Ruben Wolff - **00:00**

Right, more small talk. So who's gone skiing this winter?

Yufan Jiang - **00:04**

I haven't.

Robert Cowlishaw - **00:06**

Yep, I have. Have you?

Jason Orender - **00:09**

I have not.

Ruben Wolff - **00:10**

I still haven't gone skiing but I'm going to go ski touring o snowshoeing on Saturday in Colorado, which I'm pretty excited about.

Robert Cowlishaw - **00:25**

I was in Spard yesterday. Snow up there was good.

Ruben Wolff - **00:32**

Sounds like an island.

Robert Cowlishaw - **00:34**

Yeah, it's like north of Norway.

Ruben Wolff - **00:40**

Oh man. You were there yesterday?

Robert Cowlishaw - **00:43**

Yeah.

Ruben Wolff - **00:44**

That's crazy. Yeah, actually a friend of mine has done ski touring there. That's pretty awesome. On the map it looks super far away, but I guess that's a skew. How long is that flight?

Robert Cowlishaw - **01:00**

It's like two hours flight north of the top of Norway. So like there's 78 degrees latitude so it's a decent way up.

Ruben Wolff - **01:18**

You flew from Scotland, right?

Robert Cowlishaw - **01:23**

Yes, via Tromso. So the flight to Tromso was three and a half hours and then from Tromso to Svalbard.

Ruben Wolff - **01:46**

All right. Definitely gotta wait for asthma to still get here. We were just on a call it. All right. All right, I guess that's enough waiting. I think ASMA will still join so. And he knows anyways the introduction. So yeah, thanks for joining everybody. We are having our first meeting of the Consortium on Robust Generalization Errors. Who knows what the name will be at the end of this meeting. I have a couple note takers in here. Actually I should click record. Okay, never mind. Not allowed to record. So it'll just be an OK cursing audio. The goal of the meeting is relatively open. For me what's important is figuring out how we want to organize ourselves around the problem.

Ruben Wolff - **03:38**

Figuring out who is like where does each person feel like they can do the most contribution, be it in the direction of development like practical contribution theory organizational. And yeah, just get into some of the design decisions that I think are pretty exciting to talk about. Probably do all those. So maybe we'll do intros and try to keep it to like under 10 minutes because there are quite a few people that showed up. So maybe we'll do like each person just does like one sentence about yourself and maybe the like which direction you feel like you would be like where your strengths are in practical like coding, development theory, academic organizational. So I'll start. You guys all already know me, of course. Ruben, German American. I came from computer science, University of Texas ETH Zurich.

Ruben Wolff - **04:38**

AI consulting all the way and then a number of startups. And yeah, this problem has been bothering me for quite some time because I think people need to like look at real data that's not corruptible. And then fund startups based on that. And I think my biggest contribution will probably be through organizational and political stuff because as I am leading this meeting, I think so that's me. Let's just go down the line. Jason, do you want to do one quick intro yourself?

Jason Orender - **05:16**

Sure. My name is Jason Orger. I am a PhD student in Computer science. My specialization is high performance computing and I also dabble in crypto. My, my job right now is as the lead data scientist for AIML at a defense contractor and I am actually a former naval officer that retired and then got my master's in computer science. And now I'm really close to graduating with my Ph.D. And that's it.

Ruben Wolff - **06:09**

All right, sweet. Which, where do you see yourself like you're most interested in development theory? Organizational.

Jason Orender - **06:19**

I could probably contribute both to development and theory. I would say development is probably my, the greater of those two strengths because as I said, I only have dabbled in crypto up until now. I do have one paper out there, but it's. I, I think I would need a little time to get up to speed on the latest crypto.

Ruben Wolff - **07:00**

Yeah. All right, cool. Maybe I'm going to do a quick interjection because before we do more to interject again the, like the goal where we've talked individually obviously and half the people are actually most people here have both AI and blockchain experience. And one of the topics that we're going to go through is challenge my assumption that the way in which we need to solve this problem is like through decentralization, but not with blockchain because I think too many people in AI are afraid of blockchain. So.

Jason Orender - **07:40**

Yeah, right.

Ruben Wolff - **07:41**

The, the core problem that we want to solve is we want to solve machine learning projects, AI projects getting raided by public test data sets that they can very easily cheat on and bring that to something that is very difficult to corrupt.

And we'll get into like everyone's stories. ASMA has a lot of stories. I'm sure everyone has had their stories. But yeah, let's do more intros. So Yu Fang, you want to give quick intro of yourself? We gotta go fast though.

Yufan Jiang - **08:13**

Yeah, sure. So, hi guys, my name is Yifan and I'm now doing my PhD students here in Germany at the Kassel Institute of Technology. And my research topics actually not related to blockchain but to the applied crypto stuff. So were doing actually security notions definitions basically using the universally composable framework. And I'm also designing the multiparty computation framework for the privacy preserving machine learning tasks. So that's like my main or my main focus on my research topics.

Ruben Wolff - **08:56**

All right, sweet, let's move straight on. Robert, nice to see you back. Quick intro on yourself.

Robert Cowlishaw - **09:04**

Hi, I'm Robert. I'm based in Glasgow in Scotland. And as you can see from the screen, I'm a PhD student as well. Originally I was in aerospace engineering, but now I'm looking at satellite emergency mapping and how we can improve that with Web3 technologies and distributed technologies. And I guess the thing that maybe I would be open to look for working on in this project would be on theory side and maybe a little bit the development side as well.

Ruben Wolff - **09:36**

Sweet. Thanks, Robert. Stella, you're here, right?

Stella Wohnig - **09:40**

Yes. Hey, I'm Stella. I'm also a PhD student. I work in more theoretical foundational crypto and yeah, I think generally finding some models. I think what Yufen said also encapsulates what I do and finding some provable security to put this all on a solid foundation would be more my strength.

Ruben Wolff - **10:02**

Sweet. Saif, can you introduce yourself?

Saif Mir - **10:07**

It's nice to meet you all. It's Saf. It's Saf. No worries. But I'm a student here at Ohio State University. I'm studying computer science and engineering. My focus and my research was more in blockchain security. What were doing was trying to moderate hateful content and smart contracts. I know you just talked about not delving too deep into blockchain, but I have my skills mostly in peer review of paper and development in Python, so I think that's where I can contribute. Thank you.

Ruben Wolff - **10:34**

Yeah, well, who knows? I mean, if the majority of people think it should be done on blockchain, maybe I'm just wrong. So we'll see where the conversation goes. Siri, are you here?

Zerui Cheng - **10:45**

Yeah, I'm here. Hi everyone, I'm Siri and I'm a second year PhD student at Princeton University. My research is mostly focused on the intersection of blockchain and AI, namely how to bring AI to blockchains. And I think I can mostly help with theory side and research side. Yeah, thank you.

Ruben Wolff - **11:04**

Sweet. All right, who else is here? Who have I not introduced? Asme, can you give a quick entry yourself?

Asmelash Teka Hadgu - **11:14**

Yeah. Hi, nice to meet you all. My name is Asmale Ashteka Hatgu. I'm the co founder and CTO of Lasan AI. So at Lisan, we're building language technologies for African languages with our first product, Machine translation for Ethiopian and Eritrean languages. We're just launching an automatic speech recognition system. This problem of creating evaluation data sets that are up to date, distributed and maybe close to the real world is really close to heart because hopefully we'll discuss some of these things in detail later. But it would kind of like put everyone on equal footing whether this is like big tech companies, smaller startups or researchers in academia or elsewhere.

Ruben Wolff - **12:03**

So.

Asmelash Teka Hadgu - **12:03**

So yeah, fascinating kind of project. I'm more interested in the development aspect just to see a proof of concept of what this solution could look like. But I would also be happy to contribute to the paper.

Ruben Wolff - **12:20**

Sweet. All right, who else did I not introduce? Ramachandra, you there? Hey. Hi.

Ramachandran K - **12:34**

This is. So I'm working as a lead data scientist and idea strategic companies. Basically my area strength is to provide consultations on the A and ML solutions. Like I have given A big companies like Siemens and Facebook and other things. So I'm more interested on development side. So the similar use cases have been solved before the Facebook as well.

Ruben Wolff - **13:01**

Yeah, yeah. Thanks for coming on the call. This is actually our first call but we got connected because of your contribution on Dina Bench, which is super related. I had some other calls with other people from ML Commons. We'll go into the discussion. But yeah, really happy to see you're here.

Ramachandran K - **13:19**

Yeah.

Ruben Wolff - **13:20**

Did we forget anyone? I think that's everybody. Alrighty, so let's get back into like what the problem is. So from my perspective, you guys have heard it, but let's just repeat it again. The. The way the AI industry right now is evaluating test data benchmarks like Deep Seek came out, everyone freaks out. The entire stock market crashes because of numbers that cannot be verified. You can take Deep SEQ and you can check if that is the accuracy they get, but you can't check whether or not they have not included in their training these test results because you know the benchmark results, you know the true outcome that Human Eval and imagenet and all the other open benchmark test data sets have. And so I just don't believe them.

Ruben Wolff - **14:13**

And I want something believable and I also want something that like the world should orient itself around something that can't be corrupted like so easily as these. And let's remember that the test benchmark data sets at the beginning of machine learning they were super helpful for everyone to have the same open test data set because we could reproduce it. We can compare different machine learning algorithms for training and see which one's the best and then everyone repeats it. But now that we're in the age of deep learning where you can't reproduce the training. So you can't check that they didn't include the test data. This is really the wrong way of evaluating models and yeah, we need to change it. What we're building likely will be more difficult, but I mean there's enough money in AI to fix this. So.

Ruben Wolff - **15:06**

Yeah, that's my perspective on it. And just open for anyone to chime in, how do you guys see the problem of AI evaluation?

Jason Orender - **15:24**

Hi.

Saif Mir - **15:24**

Basically what you're saying is people are just throwing numbers out there that aren't being evaluated correctly.

Zerui Cheng - **15:29**

Yeah.

Saif Mir - **15:29**

And the purpose of this is to achieve proper evaluation rather than just attributing false numbers to things we don't know.

Ruben Wolff - **15:37**

Yeah, we just don't know if people are cheating. I'm convinced that everyone is cheating. I'm convinced. And we have, there's existing cases where it's come out for the public data sets like human eval. It's very like, it wouldn't be surprising to anyone, but there are some people who put effort into, like ARC AGI, they put effort into having a private data set and testing people. But then scandals come out showing that OpenAI somehow corrupted them and got access to that, to the results so that they could fine tune it. Yeah. So it's definitely a problem.

Jason Orender - **16:16**

Yeah, there's probably even cases where they've cheated and not known, you know, just scraping the Internet for data, you're bound to get some of the questions and answers.

Ruben Wolff - **16:34**

Yeah, that's definitely true. For large language models. It's very true. And I mean, I suppose also for language models, for translation and those things, you have a similar problem that things just get sucked in.

Asmelash Teka Hadgu - **16:50**

Yeah. If I may add, this is definitely true and we actually have some evidence. So for example, I'll tell you a story of how Facebook, I guess somebody said like somebody's from Facebook. So if you haven't heard this, you can take it from me. But basically Meta released a model, a translation model called the no Language Left behind, the NLLB model. And in it they showed for the first time using a single model, they can translate between 200 languages and so on. Which is a great kind of like engineering effort, to be honest. But my concern was really on the evaluation, many of the language they covered were languages from Africa, like language we've been working on.

Asmelash Teka Hadgu - **17:34**

And what they did is basically we saw instances like Lisan, for example, developed an open benchmark data set for researchers for anyone to build their models and check their results on. But they kind of put it as a training data and they don't mention this in their paper. They don't mention it anywhere. But if you dig on their GitHub repo you would find that some of the data sets, as Jason mentioned, because they just simply say like we scrape the web, you see like pointers, links to this kind of data sets, these benchmarks really included in them. So I think there's no doubt that many of this large language models or other like big single models, whether it's translation models or speech recognition models, are including this benchmark data set. So it's a real problem.

Asmelash Teka Hadgu - **18:29**

So I just wanted to you know, give one instance of a concrete example of where it happened.

Stella Wohnig - **18:36**

So I guess like one of the main problems, I'm not a machine learning expert, but I guess one of the main problems is like how to come up reliably with fresh benchmarking material every time now and again. Right.

Ruben Wolff - **18:51**

Sure it's an expensive thing to do, but again we have money out there. I think Duna Bench had a particular like trying to address that problem. I think the problem we want to address is the corruptability of the data generation. There are huge companies producing data sets. The AI industry is big. So scale AI is one of them. They have 5,000 employees. So in certain areas it's an issue to get the right kind of data to test on. And Dina Bench had an interesting approach where people like are actively trying to find the data that the model is bad at. But yeah, I think. And they do pretty well with that. I think we. I'm going to say that will be solved somehow mostly with money. And the part that like Dunabench doesn't so much solve is.

Ruben Wolff - **19:47**

Or at least it's not their goal is preventing corruptibility of. Yeah. So that people that are pushing for scores can't corrupt them or cheat.

Robert Cowlishaw - **20:00**

I think maybe at the moment as well people are start. Or what I've seen is people have started to sort of create bodies for test data. So almost big organizations that are potentially quite famous in the world being like we have a closed test set and we won't share it with anyone and you can bring your model to us and we can test it for you and we can give it a score. But at the moment that's very obviously centralized because we have to trust this big organization. And a lot of the time they're like somewhat international, not international, sorry, national organizations. And then you have the big thing of why would my country test your international organization and why would his country test against Yours. So you've got like a weird.

Robert Cowlishaw - **20:53**

Everyone's got their own test sets that are hidden and no one's trusting it, basically. So I think if you can bring everyone together in some sort of blockchain, decentralized system.

Ruben Wolff - **21:08**

Yeah, maybe we would actually do it.

Robert Cowlishaw - **21:12**

That could bring all these things together and actually create like an internationally trusted system for testing these data sets, which I think is a big thing that's missing at the moment.

Ruben Wolff - **21:22**

Yeah, I like that you have the perspective where nationalities are involved. So Robert, working on Space DAO has to deal with the US Space Force because US Space Force controls what American space satellite companies are allowed to do. And let me. Yeah, I guess other examples of companies that do private testing, ARC AGI is one of them. Seal, I think, is the. So SEAL by Scale AI, I think is the best that I found. And I do think that keeping the test data set private is part of the solution. Because if it's private, you can't retrain on it, you can't test on it. But indeed, this involves trust. So you have to trust Scale AI that they don't make a deal with one of the companies to get their scores up.

Ruben Wolff - **22:13**

Not only that, with seal, they don't publish the test results, which of course makes it easier for them because it allows. It means that they can reuse test data. But if, like, one of the companies can figure out who. Which of the prompts are coming from whom, they can also learn that test data. So I guess this goes to one of the big design decisions. Like, is it. Is it important to, like, I think we should kind of have all of it. We should have. There's private data, and that data then must get published later. And that data does not get reused. And you don't only have one entity making the data, but multiple, such that if you have seal, hopefully we can convince Scale to participate, then they are one of the people providing results. But there's n others.

Ruben Wolff - **23:12**

And if your model has a much better score with seal, it'll be visible. Most likely you corrupted seal and then, you know, the whole point of corruption would be pointless because people would see it. So, yeah, private data, I think is a part of it. Getting it all together, I think is second part. So I agree with you.

Stella Wohnig - **23:41**

Okay. Another question is, like, when you say that you can give your model to another company to evaluate on their data set, is there any, like, reverse issue with trusting that, like, they can maybe reverse engineer parts of your model? I mean, I know that there's this model learning type of stuff. Is that also something you want to consider?

Ruben Wolff - **24:02**

It's a good point. I had not thought too much about it. Of course, this was a big complaint of OpenAI. I think that as soon as you publish the model, you basically can't prevent this because anyone can become an open, a customer of OpenAI. And so they kind of can't prevent someone from distilling knowledge out of their model. The only types of people that can prevent it are those that sell their AIs, like B2B and maybe they're not applicable for us. So, yeah, if your model is so secret that you would not let like several independent entities test it, then maybe it's not the best use case. But maybe in the worst case it could be like an entity that's really afraid of their model getting leaked could choose like the top three benchmarkers in our system that are the highest reputation.

Ruben Wolff - **25:09**

So they could theoretically do that. But yeah, it is, I think it is a concern for some people.

Jason Orender - **25:18**

I think that you're probably. There is a use case for even models that need to remain secret. You could, you know, publish the results by blockchain address so that, you know, nobody knows precisely who each blockchain address is until you verify it with a private key. So, I mean, there's lots of opportunity, I think, for companies to show that our model performs well against these other mainstream models and not have to publish that their model, you know, exists, you know, by name. So I think there's a lot of opportunity there as well, because as of right now, none of those leaderboards contain the private models. And so there's no real way for people to check and see if I'm evaluating this private model, how do I know how it actually performs against O3 many or something like that.

Ruben Wolff - **26:52**

That is an interesting point. And I guess the way that a private B2B company could do this would be that they get tested and the test results are verified, and then when they go sell it to their customers, they can reveal only to their customers that this is them. Or even more sexy if there's a ZK proof. But I guess their customers, they would trust them that they have this certificate, but the public doesn't know who it was that was tested there.

Jason Orender - **27:28**

Right?

Ruben Wolff - **27:29**

Yeah.

Stella Wohnig - **27:31**

What I would be dreaming about, I'm not sure if that can happen, but what I would be dreaming about, of course, is that you don't have to put your whole model on shame, but rather maybe like a commitment to it so that you're kind of committed to having to evaluate it in this way and then using snarks or something. But it might be prohibitively expensive of course. But I mean that would kind of be nice if you don't have like, have to have public access in order for this to run. But then like a question is whether we can find some multi party computation way to make it so that the testing doesn't have to have the full access. I'm not sure.

Ruben Wolff - **28:10**

Yeah, I think we should investigate with the people that are currently getting evaluated whether or not how much they care. So of course all the companies that have public access to their model, like open AI and so on, they already have it so they're not worried about it. But yeah, let's investigate. Let's also quickly. Yeah, maybe we can jump into some of the architectural discussions and I intentionally didn't want to like present the architecture that had in mind, but maybe it's time to get a little bit into that. And I think most of you have already heard the design that I had in mind. But Stella, since you mentioned putting the model on chain, like that would be a different direction than I was thinking about going, my thought was that we have n validators that were benchmarkers.

Ruben Wolff - **29:03**

So one of them could be Scale Seal, another one could be ML Commons, another one could be, you know, in the case of let's take machine translation as an example because we have representatives. So another one could be actually one of the AI companies like Lasan being a validator. Now it'll be super obvious that Lasan would be biased towards themselves. But if the benchmarkers identify themselves, this is something that can easily be taken out by user who wants to get this number out. So we have n benchmarkers. Each of them are creating their own test data and this is off chain, but they need to register somewhere. Of course they need to tell the world like I am validating this task and let's say the task is just text to text, machine translation, English to.

Ruben Wolff - **29:54**

Then on the other side there are startups that all have models and probably like one of us from the foundation would run Google Translate and Microsoft Translate so that it's easy to compare. And then the, each of the benchmarkers live query the old models that have registered to this challenge and only once you are releasing the scores would you put those on chain or somewhere. And like the signing, the most important, the part where you absolutely have to sign things is signatures on the side of the validators sending the prompts and the Machine learning startups responding to it and then getting that response back, evaluating it and then signing again. The score that you gave for this one translation, often it's blue score, but it could be anything. Having those all signed and then you need to just publish these signatures.

Ruben Wolff - **30:58**

Putting it on a blockchain would make it canonical and widely available and completely decentralized. So I see all the benefits. But theoretically once these things are signed, you can put them on GitHub or IPFS or BitTorrent anywhere you have. As long as people can verify the signatures of the scores afterwards. Yeah. So on a super high level, I don't know, did anyone else have like high level, completely different ideas on how to do this?

Robert Cowlishaw - **31:39**

I've got a quick question of what you're explaining. So the validators or the test set providers, are they trusted entities in all of this or are they meant to be untrust or not required to have trust?

Ruben Wolff - **31:56**

That's a good question. So if I were to build a this in a blockchain way then of course it would be permissionless. But I think realistically it makes sense that there, these are trusted entities so we have less trust because we have verifiability. So you don't need to complete, like you said, you don't need to completely trust the U.S. Space Force or whoever is doing the evaluation because we create audit logs that anyone can look at afterwards having this in the framework. Anyone can see the scores, anyone can verify the scores, anyone can verify that the same prompt has not been reused. Those are all like that establishes authority with the benchmarkers.

Ruben Wolff - **32:47**

So basically I would say yeah, probably the need and authority in a realistic case, in a full blockchain case, we could force them, we could force the AI model providers to be agnostic to who is the person testing. But like that would be the problem if in a non blockchain world, why would, let's take lasagna. Why would Lasan want to give open access to their model? Because obviously you need to enable the API for the benchmarker from benchmarker that they've never heard of. Maybe it's indeed their competitor or. Yeah, so I'm kind of leaning towards like permissionless, but a whole framework for establishing the trust in the scores that you're giving out. But super open for other things.

Asmelash Teka Hadgu - **33:47**

Yeah, thank you for sharing your way of like maybe approaching this problem in one way. I think as you suggested at the beginning, I clearly see two things going on. I think one is maybe that we should discuss like theoretical kind of without even getting bogged down with the details of like, which are, you know, we can abstract away like as you came. What are the attributes of like the testers? What are the attributes of like the different, you know, service providers and then what are the requirements we need for the data set to kind of like then have, you know, one approach, for example that hopefully we can prototype on. And yeah, typically I'm thinking of like the core problem here is, you know, static data sets can be gained.

Asmelash Teka Hadgu - **34:37**

So we need to make sure that whatever we're working on should be dynamic. I think centralized is something that is the problem again that we're saying like from trust issue. National. It was mentioned it could be national interest. It could be from a company's perspective from the providers even there could be competition from testers as well. Right. For reputation and so on. So yeah, decentralized over centralized and then that trust issue. So I'm saying like maybe we should have like that discussion separately with like a concrete implementation. Then like maybe we can start with existing APIs on a specific kind of like ML task with a subset of like the testers to test this out in a simple prototype. Because otherwise like it would be. It might get a bit difficult to make progress. Okay, I'm done. Thank you.

Yufan Jiang - **35:40**

I do have one question. So because you mentioned that otherwise basically the public can also verify the results and I guess or I thought that the private test data set will not be published. Right.

Ruben Wolff - **35:53**

I think we. So open discussion. I think that the private test data has to get published after some time period. So you. It could be done in cohorts. I think the absolute best way would be continuous. So like every day each of the benchmarkers release like one test and then like after, I don't know, 30 days, then another maybe five days for evaluation and then they publish the whole all of the results. We could say that this is optional. I think it would just be. That would be the best thing. And, and I'm also thinking like why not if you are truly making, if you don't make new test data sets as a private benchmarker, you're just opening yourself up to the model providers, finding out what your test data is and fine tuning towards it.

Ruben Wolff - **36:54**

Having someone off the model just like find the right answers and train them in one small other alternative mitigation I thought of is that the benchmarkers could send a lot of data, like a lot of prompts and only actually evaluate like less than 1% of those. And then it would be harder to find out which of them are the test data. But it's. I think they should get published. I think that's the best way. Jason, you had another point.

Zerui Cheng - **37:36**

Oh yeah. From my perspective, I think we do need an evaluation pipeline for the benchmark, an automated evaluation pipeline, if that's right. I think for now most benchmarks don't have a very unified pipeline. You just look into their GitHub, look at their ReadMe and do it step by step to evaluate a model. But it is not automated. It needs a person to look at the readme and do it step by step. But in our case I think we need a more unified pipeline for doing the evaluation.

Ruben Wolff - **38:16**

Yeah, totally. Particularly if it's going to be happening continuously, which is what I think. Well then yeah, it has to be automatic. So for maybe this is much easier with evaluating AI models that coming from startups. So Lasan for instance, they have an API so just evaluating, getting results is already automated on his part in contrast to I suppose human eval and some others where you do it locally. So yeah, I think inherently in the process of designing the system such that it's a live system. You like. Yeah, everyone's gonna have to be, it's gonna have to be automated and for the validators or the benchmarkers, I think it's pretty straightforward.

Ruben Wolff - **39:06**

So they have some people writing new test data, put that into a database somewhere, upload a CSV and then there's just a small process going that's slowly sending out the data, collects all the responses and they have a separate process for humans to grade each of them. Yeah. So I think on the benchmarker side it's straightforward on the person that has an AI model. Yeah. Has to be wrapped in an API. And I guess there's another design question. How should the validators communicate with the AI model providers? REST APIs open API spec I think is the most like it's less decentralized, still decent. You rely on DNS. That's the only not decentralized part. But it's I think pretty decent and very industry standard.

Ruben Wolff - **39:53**

A super decentralized version would be going through like a relay network P2P like IRO or going through gossip sub, but probably just rest API calls I'm guessing.

Zerui Cheng - **40:09**

Yeah. And I have another question on the validators. I think it is a good way to involve the companies and startups like scale AI. They can provide high quality data if they want to cooperate with us. But I think there is another high quality data set that's from the AI conferences. There are thousands of papers in each AI conference each year and there are a bunch of benchmark papers. And if we can involve them as our evaluation set, I think it will be a very nice thing. But I'm not sure how to convince the authors and the researchers to come to our platform.

Ruben Wolff - **40:53**

That's a really great point and leads me to this piece. So for. As you guys know, I care about changing the industry. Like we're going to publish papers and that's cool, but I really want to improve the way AI works and for that we need to have a movement. There needs to be a community, we need to get people involved. It can't just be us. And so, yeah, part of that is academics, like you were saying. So there are people who have salaries from universities, many people, as you're many people here who are doing the work of making test data sets because that's a contribution to science. So yeah, how to convince them to participate in this is a good. Is a good question. My first thought was that if we have kind of like a paper machine. So let me.

Ruben Wolff - **41:47**

There's a good paper because every evaluating this, I think, I hope it's that one. No, it's not this one. All right, I gotta go search for it again. But basically. Dinabench ML comments they produce regularly produced publications as a result. Sorry, it's not this one. I'll go find it a bit after. But they regularly produce publications where they just show the results and some of them are super high impact because they. So in their case, they're saying they're building different test data, but it is like an interesting result if most of the industry thinks that model A is the best and then through our independent testing, model B was the best. And so I was kind of thinking, yeah, maybe that is just like we can help people get another paper out of it if they contributed to the testing.

Ruben Wolff - **43:00**

And then there's like a regular machine in the NGO that's testing putting these papers together and publishing them. That's one thought. We're going to have a lot of thoughts. I think I want to point out two other groups of people that are going to be interested or that we should get involved. So we have one, we have academics. And then the second one I think is super important is test data companies like Scale AI, We've already mentioned them. There are a number of companies who it's their business to make data sets and we want them to be involved. And then of course, what's this one decision to make? Attract people that are in the Process of releasing. Okay, this is actually kind of related to the academic one. This is also related to what you were saying.

Ruben Wolff - **43:49**

I think maybe third one here is AI startups. So like how do we get AI startups to participate? I'm not too worried about it. If we start publishing papers and if we get the other people involved, then we will just establish ourselves as the new norm and people will follow like they have to get benchmarked by us if it's the new norm. But yeah, there could be these other considerations that we've mentioned. So privacy of B2B models could be considered. Yeah, yeah. Anyone have other points on how to like establish the movement?

Jason Orender - **44:39**

Well, I think once we publish a paper that's probably the gonna have to be the opening salvo because we have to establish our credibility first. And then I think the academics would probably want to get involved since, you know, their business is publishing papers. If we can offer them the chance to publish another one, I'm sure they'll jump at it.

Ruben Wolff - **45:11**

So yeah, I do think we obviously need to publish a paper. But I would say that for instance Duna Bench, they existed like as an open source framework first before publishing the paper. I have to go look at the exact like timeline. But it was like, yeah, it was like write the code, publish the paper and then create the ngo like ML Comments like grew out of this and then got all this money from different AI startups because they need better evaluation. Yeah. Asma, what do you think AI startups care the most about to get tested by this?

Asmelash Teka Hadgu - **45:56**

Yeah, I can talk about that. But like maybe following up on the previous bit, the academy involving academics or getting those data sets anyway gathered is a great idea because even if we're coming up with a new approach, it has to also benchmarked against existing kind of approach. Right. And existing approach is basically these scattered data sets from maybe AI papers across different tasks and so on. I'm sure many of you know, like many papers come with data sheets for data sets. So by following those we can unify and kind of like have of, let's say what the academic benchmark data sets would tell us about existing AI systems. Now of course the hypothesis is that they're not good because all these, you know, these big companies are, you know, they're scraping the web.

Asmelash Teka Hadgu - **46:54**

And by that whether I personally believe this is intentional to kind of like, you know, hype up their models, but it's just a good idea to also show why this is broken. Right. And yeah, in terms of like what would be useful from AI startups kind of perspective? I think if you're like a tiny startup like ours and you feel like you really have developed something great, but you're at a disadvantage because there is no fair kind of evaluation data set, you're up for anyone who could come up with a good benchmark because if you truly believe you have something better than, let's say Google Translate, and yet there doesn't exist ungamed kind of evaluation data set that hasn't been used by Google Translate, we would absolutely love it.

Asmelash Teka Hadgu - **47:45**

And I think I can speak for many of the people building core AI models, whether it's translation, speech recognition, even large, or we call it in the African context, small language models. This is actually one of the pain points we have, is like a fair evaluation data set that's not tampered by big tech because we definitely believe, and we have evidence that they have used all the evaluation data sets as training data. So their results are always better than anything you could build if you follow just like standard kind of benchmarks.

Ruben Wolff - **48:24**

So you're saying basically the core thesis, if we succeed at that, then they're going to love it. Like it's already designed for smaller AI startups.

Asmelash Teka Hadgu - **48:34**

Absolutely. And this is something they cannot pull off only themselves because they run into that problem. Right. Listen, kind of creates a very independent data set, even independent that we don't touch and so on. But nobody would believe it because. Because you're saying the big tech companies are doing that. Why would we believe you that you're not using part of your data like your evaluation data set as your training data set it. So for us it's actually attractive if an independent kind of organization does it this way, you know, it's fair for everyone.

Ruben Wolff - **49:08**

Jason, you had another point.

Jason Orender - **49:11**

Yeah. I just wanted to point out that it could also be a way to evaluate the questions that professional organizations create to evaluate models. Because if they can submit their question, if they're a validator, for instance, then and their question performs comparatively with the ones that we're generating, they can say that their date, their test data set is good. So that may be another market.

Ruben Wolff - **49:55**

So these are these academics being test data sets or these are people like scale that have, you probably have to.

Jason Orender - **50:00**

Start out with academics, but I think that ultimately you could convince professional organizations to be validators and submit their own questions as part of the process and then they can point to the fact that their questions performed well in comparison with the other questions.

Ruben Wolff - **50:25**

So how do you define a question performing well versus bad well.

Jason Orender - **50:31**

I mean, there's several statistical tests that you could use that would evaluate how each model performed on a, on a test that would indicate whether the test is accurately predicting how well it generalizes. So you would need enough for a statistically significant sample. But, but once you have that, you could evaluate the questions themselves.

Ruben Wolff - **51:14**

Yeah, yeah. So in between them, which is this is where if I want to do this statistical evaluation of different portions of the data set, that's where again, I think it's important that at some point in the future the, all the test data that everyone has made and the results that these are published, so you can then do the statistics yourself.

Jason Orender - **51:36**

Yeah, ultimately that would be required to get trust, I think. But, but the fact that some people make their, or some companies make their living, you know, creating tests and benchmarks, it would be good for them to have a way to prove that their tests are actually comparable with the best tests out there.

Ruben Wolff - **52:08**

Yes. So for the companies like scale, probably the smaller ones, because scale has established itself as a leader, and the smaller ones making test data sets, showing that their test data is good and then maybe scale costs you like, I don't know, a million to have them make a data set for you and then you can establish yourself as a commercial test data set producer that costs less than this. And in this, the on chain results that are very hard to falsify, you can see that this, you're producing good test data.

Jason Orender - **52:38**

Right?

Ruben Wolff - **52:38**

Yeah, I see that. That's cool. I didn't think about that angle. Stella, you have another point?

Stella Wohnig - **52:44**

Yeah, I think the point that I had actually ties in very well with what Jason just said. So I'm trying to blame blockchain again because I only have one hammer. So the thing is that blockchain gives you timestamps, right? So the thing is like why is it a problem to release your test data set when you have it? If you have to commit on the chain already to which model version or which API kind of you have to evaluate from and then you freeze.

Stella Wohnig - **53:13**

And then multiple people and in a similar marketplace fashion can publish different questions, different tests, and then maybe, you know, you run them on all of them and maybe there should be some process to bet on the chain, maybe with some kind of like committee voting on which of these questions they actually think are like viable so that nobody interjects some bullshit questions there in that place. But yeah, so if you, then, if there is a mechanism, like Jason said, where you can actually evaluate how good the questions are, then you know, you can maybe like have a similar thing as you explain the potential where you're rating both.

Ruben Wolff - **53:56**

I think an important point. Sorry to interrupt. Just thinking about time. This is a super big question is do we want to evaluate open source models or closed source models? Like things that are APIs versus things that are like something you can download. Because if it's an API, there's nothing we can commit on train that would prevent the API, the person that controls the closed source hosted model, that prevents them from changing it in the background. So I think that's an important question for us to ask. Do we, do we want to focus on closed source or do we want to focus on open source or let's say, let's call them open weights, like downloadable. And I'm leaning towards the closed source because I think that's where the problem is bigger. And the open source, it's kind of like it can follow from that.

Ruben Wolff - **54:50**

You can package an open source thing behind an API. Yeah, but it's. We, we have to ask people to see whether or not they care about this. I have people, the guys from ML Commons that I talked to, they said, they asked me the same questions like how do we know that the AI model wasn't changed behind the fact. Because in Duna Bench they have this mechanism. Everything must, all the models must be uploaded to them and then they manage it in a, their own little trust execution environment. It's just a question that we have to do. We have more discussion coming up. I do want to make sure that we get a little bit of organizational stuff in also though make one more point very quickly.

Stella Wohnig - **55:39**

I think there's a fundamental issue potentially with wanting to have decentralized contribution to the tests, but also wanting to use just the API, which you can change. Because I think if you don't put like some commitment, some time restraints in place for having to use an old model, then what keeps you from, you know, coming up with your own things, then training your own model on it and you know, kind of waiting before you give it to other people. So like somehow you have to guarantee the novelty of the testing sets and I think that would.

Ruben Wolff - **56:13**

So the novelty of test sets we can guarantee of the models is, that's the part where it's difficult and I think committing the novelty of the test that there's no problem. You take the, each of the test prompts, hash them, put them on chain and then we can see later on that you use those.

Stella Wohnig - **56:30**

But if it's decentralized, what keeps companies that have I models from training their own AI models with the data that they're coming up with themselves and then putting the tests that they put into their own models out there to test other companies?

Ruben Wolff - **56:46**

Yeah, they can do that and I think I mentioned it as an example. Like I assume that Lasan AI would put out one data set and test the other model the other companies. But that's why we have multiple benchmarkers, there can't just be one. And that's. Yeah, that's fundamental of that decentralization part. So three minutes left. I want to a little bit back to organizational stuff. So one question is like org structure. So there's writing a paper which I suppose is more informal and we just figure it out ourselves. But I am interested if anyone has opinions on like how to organize the NGO kind of long term part of this. So Duna Bench originally was just funded by Meta and then was donated to ML Commons. I've worked in the decentralized at any foundation and I, I kind of like their model.

Ruben Wolff - **57:41**

So you make a foundation, it has a mission and then there are working groups. The working group puts out the gold that they have. So I could see us making a new foundation. I could see us firstly just like write the code. Also you can look at Conda Bench. They seem to just have written the code and people just open source contribute. Yeah. So let's maybe. So option one would be there's just a paper and there's open source contributors and there's no real Org structure. Another one would be making a new foundation or consortium and then start getting NGO money into that for the long term sustainability. Another one would be to join an existing foundation like ML Commons or the Decentralized Identity Foundation. So yeah. Does anyone have opinions on org structures or the first option? No Org structure, let's just build.

Robert Cowlishaw - **58:40**

For me build and prototypes seems to be the best way to push things forwards. Even if they're terrible at the beginning. Just having something to show people is really nice. So I'd maybe be more on the side of build something and then we go from there. But I mean I'm academic, I don't know how to do business.

Ruben Wolff - **59:04**

For sure we should build and we should just go. There should be no stopping. I'm just thinking that again to have a movement you need to have people involved and so being part of ML Commons would already get us direct access to the top researchers at Meta, at Google, at all the big Companies because they're already using that to test their benchmarks. Yeah. So I'm kind of thinking about it from that perspective. It gives even more reputation to the project, which maybe will get more people involved. And you know, at the beginning it could just be a website at a name and it just makes it sound more grand. Or joining an existing one. Yeah. Code first. I agree. Code right away. Never stop coding. Anyone else care about Org structures?

Asmelash Teka Hadgu - **59:57**

I think maybe what I wanted to just agree with Robert, I guess maybe what he also wanted to mean if I'm right, is basically that we should start kind of putting a prototype of this idea and maybe also have a, you know, writing on the site just to kind of clean out theoretical kind of things that we're discussing. Have, you know, strip out all the jargons and have a unified way of referring to all components involved in this evaluation. And then kind of from there, if there is something substantial we can, you know, it would be clear whether, you know, we have aligned kind of organizations we want to join or this is something independent. But that's my view.

Ruben Wolff - **01:00:42**

Yeah. Cool. So I guess I should have continued the question, which is that it kind of leads into how we work together. So like at the Descendraden foundation we had bi weekly calls and there's like a chair and organize what's going to be discussed at this point and there's like deadlines for when we need to have something decided. I would say that actually the way that works and also how the Linux foundation works, it's quite slow, even though they pride themselves in being fast. So we should probably start with just working on something. Maybe then the question is more broadly like, yeah, how do we organize ourselves?

Ruben Wolff - **01:01:23**

Bi weekly meetings, asynchronous communication through GitHub discussions maybe like I can figure out who wants to work on what and do a bit of coordination and then make smaller groups like the people who said they want to work on development. But we need to of course have interchange between development and theory. Yeah. How to work together.

Saif Mir - **01:01:51**

I just wanted to say I think I'm on board with Azme and Robert is that I think if we build something we'll have a far stronger idea of what our tangible goal is. Right. If we kind of go straight into organizational stuff and you talked about ML Commons having connections top Meta and Google researchers. If we come there and we don't have much to show, I think it'll be a weaker argument than if we had just started from the beginning and built versus, you know, Actually having something to show. Right. That's, that's where I stand at least. And Maria had contacted me about, I don't know about everyone else but about co authorship on white papers. And you know my skills are best applied in peer review of papers. I come from more academic background and also in some technical stuff.

Saif Mir - **01:02:38**

So I think it'll be build first we can find out what everyone is best at and then apply those skills in an organization. Those are just my thoughts.

Ruben Wolff - **01:02:47**

Okay, so maybe I'll throw out the idea that you know, these idea ruminates. I. So I actually have like a list of 22 design questions that we didn't have time to get into. But perhaps. Yeah we could just. The minimum is like we could have another session like this, see who is still interested after the first session and then just get into the nitty gritty design questions. Maybe that can happen simultaneously with a like session that's specifically around coding. Yeah.

Robert Cowlishaw - **01:03:27**

Is there somewhere we can see the design questions?

Ruben Wolff - **01:03:30**

Yes, I'm totally going to.

Robert Cowlishaw - **01:03:32**

Then we could prepare some answers and think about it a little bit.

Ruben Wolff - **01:03:38**

Yeah, I didn't send anything out yet because again didn't know how we want to organize. I think a lot of people said they want to organize through Slack. I don't really want to pay for Slack but I'm down for doing that the other way. I think organizing through GitHub is really good. But yeah, I'll just like send this stock to everyone because there's also the. My structured notes which will be in addition to the AI automated note takers notes. So yeah, I also like to work inside Google Docs. Maybe there's here like do we prefer working inside GitHub or working inside Google Doc? That could be another thing. We just make a doc and all of these questions, you just like hit enter and put your opinion inside.

Ruben Wolff - **01:04:26**

GitHub is a little bit more structured where you can make discussion points and then report apply to them and form a spec through that.

Jason Orender - **01:04:36**

I think I would prefer GitHub simply because that allows whenever we start developing, you know, who does what. It'll allow us to build you know, assignments and get, you know, create cards for different aspects of the, you know, problem. It'll just be more organized.

Ruben Wolff - **01:05:03**

I also like GitHub. I like it because it's, you know, you see contributions of people. It's like your public profile. Even before anything is published you can see people worked on something. So I really like that. And I mean for sure we're going to have the GitHub repo that, you know, competes with Contra Bench or Dina bench. Yeah. So I think it's kind of the de facto. Yeah, maybe. Does anyone oppose us having most of the discussions centered around GitHub?

Because then I could make like, do they have this here? So there's like, you can have these discussion sections. So this is pretty cool. I'll try to put something together and you guys can fire at me if you prefer anything else. We are five minutes over time, but I think we got most of the points in.

Ruben Wolff - **01:06:04**

We didn't have too much time to discuss the open points, but that will be the next step then. So, yeah. Thanks for coming, everyone. Pretty happy about the turnout. Excited to see that people care about solving real problems. I'll start a discussion group. I'll start a GitHub and also circle back individually with each of you to see what part you want to work on or if you want to make a private statement on something. Super exciting, though. Great Internet.

Robert Cowlishaw - **01:06:43**

Thank you. Cool.

Ruben Wolff - **01:06:44**

Cheers.

Asmelash Teka Hadgu - **01:06:47**

Thank you. Bye.

Saif Mir - **01:06:48**

It was a pleasure to meet you all. Have a good one.

Ruben Wolff - **01:06:51**

Bye.