results on 3. And this kind of showed, ARC-AGI think, hey, there's a real demand for what ARC is providing. There's a real demand for benchmarks that look like this, that ascertain these like capability gaps between humans and computers. And so we set up this foundation in order to basically be the North Star for AGI and continue to produce useful, interesting, durable benchmarks in the spirit of trying to discern what are the things that are simple, straightforward, easy for humans and still remain impossible or very, very difficult for AI. We're gonna carry that torch all the way until we get to AGI.

Tim Scarfe 00:24:25

As you can see now, all of the large AI labs, they're focusing on reasoning, and I'd like to think that ARC was at least a small part of that. And you folks are very focused on open source as well.

Mike Knoop 00:24:36

Mark Jen said specifically on the podcast that it was that they'd been thinking about Arc v 1 for years.

Tim Scarfe 00:24:40

There you go. Well, yeah. Exactly. But just tell me a little bit about that. So there's the industry impact, but you guys are really focused on open source as well. So how do you see those 2 things?

Mike Knoop 00:24:51

So my sort of overriding philosophy at this point is AGARC-AGARC-AGI is the most important technology that humanity is going to develop. And if it is true that we are in an idea constrained environment, we still need new ideas to get to AGI, which ARC-AGI think our AGI 2 shows is true. If that's true about the world, then ARC-AGI think we should be designing the most innovative sort of ecosystem and environment across the world that we possibly can. This is 1 of the reasons why we launched Archiprise originally internationally to reach solo researchers, to inspire researchers again to go try and work on these new ideas, get past this pre training regime, try something that we knew that it needed to be something beyond this and even beyond what we have today. And ARC-AGI think if you look at a really healthy, strong innovation ecosystem, you're going look at 1 that is very open and there's a lot of sharing and there's a lot of diversity of approach. And this is in contrast to an ecosystem that would be very closed, very secretive, very dogmatic, very monocultural. And so those values of openness, those values of sharing are what the ArcPrice Foundation stands for in order to sort of increase the chance that we can get to AGI soon.

Tim Scarfe 00:25:53

So talking about the the version 2 of of the Arc challenge, can can you just give us the elevator pitch of that?

Francois Chollet 00:25:59

Sure. So Arc 2 is basically a new version of Arc that keeps the same format but tries to address the main flaws that we saw in ARC 1. So for instance, in ARC 1, we knew that there was a little bit of redundancy across tasks. So we we saw that actually very early on, as early as the the 2020 kind of competition. And also, Ark 1 was way too brute forceable. So back in 2020, what we did after the Kaili competitions that we tried to look at all tasks that were sold at least once by 1 entry in the competition. And we found that half of the private datasets could be sold, in fact, just just, yeah, the sort of basic brute force program search methods that were that were deployed during the first competition. And so that means half the dataset actually doesn't give you very good signal about AGI at all. So the other half was actually good enough, required enough generalization that the benchmark overall was still useful and still lasted quite a few years after that. But it told you from the start that there were some pretty significant flaws, which is expected, by the way. Like, you know, when when ARC-AGI started creating ARC back in 2018, 2019, ARC-AGI was flying blind. You know? ARC-AGI was ARC-AGI was trying to capture my own thoughts, my own, you know, intuition about what does it mean to journalize, what what what is abstraction, what is reasoning, and that turned into this benchmark. But ARC-AGI could not anticipate what kind of AI techniques would would be used against it. And so, yeah, as it turns out, a lot of it could be brute force. So ARC 2 completely addresses that. You cannot score,[5]

Program Synthesis and Test-Time Adaptation

Francois Chollet 00:27:30

you know, higher than 1 1 or 2% at most using brute force techniques on ARC 2. So that's good news. And other than that, we generally try to make it a little bit harder. So what we saw with ARC 1 is that it was very easy to saturate for humans. Like, if if you're if you're, you know, like a a STEM grad, for instance, you could very easily get 100% or within within noise range of 100%, like something like 97, 98. So that means that you were not getting a lot of useful bandwidths to compare AI capabilities with the capabilities of of smart humans. And if you just make it a little bit harder, then you get more range when if you're not if you're not very intelligent, you score lower. If you're very intelligent, you score higher. And you're not super likely to completely saturate it until you are at the very the the the very top end of the distribution. So that's that's what arc 2 is. Same format, same basic rules, so we're only using core knowledge. You you have this input output pairs of grids that are at most 30 by 30, but the content is very dif-

[5] ARC Prize 2024: Technical Report — Paper

ferent. You're not going to find tasks where you only have to apply 1 basic rule that could be anticipated in advance, like, you know, some kind of gravity, things falling task, or symmetry task. All the tasks are very compositional. So you have multiple rules. You have more objects, the grids are generally bigger, the rules can be chained together or can be interacting together, and that makes it completely out of out of reach for brute force methods. And as it turns out, it also makes it out of reach for the the the the base LNP training paradigm.

Tim Scarfe 00:29:11

You're saying that you've made them more compositional and iterative and harder for humans.

Francois Chollet 00:29:17

That's right.

Mike Knoop 00:29:18

So Could you

Tim Scarfe 00:29:18

give me a little bit more detail on that? ARC-AGARC-AGARC-AGI mean, if you think about it, there are there are different dimensions of things that AI models can do and there are different dimensions of things that humans can do. Have you sort of quite diversely explored that? Or ARC-AGI mean, could you just give me a bit of a breakdown of the task characteristic?

Francois Chollet 00:29:34

So in ARC 1, you had many tasks that were very very basic where you just had 1 rule. Let's say for instance, you have a few objects and you have to flip them. Right? So this is an example of a task that's easy to brute force because flipping is something that you can you can acquire via pretraining as a concept or that you could just hard code in a in a brute force program search system. So if that's the only rule you have to apply and you just apply it once, that's not compositional. That's actually pretty easy to anticipate. That's easy to brute force. So a compositional task is gonna be a task where you have more than 1 concept and typically they're gonna be interacting together. Like an example of a very simple compositional task is let's say you have object flipping, but also the objects are falling. Right? So you have 2 rules to apply to each object at once. But that again is a kind of task that could still be found via brute force program search. If you have as as key elements in your DSL, gravity and and flipping for instance. And so you want to create tasks that are that are where where the rules are chained to a to a sufficient level of depth that there's no way you could you could find that chain by just trying every possible chain, where where it will become too expensive. Of course, humans can still do it because humans are not just trying every possible combination of things they know on every problem that they see. They just have this very efficient, very intuitive way of searching for for a theory of explains what they see.

Tim Scarfe 00:31:02

You know you know, my cohost Keith, he had this idea of doing a recursive version of of ARC. But the thought occurred to me is that even though we do this systematic compositional reasoning, we still have some kind of cognitive limit. So if we nested, let's say 4 levels of of our challenges with within the same problem, wouldn't you find very quickly that humans just can't solve it?

Francois Chollet 00:31:24

Right. So if you just like concatenate 2 arc tasks for instance, you get something that's much less enforceable, that's much harder because there are more rules going on. ARC-AGARC-AGIt's not quite what ARC-AGI would call compositional though. Because even though you have 2 rules at once, they're not interacting with each other. Right? They're they're they're you you can solve them separately and they can get in the solutions. And ARC-AGI think it's it's not it's not a bad idea at all. Like, it will work as a way to make ARC more difficult with, again, this caveat that you're not actually testing for depth of compositionality. 1 1 issue though is that it will only really work once because as soon as the person developing the AI system notices that the task can actually be decomposed into subtasks, then it's game over. You know? So ARC-AGI think it's actually more interesting to to have multiple rules at once, but they're actually being chained together or they're interacting together in some way. For instance, 1 rule might be writing some information on the grid that needs to be read by the second rule. Right?

Tim Scarfe 00:32:28

What peARC-AGIformance do the frontier models get on ARC v 2?

Francois Chollet 00:32:31

So what we saw was a big gap between models that don't do any kind of test time adaptation, like any kind of test time search or test time training and models that do. And the base LLMs, even models like GPT 4.5, they're basically scoring 0. ARC-AGARC-AGARC-AGARC-AGI think 1 of them, ARC-AGI think it was ARC-AGI 1 maybe, score like slightly above 0, something like 1%. But, you know, it's it's within within noise range of 0. So any model that cannot do a test annotation that is that does not possess fluid intelligence starts effectively 0. In that sense, arc 2 is actually a very strong sign that you have fluid intelligence, better than arc 1. I think arc 1 could already tell you that, but less perfectly. So in arc 1, if you do not do test neural adaptation, you can still do up to roughly 10%. Right? So on arc 2, that's actually 0. So it's a better test. Now when it comes to model that do test time adaptation, so we tried, for instance, some of the the top entries from the KALY competition last year in the models that we are doing test time training in particular or certain some kind of program search. And the best model, the model that actually won the current competition can do, I believe, 3% on on arc 2. And if you take an ensemble of the the top entries from the competition, you get to 4%. Right? So that's not very high. We also estimate that so so all 3 would be the the current state of the art in terms of an AI model that does exhibit fluid intelligence. And so we haven't been able to test o 3 on low compute settings on

all of the tasks that we wanted to test it on, but we've tested it on a subset and so we can extrapolate which to score on the on the on the on the on And it sounds like it's gonna be about 4%. Right? So not not not super high. There's a lot there. There's a lot of room to go higher than that. And we haven't been able to test all 3 on high compute settings at all. So, you know, the model that we're scoring 88% on on arc 1. So I can I can make a guess? I guess based on what we saw from o 3 look and and and other models, I think you might get up to like 15, maybe even 20% if you were really maxing out the compensating and spending like, you know, 10 k per task for instance. But that that we we we see be like far below average human performance, which would be more like 60%.

Tim Scarfe 00:34:59

So that 4% that o 3 gets on ARC v 2, do do you think of that as fluid intelligence or do you think of that as a potential gap in I mean, you could have designed ARC v 2 if you if you selected the correct sets of human calibrated challenges. You could have found a set which was still 0 1 0 3.

Francois Chollet 00:35:18

Yeah. Absolutely. No. You you you could have, you know, adversary selected against all 3 and then all 3 would do would do 0. It would be very very easy to go from 4% to 0%. Right? It's just a few few a few tiles that you need to change. So we not we not actually try to do that. Yes. I do believe that 4% does show that you have non 0 fluid intelligence, which is also something that you could you could get as a signal from arc 1. And I think the the the the sign that you see fluid intelligence in these models is the performance gap between the huge pretrained only models that don't do test time adaptation that score effectively 0, maybe 1, and you could say that that 1% is in fact a flat dataset. Sure. It it should in practice be 0. And the models that do test time adaptation and do non 0, that means 3%, 4%, maybe 5%. Right? And that means that there's something like 95% of the dataset that will actually give you these useful bandwidths for measuring how much fluid intelligence the model has, and that's something you were not getting with ARC 1. ARC 1 was more binary where if you don't have fluid intelligence, you're gonna do very, very low, like below 10% roughly. If you do, you're gonna score significantly higher and getting above 50% would be very easy. But because the the the the measure would saturate very quickly as

O3 Failure Modes and Reasoning Limitations

Francois Chollet 00:36:40

as soon as you start having non 0 fluid intelligence, you did not get that that useful bandwidth that you're getting with ARC 2. So I think ARC 2 should allow for, you know, answering the question, is this model actually as fluidly intelligent as the average human? Which is something you could not get out of ARC 1.

Tim Scarfe 00:36:59

I guess it's just an economics thing at this point. So if you spent, let's say, 1000000000 dollars or half $1,000,000,000, you could saturate arc v 2. I'm not sure if you would agree with that. But if if that isn't the case, I mean, do you think are the specific things that are missing from o 3 that that are stopping it from doing it doing better?

Francois Chollet 00:37:16

So it's never just an economics question because intelligence is not just about capabilities. It's also about the efficiency with which you acquire and deploy these capabilities. And sure, if you spend billions and billions of dollars, maybe you can saturate arc 2, but that would already have been true back in 2020 using like extremely crude brute force program search. If you have a DSL that's actually train complete, then you know that for every arc task, there exists a program that may not in fact be all that long that will solve the task, and all you need to do to find it is iterate over all possible programs in order of length. And then the first 1 you find is the 1 that's gonna that's gonna generalize. Right? Because it's it's the it's the shortest. It's the most the most parsimonious. So if you spend unlimited resources, you already have AGR in that sense just just in the pure skill sense. You can always just strive every possible program until you find 1 that works. But that's not what intelligence is. Intelligence is about finding that program in very few hops using actually very little compute. Like, look at the the amount of energy that a human expense to sort of 1 hour task over, you know, 2, 3, 4 minutes. It's it's it's almost 0. Right? And compare that to model like o3 high compute settings, for instance, which which is going to use like over $3,000 of compute. So it's never just an economics problem. Efficiency is actually the question we're asking. Efficiency is a problem statement. It's not capability.

Tim Scarfe 00:38:47

So intelligence is knowledge acquisition efficiency. O 3 did very well on arc v 1. And now that it does so badly on arc v 2, the whole point of of your definition of intelligence is that given some basis knowledge, you efficiently recombine, you produce new skill programs. You're saying that in the absence of the base knowledge in v 2, there is no intelligence. Therefore, is o 3 not actually as intelligent as we thought it was?

Francois Chollet 00:39:13

I think o 3 is 1 of the first models, perhaps, arguably in the first model that does show fluid intelligence. So now what the results on arc 2 are telling you is that it's not human level fluid intelligence. Right? But still, I would I would consider all 3 as a kind of proto AGI with with 2 2 big flaws, 2 big caveats. 1 is, of course, efficiency. Efficiency is part of the prime statements, in fact, the central point. So as long as you're not as efficient in terms of, for instance, you know, data efficiency, compute efficiency, energy efficiency, then it's only a temporary solution. We'll find a better solution in the future. And also, it's not it's not quite human level. If it were human level, you'd expect it to score, you know, something like Osrido should score like over 60% on arc 2. And we don't know what the exact number is gonna be, but, you know, probably probably around like 4 to 5%. Right?

Tim Scarfe 00:40:10

Do you think that general intelligence is a category or a spectrum?

Francois Chollet 00:40:15

So general fluid intelligence, it's I would say it's both because there's a huge difference between just having memorized a bunch of skill programs that are static and and and knowledge factors versus being able to adapt novelty to a nonzero extent. So that is a binary distinction, either you have fluid intelligence or you don't. Right? And and arc 1 could answer that question for any system. But once you have non 0 fluid intelligence, then the question is how much do you actually have and how how how it compares to to humans. And that's related to the notion of recombination of the ski programs that you have, the knowledge that you have, and depths of recombination. So if you do no recombination at all, you don't have fluid integers. If you do some recombination, you do. But then the question is how deeply can you recombine? Like, for instance, if using a program synthesis analogy, question is how how big of a program can you write on the fly to adapt to a new problem. Right? And and of of course, as well, how how efficiently how fast and how efficiently you can write it. Right? So it is a binary, but it's also a spectrum. And arc 1 was trying to ask the binary question, does this system has any fluid intelligence at all? And arc 2 is is more on the side of trying to measure how much re intelligence you actually have compared to humans.

Tim Scarfe 00:41:49

How long do you think it will take for V2 to be saturated? And do you think it will survive until V3 comes up?

Francois Chollet 00:41:55

So that's a question where you have to take into account resource efficiency. So if you're asking how long it will take before we have a system that can score higher than, let's say, 80% on ARC 2 using using less than $10,000 of compute, for instance, I think probably around a couple years. So it's very difficult to make predictions here. I think if you're just looking at current techniques and and scaling up current techniques, I think it could take a while. I think that arc 2 is actually way out of of reach of current techniques. But, of course, we are not limited to current techniques. You know? In 2025, we're probably gonna see new breakthroughs in the same way if we saw we saw new breakthroughs last year. And these breakthroughs are actually very difficult to predict. I was personally very surprised with the performance that all 3 could get on ARC 1 last year. That that that came as a surprise. So maybe we'll have new surprises this year, but I would be extremely surprised if we see an efficient solution that's human level on arc 2 by the 2025. I I I would basically that out. By the 2026, maybe. Right? Which is why we have we have arc 3 coming, of course.

Tim Scarfe 00:43:09

On analysis of failure modes, I'm I'm I'm sure you saw the blog post that I read where it went through all of the different failure modes of of o 3. And of course, it was solution space prediction which made it more surprising to me. My my take on it was I was really impressed that even when it failed, it was because the solution space got too big or it was just getting minor mistakes. But broadly, it it got the direction of of many of the problems quite well. You know, similarly, tell me about the the failure modes on v 2.

Francois Chollet 00:43:37

Right. So we were not able to test all 3 as much on v 2. But I can tell you about failure modes based on what we saw on on on v 1. And, well, there are many, but generally, this is a model where reasoning abilities can can decrease exponentially with this problem size. If you have more objects in the scene, if you have more more rules, more concepts interacting, you you see this exponential decrease in in capabilities. It's also you know, because it's a model that needs to it works by writing a kind of natural language program that describes what it's seeing, that describes the problem and the sequence of steps to solve it. So in that sense, it's in that person a natural language program. And that means that in order to solve a problem, that's to talk about it using words. And as a result, if you have a task where the rule is very simple to grok for a human but in a nonverbal way, but it's very difficult to put it into words, it has no verbal analogy, that's actually much harder to solve for for for this direction of thought model. So other than that, we saw that just 1 of the big challenges is, you know, compositionality, having multiple rules interact. There's also it seems there's a bit of a locality bias going on as well, where if you have to combine together bits of information that are spatially collocated together on the grid, that's easier for the model than if you have to do the exact same thing. But the 2 bits of information you have to to synthesize are are are pretty distant. So having to to, yeah, combine together bits of information that are that are separate, having to so it seems as well that the the model has a trouble simulating the execution of a rule and then reading the results. Like, for instance, if you're solving an art task and you a certain rule and then you start applying it, let's say it's like you're you're continuing using line line continuation or something. And then you have to take another rule and use that rule to read a bit of information that you have written in the process of executing the first rule. That sort of thing is completely out of reach for the chain of thought models.

Tim Scarfe 00:45:58

How multidimensional do you think intelligence is? You know, 1 school of thought, and I and I think you you might subscribe to this, is that the universe is kind of almost made up of platonic rules that that are disconnected from the the the world that that we live in, and then there's this kaleidoscope idea you talk about and they get combined together and and that's what we see. But another school of thought is that there'll always be another dimension of intelligence. We'll always need arc v 4, v 5, v 6, and there'll always be something missing.

Francois Chollet 00:46:27

Each step of generating that you cross, you you you're gaining a nonlinear amount of capabilities. Right? And so after a few steps, you are so overwhelmingly superhuman across every possible dimension

Intelligence Dimensions and the Road to AGI

Francois Chollet 00:46:40

that, yeah, yeah, you can you can say without a doubt that you you have you have a GI, in fact. In fact, you have super intelligence. But, yeah, intelligence is, in a sense, multidimensional. And what ARC is trying to capture is really just this fluid intelligence aspect, this ability to recombine core knowledge building blocks. So in my definition of intelligence, intelligence is about efficiently acquiring skills and and knowledge and recombining them to well, again, efficiently efficiently recombining them to adapt to novel tasks, to novel situations that you cannot prepare for explicitly. And purely the ability to take a bunch of building blocks and and recombining them, doing kind of program synthesis, that's 1 aspect of that. That's probably the most central aspect, which is why, you know, this is why we're we're focusing on with ARC. But it's not the only aspect because this is assuming that you already have this pile of knowledge available. So it's not it's overlooking the acquisition of of these abstractions. It's also overlooking the acquisition of information about the task. In ARC, you're provided all the information about the task at once. But in the real world, you have to collect that information. You have to take actions, set goals, to discover what what your environment is even about, what you can do within it. And you have to do these things efficiently, of course. And and that efficient efficiency aspect is very important because, you know, intelligence was developed by by evolutions in evolutionary adaptation. And when you're exploring the world, you are taking on some risk. You know, you might get killed by a by a predator, for instance. And so you want to be, being you want to gain the maximum amount of

information and thereby power over your environment by taking on minimum a amount of risk and expanding minimum amount of energy. That's not something that you can measure. That's not something that you can capture with ARC v 1 or v 2 alone.

Tim Scarfe 00:48:47

Can you just expand on the significance of the solution space prediction with with o 3? Because that rather suggests to me that it's almost this rich Sutton idea where it's nearly a blank slate and it's very empiricist and we just take the data in and the neural network does all of the things. I always imagined that we would need to have some kind of structured approach which took, you know, the core knowledge in into account. Do do you think that, you know, it's it's actually simpler than than we thought?[6]

Francois Chollet 00:49:15

Try trying to directly predict the output versus trying to write down the steps to get the output. They are not entirely separate things because of course, once you've written down the steps, you can do what looks like transaction. And o 3 is not actually a a real transaction model because it's a it's it's it's much closer to program synthesis model where the the it's searching for the right chain of thought to describe the task and list the sequence of steps to solve it. And once you have the chain of thought, you can just use the model to execute it and it gives you the output. So from the outside, if you treat the entire system as a black box, it looks like like transduction, but the same would be true of any program search system. What it's actually doing, and the reason why it's able to adapt novelty so well, it's it is because it's synthesizing this chain of thought, which serves as a recombination artifact for the knowledge and the skills that the model has. A recombination artifact is adapted to the partial task at hand. So it's much closer to a Bronze's model.

Tim Scarfe 00:50:23

This is something that the community found very confusing because in the last interview you you were describing, I think it was o 1 pro as being a kind of explicit search process. And Yeah. What seems to be the case is that it's, you know, there is some kind of reinforcement learning thing in in the pre training, and then it maybe does some sampling inference time. So we, you know, we're doing a whole bunch of completions. And are you saying it's as if it's doing a program search or are you saying it's somehow explicitly doing a a chain of thought, you know, program search?

[6]The Bitter Lesson — Paper

Francois Chollet 00:50:52

It's it's searching over the space of possible chain of thoughts and finding the 1 that seems most appropriate. So in that case, it's entirely analogous to a program search system where the program you're synthesizing is a natural language program. Right? A program written in English.

Tim Scarfe 00:51:08

Okay. It just seems a bit strange doing auto regression on a language model, how that could be characterized as a search process.

Francois Chollet 00:51:15

It's so a model like o 1 pro for instance or o 3 is not just autoregressive. It has actually this test time search step, which is why it can adapt to novelty much much better than the base models are purely autoregressive. And that's why, again, you see on on the so in general, ARC, even ARC 1, has completely resisted the pretraining the purely autoregressive pretraining scaling paradigm. Like, from 2019 to 2025, we scaled up these models by, like, 50,000 x, like from from GPT 2 to GPT 4.5. And even on arc 1, you went from 0% to something like 10%. And on on arc 2, you're you're going from 0% to 0%. Right? And meanwhile, if you have any system that's actually capable of doing test time adaptation, like test time search, like o 1 pro or or o 3, then you're getting you're getting much, much better performance. There's this huge performance gap. So in general, you can tell the difference between a model that does not do test time adaptation and the model that does by looking at this performance gap, this generalization gap on ARC, also by looking at latency and by looking at cost. So of course, a model that does test time search is gonna give you answer. It's gonna take much longer. Like if you look at o 1 pro for instance, it's taking 10 minutes to answer your queries and it's gonna cost you much more as well because of what this work is doing.

Tim Scarfe 00:52:39

So I could download the Deepsea car 1 model and I'm running it on my machine and as far as my machine is concerned, it's just a normal LLM. It's doing greedy sampling

Francois Chollet 00:52:48

That's right.

Mike Knoop 00:52:48

Auto regressors.

Francois Chollet 00:52:49

That's right. Which is why which is why it does not adapt to novelty and it scores basically 0 on arc or like 1% maybe.

Tim Scarfe 00:52:54

Oh, so so so you're saying there is something different about

Francois Chollet 00:52:57

o 3 is qualitatively different. That's that's correct. It is qualitatively different from all the other models that came before. It is actually a model that has fluid intelligence. It has non 0 amount of fluid intelligence. And and and r 1 for instance is not.

Tim Scarfe 00:53:12

Okay. So categorically it's doing some kind of active search process at inference.

Francois Chollet 00:53:16

That's what it looks like. So of course, I don't actually know how it works, but that's that's what I would speculate it looks like. Yes. And you see it in in the latency, in the cost, and of course, the arc performance.

Tim Scarfe 00:53:25

Would you be shocked and surprised if you if it came to light that it was just doing like auto regressive greedy sampling?[7]

Francois Chollet 00:53:32

Honestly, I think it's very very unlikely because it's completely incompatible with the the characteristics of the system that we know of, that that we were exposed to when we tested those 3.

Tim Scarfe 00:53:41

Awesome. And do you think that there will always be human gaps?

[7] Decoding Strategies in Neural Text Generation — Paper

Francois Chollet 00:53:46

Probably not always. Today, they are very clear, very significant gaps. Right? Like, we are not we're not actually that close to a GI right now. But eventually, you know, as we get closer and closer, there will be few additional gaps. And at some point, you know, we're we're gonna have AI system that are just overwhelmingly superhuman along every possible axis you choose to look at. So, you know, I don't think that there will be gaps forever.

Mike Knoop 00:54:11

Alright, Tim. Thank you so much for doing this. I really appreciate it. Looking forward to seeing you in a couple weeks.

References

[1] **On the Measure of Intelligence**
https://arxiv.org/abs/1911.01547
Paper

[2] **ARC Prize Foundation**
https://arcprize.org/
Organization

[3] **OpenAI o3 Performance on ARC v1**
https://arcprize.org/blog/oai-o3-pub-breakthrough
Paper

[4] **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**
https://arxiv.org/abs/2201.11903
Paper

[5] **ARC Prize 2024: Technical Report**
https://arxiv.org/abs/2412.04604
Paper

[6] **The Bitter Lesson**
http://www.incompleteideas.net/IncIdeas/BitterLesson.html
Paper

[7] **Decoding Strategies in Neural Text Generation**
https://www.mdpi.com/2078-2489/12/9/355/pdf
Paper