Brian Gracely (00:01.774)
Good morning, good evening, wherever you are, and welcome back to the Cloudcast. We are coming to you live from the massive Cloudcast studios here in Raleigh, North Carolina. Hope everybody is doing well. We are going to try something just a little bit different this week, and probably something we'll try and continue to do going forward. I talked a few weeks ago about moving the Sunday Perspective Show to Saturday for a couple of reasons. One, it keeps me from having to think about the show all weekend. I can sort of knock it out during the week, so selfishly for me. But more importantly,

I think we were finding that sometimes you think about how often people are consuming your show, how often they're having to do a lot of things for a lot of shows. And we kind of felt like the Sunday and Wednesday was maybe a little too tight at the beginning of the week. If we put them out on Saturday, it gave people a little more time to consume it. If they were doing it on Saturday, doing sort of with hobbies, kind of running, mowing the grass, those sort of things. Sunday, a lot of times, is family time. And we didn't want to necessarily, you know.

conflict with those types of things. So I think we're gonna kind of move this to Saturday, call it Saturday perspective. First and foremost, putting this one out a day early for those that will be hopefully celebrating Mother's Day here in the United States. I know Mother's Day I learned a little bit this week. Mother's Day is a little different in different parts of the world. But I'm going to just pick one day which happens to be US Mother's Day. So first and foremost, happy Mother's Day to all the mothers out there. Happy Mother's Day to my mom. I know she periodically will listen to the show and tell me things like I have no idea what you were talking about, but listened in.

But you know, hopefully everybody is doing this especially on Saturday gives you some time if you have not yet gotten a Mother's Day gift or planned things with your kids for Mother's Day Haven't had a chance to schedule some time to call your mom reach out to your mom via You know FaceTime or some other way so, you know Hopefully as you're listening to this it gives you a day heads up if you if you're busy and it may have slipped off your calendar but make sure to call mom make sure to you know, take care of

Kind of all the all the women in your life not that you shouldn't be doing that on any given day But you know reach out to moms reach out to grandmothers reach out to wives and try to be a good role model for for the women in your life whether those are young kids or just your colleagues and co -workers so This week I want to talk about something that is I try and avoid as much as possible sort of Bringing things into the podcast that are sort of directly related to to my specific work just if for no other reason then I

Brian Gracely (02:25.134)
There's plenty of chances I have to talk about my own work during the week. And we try and keep the show a little more community focused and just separation of work and non -work. But I wanted to bring something this week. I know Aaron and I talk quite a bit about AI. It's sort of the talk of the industry these days. And there was something related to my work that happened this

last week at Red Hat Summit that I thought would be interesting to bring up to the broader community, because it not from a Red Hat perspective specifically, but just.

from a community perspective. And so I'm going to try and dive this week a little bit into some things that are happening that I think might be interesting to a lot of people, especially anybody who is interested in AI, who's interested in trying to figure out how do we bring some open source principles to AI? We'll talk about that a little bit. As well as how do we, as we're looking at AI for your business, for certain aspects of your business, how can we bring some knowledge and skills and some customization capabilities?

to your business in a way that is cost effective and also preserve some of the security and some things that you specifically do with your business. So we're going to get into that right after the break.

Brian Gracely (00:02.638)
And we're back. Welcome back to the Cloudcast. I am your host Brian Grace Lee. And as mentioned at the top of the show, first and foremost, Happy Mother's Day. Make sure to remind your reminder to to reach out to mom, let mom know that she's important to you that you're thinking about her and you know, just thank her for thank her for bringing into bringing you into the world and hopefully, you know, your your relationship with mom is good or an opportunity to make that better. So what I want to talk about this week, and as I mentioned, you know, both Aaron and I try very, very hard to not

Use the show as a vehicle for the things that we do in our day job We've got plenty of opportunities to do that but there are some times when you know the things that we're working on we think might be applicable to a broader audience and You know we try again like we try and separate those things but you know this week I thought you know this might be an interesting thing to kind of bring to the forefront because Aaron and I talked quite a bit about AI topics obviously a lot of people right now are talking about that as a as an important thing or an exploratory thing both for their business or

for what they're working on. How can we take advantage of AI technologies? Is it appropriate for us? How can we go about doing that? And one of the challenges with AI has been kind of the central question for a lot of people, especially as they're exploring things and they're dealing with the fact that it feels like every day, every week, every month, there are new advances in AI, new models that are coming out. And people are trying to figure out, one central question that we've talked about a lot is will the...

things in the future world, will your consumer experiences, your business experiences be sort of dominated by one gigantic single model? Maybe that's a model that's coming from open AI or Facebook or Google or something along those lines, one sort of giant model to rule them all that all applications are built on top of, or will we live in a world in which there is a pretty good diversity of models? And the trade -offs there, the reason that conversation tends to come up is, is a lot of different reasons. First and foremost is control, because as with all things in technology, we do see certain large companies or large entities, but we also see a lot of competition. And where there's competition, we often see interesting new innovation happen. So there's always an aspect of kind of control versus competition. There is a consideration for how much should your data influence a model.

Brian Gracely (02:26.606)
And how much should your data be in your own control? Right. And so, you know, as we talk about some of the gigantic models, yes, they have the ability to go out and basically, you know, look at everything that's in the public internet. And they're doing all sorts of interesting things about, you know, business relationships with private data and so forth. But, you know, for a lot of companies, you know, their secret sauce, their data, their information, they want to keep fairly private and that that provides differentiation for them. It's something that

They're able to build new services upon. And so there's that question of, in the future, from a business perspective, how much of your business will be impacted or used within models to try and create new business services, create differentiation? And how much of that will you have to kind of do yourself versus being able to use one of the gigantic models that are out there? There's questions about.

just basic things that come into play in the enterprise that don't always get news coverage about things like, well, what happens if our service can't be public to the internet? So there's lots of things that are regulated, things that live on premises, things that are behind firewalls or are sort of air gapped from the internet, whether these are government agencies or military organizations or, again, regulated services and things like that. So those types of things come into play. So there's a lot of questions about these sort of super gigantic Uber model.

that's controlled kind of by one company versus, you know, a diversity of models or a lot of models that might happen. There's another question of, you know, there's obviously been a lot of conversation recently about open source, but, you know, is there a role for open source in, in AI? And where this usually comes into play is we've seen a number of models that have been released with a open license or with the word open associated with them. But, you know, there's some debate about,

what do those licenses mean? So in the case of, let's say, something like Llama 2 or the Facebook model Llama, some of them are released with things that have openness but then restrictions. So you can go take a look at those models. But for example, one of the models was released with, they were open for public consumption, but only if they were used in smaller set services. There have been some.

Brian Gracely (04:46.638)
models that have been launched with Apache 2 licensing, so more open licensing, but not necessarily the data being necessarily open. The sources of data, the siting of data may not

necessarily be open. And then the last piece of it is, if a model is out there and it's considered to be open, how does contribution happen for those models? And that's a conversation that, for the most part, what has to happen today is if a model is released as open,

Somebody can go fork that model, you know sort of pull a copy of it if you will Do whatever sort of retraining tuning alignment they want to do But there's not necessarily an open contribution model in which your contributions or your changes to that model would then sort of be Rebased back upstream into the main model, right? So, you know if you if you try and draw a perfect comparison a perfect parallel between how open source software Tends to work today in which you know, the software is open and

Contributions are done in the open. You can look at contributors and all those sort of things. We don't have an exact model of that in the world of AI as of today. So the great thing about 2023 and early 2024 from an AI perspective is we have seen more openness happen. And again, how much open matters to companies will be determined. The same way that how much open source and open standards matter to companies will be determined.

There will always be in the technology world proprietary offerings completely closed offerings and then offerings that have various levels of flexibility Beyond that proprietariness whether it's you know free whether it's open source whether it's open core all those things are gonna be out there and so What it was interesting this this past week is? Is red hat and IBM research? Came out with some a couple of things a couple of announcements, but basically what they were going after is they said

you know, essentially Red Hat saying, you know, there should be some sort of parallel between what happens in open source communities today in terms of open contributions, really knowing where the source code came from. And in this case, you know, availability of where the data came from, and then trying to create a contribution model for, for LLMs that looks like what you're used to with open source in which you don't have to fork the main thing. You can just make contributions to

Brian Gracely (07:12.59)
kind of the main project, if you will. So with those things in mind, and then one other thing in mind that sort of came from this project, and again, I'll get into the details of it, was to sort of look at four or five areas in which the cost of AI today is fairly expensive or somewhat prohibited for certain things. So if we think about the life cycle of AI, typically it's going to begin with data scientists taking a whole bunch of data, so data collection.

But data scientists working on very very large sets of data. So data scientists today wonderful people do fantastic work Not necessarily the easiest people in the world to find us so the easy people easy people in the world to hire and The cost of them is fairly significant. So because the work they do is fairly complicated. So You've got you've got sort of cost of data scientists. You've got cost of data collection, right? You know, you need to be able to collect enough data to be able to build a model. The third thing you have is the cost of training that model. Typically, to do these large models, and we're starting to see models with trillions of parameters, billions of permutations within the models, those are fairly expensive. Built on hundreds and hundreds of GPUs over many months, we've seen numbers like.

Training this model or that model is costing tens 20s 30 hundreds of billions of millions of dollars, right? So so the cost of training the cost of acquiring GPUs from a hardware perspective and then the cost of Tuning and alignment on an ongoing basis as you're as you're getting more data into the model You know can be timely and costly. So there's a whole bunch of aspects to to AI that You know are expensive today you know, I know there's a number of

groups and organizations that are looking at how can you begin to drive down those costs to make it more applicable. And we've seen this with things like small language models. We're seeing different types of acceleration hardware that are coming out there. Obviously, we've seen people like Intel and Nvidia are trying to make sure that, I'm sorry, Intel and AMD are trying to make sure that Nvidia doesn't have a complete monopoly on that. And right now, Nvidia has a pretty good moat around GPUs. But.

Brian Gracely (09:29.71)
you know looking at just a lot of different ways in which the overall costs of AI from a life cycle perspective can be brought down and again the reason for that is you know in order for any technology to be widely used the cost of it can't be so prohibitive that nobody receives any value for it or the cost of entry is is so high the first step is so high that is prohibitive for you know too many people to be able to take advantage of it.

So anyways, the thing that Red Hat and IBM research announced this past week was really two things. The first was they took a number of models, both large language models, in this case initially English language models, as well as a number of code models, so LLMs for code, that are called the granite models. And I'll put some links in the show notes. These were models that have been trained over a number of years. They are used by Watson X.

within that service as well. But they open source a number of models. And so first and foremost, what's I guess interesting about these is they're not the largest models in the world. They're not necessarily Lama 3 size models or Gemini size models, but reasonable size models. The second thing is they are released not only under Apache 2 license in terms of being open, but one of the first, if not the first, license model to be

open in terms of weights and biases and where the data, the siting of where data sources came from. So open in the way that you would think about open source projects, right? In terms of sort of everything is open. So that's the first thing is they released a number of open models with sort of full openness in terms of the date, the weights, the biases, how the model was built in essence. And the second thing they did was they released something called InstructLab.

And InstructLab is a really sort of interesting concept. So what InstructLab does is it says, its premise is there needs to be a way in which individuals who maybe aren't necessarily data scientists can contribute to a model. And so what they've done is they've created something, and I put a link to the research paper if you want to kind of dive into it in details, as well as the GitHub homepage that will track this and so forth.

Brian Gracely (11:50.702)
But in essence, what InstructLab does is it says, let me provide a more simplified user interface to be able to contribute to a model. The framing of that or the framework for doing that is a taxonomy that is based on skills and knowledge. So the things that you're going to contribute to a model, you're going to start with some sort of base model. That base model is going to be fairly capable from an LLM perspective, if you wanted to build a chat bot or interact with it through language and so forth.

But the thinking is, there are going to be things, especially around businesses, that the skills and knowledge that are specific to your business aren't necessarily embedded in the model. And while there have been some ways to augment models in the past, so if you've listened to the show a little bit or you're following AI, there are things called RAG, or Retrieval Augmented Generation, which was a way of saying, I have a model. That model doesn't necessarily know about my specific data.

And I want to be able to do that. And so what happens is when you make a request to the model, the application that sort of fronts that model says, OK, I looked at the model, didn't necessarily have the information that I wanted. So I'm going to use this rag method, this retrieval augmented generation, which means I'm going to essentially query an external vector database, check in that database, see if there's a way to augment the query or the part of the thing that you asked with.

specific data that has been put into this database and then mash that together with maybe the output that would have come from the LLM. So this isn't a unique new concept that businesses haven't been able to augment their LLMs. But what the InstructLab model does is it says, let's provide a simpler interface, sort of a YAML -based interface that allows skills and knowledge that can be specific to a business.

to be brought to the table, augment the model. So that's the first thing is the ability to bring skills and knowledge that are specific and may not necessarily be present in the model to the model. That's the first thing. So the second thing that happens within InstructLab that's interesting and begins to address some of the cost aspects of it is in the past, in order to train a model, to tune a model, to align a model to your specific things you wanted to do,

Brian Gracely (14:15.598)
you typically had to bring a lot of data and orders of magnitude, a large amount of data that's specific to what you're trying to do. What InstructLab does in that second phase, there's four phases of what happens. First phase is taxonomy driven. It's you inputting skills and knowledge.

And there's links in the thing about how you go about doing that. But the input is a fairly small amount of knowledge and skills.

to kind of seed the process of bringing that skills and knowledge to the model. The second step is synthetic data generation. So what the InstructLab then does is it takes your number of inputs to create skills and knowledge. It then uses synthetic data generation to create essentially thousands of like skills and knowledge inputs.

to what you did, right? So instead of having to physically go out and collect all that data, have to have somebody, you know, sort of data science type of skills, data cleaning, go about dealing with that, you're able to synthetically generate data from that input, small input, turn that into large amounts of input. So you're able to sort of bring a lot of data to a model. That's the second phase. That becomes what's called sort of a teacher phase. The third phase of it is,

you've generated a lot of new data, but there's a chance that, as we've seen with things like hallucinations and other things, that some of that data that got generated, synthetically generated, might be bad, misleading, incorrect information. So the third stage of it is there's essentially a critic model that is used to validate the information that's come along. And once that third phase has happened and you've eliminated the bad or misaligned or incorrect values,

then it moves on to doing the actual training of the new model. So four steps, input through the taxonomy, skills and knowledge, second phase, synthetic data generation, third phase, critic model, getting rid of the bad stuff, and then the fourth phase is the training. So when I bring this up, and I mentioned earlier that there's a lot of stages to AI that can be costly, the way to think about those is the first stage in which,

Brian Gracely (16:39.982)
you needed only data scientists could kind of contribute. This is now sort of unlocks this for really kind of anybody, right? So myself, Aaron, you listening at home, you could contribute to a model. Your business could contribute to a model. Your group that you work with could contribute to a model without necessarily having to have data scientists. The second thing is that synthetic data generation, because it's a fairly low cost task to do,

Again reduces some of the cost that you would typically have in in training or tuning a model, right? You can do a lot of this work on a laptop on a high -powered sort of single server with a few number of GPUs So it addresses some of the cost of being able to not necessarily train the entire model But fine -tune and align the model so bring down those costs the third piece of it is you know since the models originate as a smaller model fairly smaller model and

the synthetic data generation and then the sort of getting rid of the bad things keeps the model still fairly small, but tuned to the things that are specific to you. You're not necessarily dealing with gigantic models. So again, once the training phase comes along, you're not having to do nearly as much training with nearly as big a footprint. So again, addressing some of those training costs and so forth. And then finally, you end up with models that you can...

either take the skills that you have contributed and bring them back to the broader community. So if you're interested in sort of being part of the broader community around these granite models, you can go about doing that, right? So there may be things that you feel you want to contribute. But also, you know, you can keep those things to yourself, right? You can keep the things that you've contributed to yourself as private, right? So you have complete control over what you contribute.

Does it go back into the community? Maybe as a foundational thing that you think would be helpful to grow the community or does that stay within sort of your company's own domain? Right. So the other last thing that's sort of interesting is the instruct lab tooling While it it it was announced in conjunction with the granite models Can also work on other models. So I'll put some links in the show notes. You can see some examples of it training

Brian Gracely (19:01.55)
Doing this sort of lab process on like Mistral models and llama models as well So it's not confined entirely to the granite models. Although the granite models are sort of the first ones to be fully open But you can sort of make these contributions to a say forked version of your models as well if you're building something specific with llama or Mistral 2 or something else so You know, I thought it was sort of interesting just in terms of it's another step in

you know, where open meets AI and now open source and some of the things that a lot of us think about in terms of open source, how that process works, but also allows, you know, kind of anybody to contribute to that model. And it does begin to open up some very interesting possibilities that might be specific to your business, maybe how your business goes to market. Maybe you're going to start to embed an LLM or some generative AI into your own products that you build or services that you deliver.

It might just be relevant to your organization if you are building some capabilities that will always be internal and you're not trying to share with the rest of the world. Or if you are looking to participate in broader communities, this is a starting point for being able to participate in broader communities. So anyways, all this stuff had nothing to do with Mother's Day. We just sort of had that intersection of it. And again, I tried to keep this very focused on sort of the open and community capabilities as opposed to.

sort of the specific things that Red Hat or IBM might be doing from a commercial perspective somewhere down the road. So anyways, with that, again, happy Mother's Day to all the mothers out there, anybody who is needing to talk to their mom, talk to their wives, talk to their grandmothers. But anyways, hopefully this gives you a little bit of perspective on maybe a next step in where open source meets AI and some of the things that might be possible with that. And,

If you're interested, I put some links in the show notes in terms of how you can take a look at what's going on, take a look at the research paper, take a look at the open source projects, take a look at the open models, and see if they make any sense for you. And I would love to hear your feedback on those things. I know there will be a lot of things from a community perspective that will be happening, again, above and beyond what Red Hat and IBM are doing to participate from an open perspective. I would expect to see just a number of people participate as an open source project in an open source community. So.

Brian Gracely (21:23.694)
That anyways first Saturday perspective, I guess probably be moving to these to Saturday unless there are some big travel things that prevent me from doing Saturdays. But anyways, hope you're all are doing well. Hope you're enjoying May as we get into sort of the middle of May here fairly soon and hopefully the weather's enjoyable for you are you are hopefully for those of you who've got kids that are graduating from college and things like that. Congratulations to you. Congratulations to your kids with that. I'm going to wrap it up. Thank you all for listening. Thanks for telling a friend. Thanks for.

giving us feedback on the show and in all the places you do it, whether it's, you know, via your podcast players, whether it's via YouTube. But with that, I'll wrap it up and we'll talk to you next week.