# SDS PODCAST EPISODE 706: LARGE LANGUAGE MODEL LEADERBOARDS AND BENCHMARKS

| | | |
|---|---|---|
| Jon: | 00:00 | This is episode number 706 with Caterina Constantinescu, principal Data Consultant at GlobalLogic. |
| | 00:19 | Welcome back to the Super Data Science Podcast. Today I'm joined by the insightful Caterina Constantinescu. Caterina is a principal data consultant at GlobalLogic, which is a full lifecycle software development services provider that is huge and has over 25,000 employees worldwide. Previously, she worked as a data scientist for financial services and marketing firms. She's a key player in data science conferences and meetups in Scotland, and she holds a PhD from the University of Edinburgh in Scotland. In this episode, Katarina details the best leaderboards for comparing the quality of both open-source and commercial large language models, and the advantages and issues associated with LLM evaluation benchmarks. All right, let's jump right into our conversation. |
| | 01:01 | Caterina, welcome to the Super Data Science Podcast. It's nice to see you again. So where are you calling in from today? |
| Caterina: | 01:09 | Edinburgh, Scotland, actually, I am delighted to be here, by the way. |
| Jon: | 01:13 | Nice. Edinburgh is a place that, as you know, from our, the time that we met at the New York R Conference that Edinburgh was a place that I spent a lot of my time during my PhD. I had a research collaboration there that led to my only, like really top machine learning journal paper I had, I had a paper in NeurIPS from my collaboration at the University of Edinburgh. So there's a lot of amazing computer science faculty at Edinburgh, in particularly in AI. And there have been for decades, like, it's like, it's a powerhouse school for AI. It might be one of |

the oldest AI schools around. I mean, I don't know what stretches back further.

Caterina: 01:55 That's so interesting. Yeah, that's definitely a draw to Edinburgh, which is I feel like it doesn't really even need it. It's such a gorgeous, gothic-looking place. But for me, my trajectory has been quite different. I actually came here to study psychology and then sort of seamlessly segued into data science through I don't know, some discoveries along the way that actually during my PhD I was becoming more and more interested in the data design sort of aspects and the experiments I was running, the data analysis as opposed to the psychological theory per se. But then also some accidents happened along the way. I found myself running the R meetup in Edinburgh, met up with a lot of people who were doing data science, and slowly but surely I ended up working for the data lab for a couple of years. And that was my first proper data science gig. And I've just stuck with it ever since. And I'm also still in Edinburgh. This is maybe 10 years later after having appeared on the scene here. So yeah, here we are.

Jon: 03:05 It's a beautiful city. Very dark in the winter, but it's a beautiful city.

Caterina: 03:10 That's for sure. That is, that is the tough thing about Edinburgh. I think in winter, the sun sets around 3:00 PM which is, which is a big grim, to be fair.

Jon: 03:20 But yeah, your affiliation with that R meetup in Edinburgh is I guess what ultimately brought us together. Because that's how you ended up having a connection to the New York R meetup. The Jared Lander runs. And so, yeah, you had a talk at the R conference. We filmed a Super Data Science episode live at the New York R Conference, and that was recently released as episode number 703 with Chris Wiggins. That was an awesome episode, and you had a great talk there as well on

benchmarking large language models. So I wanted to, I wanted to have an episode focused specifically on that today. So big news, at least at the time of recording, and hopefully still quite relevant at the time that this episode is published because this space moves so quickly. But very recently at the time of recording, Llama 2 was released, and Llama 2 came published by Meta with 11 benchmarks where, so there's three Llama 2 models that were publicly released.

04:27 There's a 7 billion, a 13 billion and a 70 billion parameter model. And even the 13 billion parameter model on these 11 benchmarks that Meta published, it's comparable to what I would've said was previously the top open-source large language model for chat applications, which was Falcon, 40 billion parameter model. So all of a sudden you have this Llama 2 architecture, that's a third of the size with comparable performance on these benchmarks. But then when you jump to Llama 2, the 70 billion parameter model, it blows all of the preexisting open-source LLMS out of the model, out of the water. And so yeah, so do you, should, should we believe this? Can we trust these kinds of benchmarks? What are, I mean, yeah, dig in for us into why these benchmarks are useful, but also what the issues are.

Caterina: 05:28 Cool. Yeah. So this is a really good starting point for our entire conversation because this example, I think, pulls in various aspects I really wanted to talk about. And I think the first one I'm gonna dive into is what, what does all of this mean? How can you in a way that really does justice to all the effort that's been ongoing for the last few years in this LLM space unpack this idea of performance and what does it even mean? What are all the facets that are involved? And at the end of the day, once you do start to dive into all of this detail with all the benchmarks, all the metrics, all the particular domains that are involved in a particular data set used within these test suites if you

want, how do you kind of drill back up again to come up with some conclusions that actually make sense across this entire field, especially as it's moving so fast?

06:26 So I guess something that I would probably point towards as a risk, first and foremost, is we're immediately placed within this arena of academic research, and it's obviously an extremely well-developed area already. We are talking about all of these benchmarks as you mentioned. But what I wanted to kind of flag beforehand as well is at the end of the day, the idea is that these models are gonna be exposed to some lay person, some user, and their idea of performance may not really overlap, particularly with what's in these all of these benchmarks. I think a good example to really drive this message home would be something like, maybe as a random average person, I might be looking to interrogate ChatGPT as an example on what a suitable present would be for my niece. And my entire experience, and my idea of performance might rather have to do with, are the answers creative enough?

07:40 Creativity is not something you typically see in these benchmarks, and how would you even begin to measure creativity? So that's one aspect. It might also have to do with, is the interface that surrounds these models making it easy enough for users to interact with the models per se. So yeah, I think that's something that's definitely worth pursuing a lot more in conversations, especially as the, the area develops further. But to kind of return to the more academic research angle as well. Then what I'd probably dive into at this point, because it's a really good solid effort of trying to incorporate a lot of facets of measurement metrics, data sets, is the whole effort surrounding the HELM paper. So rather than immediately talk about, is this model better than that model on this task or that task, or this metric or that metric in HELM, I think the-

| | | |
|---|---|---|
| Jon: | 08:51 | Sorry to interrupt you Caterina, but quickly, let's define what HELM is, at least like the acronym for our listeners. So it's the holistic evaluation of language models which yeah, I'm sure you're gonna go into is this comprehensive benchmark. But just before we get there, there was another aspect that you mentioned to me before we started recording related to issues with any of these tests. And maybe you were gonna get into it with HELM anyway, but it's this issue of contamination. |
| Caterina: | 09:17 | Yes. So one aspect that I think isn't maybe as obvious, first and foremost whenever we talk about evaluation risks, is this idea that especially models that are considered to be state of the art and have broadly speaking, "good performance", air quotes, they tend to be closed-source. So what happens there is we don't have a very good grasp on all the types of data that went into these models in the first place. And therefore, the outcome of that is we have some degree of uncertainty in terms of are we actually exposing these models within our test to data they've actually already seen before? And then if that's the case, then obviously any performance we see might end up being inflated. |
| Jon: | 10:08 | This relates to, so if we're using GPT-4 and we're blown away that one, it gets amazing results on these kinds of metrics, but it's been trained on all of the internet. And so these, these test questions, the test answers, they're all in there. And so it's a classic situation where when we're creating our machine learning model, we wanna make sure that our evaluation data don't contain the training data, but, if the algorithm's been trained on everything on the internet, probably the, the questions on any evaluation and the answers are already in there. Even more so it's interesting because we, there's this huge jump from GPT-3.5 to GPT-4 with respect to performance on things like the LSAT. Or, or I don't know if it was specifically the LSAT, actually it was, it was some kind of |

general bar exam, which actually, so that's, so LSAT I guess is to get into law school in the US. The general bar exam is once you have your law degree and you wanna qualify in a whole bunch of different states in the US there's this general test, and I can't remember the exact numbers, but like GPT-3.5 was like, you know, nine out of 10 humans would outperform it. And then with GPT-4, it was the other way around, only one out of 10 humans would outperform it on this bar exam.

Caterina: 11:23 Yeah. So that's, that's actually a really good example because LSAT is definitely part of these benchmarks. So if something like GPT-4 was trained to actually perform well on that, then if you come in and try to test it again on that same sort of benchmark, then that's slightly pointless because you're not gonna really find out anything new about its performance. And that kind of brings us to a different point that I'm, I'm glad we're able to make at this point. There's this whole idea of there's probably never gonna be a particular point in time where we can stop refining and updating these benchmarks because well, first and foremost, we don't know exactly what's been incorporated in the training sets in the first place. So the only real way around that is to kind of find clever and cleverer ways to test the performance on models and keep updating the benchmarks themselves.

12:25 But separately as well, as performance evolves, then benchmarks actually might become obsolete, and relatively speaking, too easy. So from these two points of view, there's been this effort to keep adding new tests. For example, BIG-bench I think started off with 200 tests or something of that nature, but now has 214 for this exact reason. So that's why there's probably gonna be a lot of movement also from the perspective of any type of standardization that might increase over time, because currently performance can mean a vast number of things. It could mean accuracy, it could mean fairness, it could

mean lack of toxicity. So a big measurement problem is how do you incorporate all of these different aspects and do you even need to, because there is some indication there is, there are some pieces of research that would suggest actually, despite being substantively quite different things, all of these facets end up being very highly correlated, which is also an interesting idea. So yeah, for all of these reasons, I don't think the research in this entire area is gonna stop anytime soon. So another big problem is how do you even keep yourself up to date and digest everything that's been happening in this field.

Jon: 14:06 Yeah, this does seem really tricky, this problem of constantly having to come up with new benchmarks to evaluate. And that's gonna become a bigger and bigger problem because, presumably in the same way that when you do a Google search today, you of course are getting information that's minutes or hours old from across the internet, and it seems conceivable that in the not too distant future while models like GPT-4 today are trained on data that stopped several years ago, presumably people are working on ways of constantly updating these model weights so that you have the LLMs right there in the model weights using up-to-date information about what's going on in the world. And so somebody could publish a benchmark and then minutes later an LLM has already memorized the solutions. So it's yeah, moving goalposts, I guess is the definition.

Caterina: 15:03 Exactly.

Jon: 15:03 Now, on the other hand, we can certainly say then these models are getting better. So despite all these issues, like I feel very confident that when I'm using GPT-4 relative to GPT-3.5, I am getting way better answers than before and much less likely to have hallucinations than before. And so these tests should measure something like there's, you know, these, these tests I think do have value. They have,

they have tremendous value, and you know, they, they should correlate. I would hope that they would correlate I, or at least it seems like when, when these papers come out and, you know, Llama 2 comes out and I see that, wow, it, the 70 billion Llama 2 model, it outperforms Falcon and Vicuña and all these other previous models. And then I go and use the 70 billion Llama 2 in the Hugging Face Chat Interface, and I'm like, wow, this is actually pretty close to GPT-4 on some of these questions that I'm asking it that I feel like are questions that it, that it hasn't encountered before. So there is this underlying real improvement happening and it does seem to correlate with these quantitative metrics, but yeah, the thorny problems, lots of thorny problems. I don't know. Do you think that HELM, it seemed like you felt like HELM could be a solution that you started talking about earlier?

Caterina: 16:27 I, I think the way they went about trying to systematically unpack performance and try to cross various factors is probably the way I would've ended up organizing this research. So that's why it really stuck out to me. But yeah, the, the sheer scale of effort that went into it does make it very difficult to really at some point see the forest for the trees. And I want to dive into this idea a little bit more, but yeah, we're talking about for example, I think five or six core types of tasks from things like summarization, information retrieval, it's sentiment-

Jon: 17:14 I've got the, I got the page open in front of me. So again, HEML, it's holistic evaluation of language models, and it's a Stanford University effort from the Center for Research on Foundation Models, CRFM. And there are 42 total scenarios that they evaluate over a bunch of categories like you were describing. So like summarization, question answering, sentiment analysis, toxicity, detection, it goes on and on and on. Knowledge reasoning, harms, efficiency, calibration. And I'm not listing all the individual tests, I'm listing the categories.

| | | |
|---|---|---|
| Caterina: | 17:49 | Yes, exactly. |
| Jon: | 17:50 | And those categories, there could be half a dozen to a dozen different tests. |
| Caterina: | 17:54 | Yes. And multiply all that by the tens of models they're considering. So very quickly, you arrive at this wealth of information, and if you take a step back, you naturally ask yourself like, well, what does all of this mean? Now the authors helpfully try to sift through this volume of information by creating a leaderboard on the website. And this is another really interesting tool because it's not a unique concept. We have leaderboards on Chatbot Arena, and we also have one on Hugging Face. But here's the thing, my initial thought process was, oh, great, I don't have to keep up with individual models necessarily. I can just take a glance at these leaderboards, get the gist of what's been happening in the area, and then anything that kind of leaps out at me, that's what I'll dive into a bit deeper. |
| | 18:48 | But I kind of started to realize that's not, it's not quite so simple. Because even with these three leaderboards, the reality is they're evaluation criteria. The models themselves that are included don't overlap. So looking at three different places is already kind of creating a hazy picture of what's really going on. So connected to this idea, I kind of realized that actually papers as vast as the HELM one, kind of subtly introduce this concept of time horizon you're interested in. Because if you're interested in models in the here and now, because maybe you want to pick one for a particular application that you wanna create, then sure, you're gonna dive into these and think, okay, for this task, this metric, I want to see which one does best and I'll just go with that one and test it further myself or whatever. |

19:54 But maybe there's more to the story. If we have a longer-term view, then maybe what we're gonna be interested in is nothing to do with the particulars of this model versus that one, but rather issues like, what is a good standardized way that we can even think about measurement of these things because it's so vast, because it involves so many different aspects. Maybe at some point in the future, rather than checking tens of different benchmarks multiple leaderboards, maybe there's gonna be a distillation of fewer places to actually check, or at least we can hope. And there's also an extra longer-term focus. Because at the end of the day, once we get all of these metrics right, like accuracy in terms of, I don't know information retrieval or Q&A and, and any associated metrics that get computed for tens of models, what we can do with those is start to frame everything as a prediction problem, which is where things get really interesting.

21:08 Because if we keep collecting these types of metrics, we're finally gonna get closer to this point in time where we get to say, okay, what are the ingredients from various models that actually go into this observed level of performance? Is it the fact that they have this many parameters? Is it the fact that they had this training objective? Or like, generally speaking, is there some sort of recipe of success that tends to lead to better performance? And if so, what is it? And we won't really know the answers to these types of questions unless we do all of these evaluations, but look at them from this much broader perspective of not this model or that model, but general laws that somehow govern how LLMs operate on a general level.

Jon: 22:01 Yeah. All really, really great points and very thoughtful to think that we could eventually converge and have kind of one state of truth for you know, to go to. It is interesting going to the Open LLM leaderboard from Hugging Face at

time of recording, we do have various variants of Llama 2 that are generally near the top. Looks like some groups have kind of retrained it with more instruction tuning. And yeah, Hugging Face is trying to do an average over some different evaluations, like HellaSwag, like MMLU, like Truthful-QA, but those tests are just three of the 40 tests that HELM ran, for example. Yeah. So I guess, I mean, it's, it's nice to think that we could maybe go and kind of have one absolute answer, but I think on the other hand, depending on specific use cases that you, that you're gonna have for you or your users, maybe these different kinds of benchmarks, this kind of level of granularity is useful.

23:09 So with Llama 2, for example, I've actually not tested this myself, but I've read that Llama 2 doesn't perform as well on code tasks or math tasks as something like GPT-4 even though it can be comparable in a lot of just plain natural language situations where it's just human language. So yeah. So that kind of distinction could end up being important depending on your use case. Like you wouldn't wanna, I guess, take Llama 2 and make something that's kind of like a GitHub co-pilot with it.

Caterina: 23:42 Yeah.

Jon: 23:42 You might wanna start with something else.

Caterina: 23:45 That's to be fair, yeah, I, I do agree actually with that point. And it does bring to mind all sorts of really interesting tests that are part of BIG-bench. And we're dealing with things like finding anachronisms and anagrams and stuff like that, which depending on the application of a model might really be completely irrelevant, so, yeah.

Jon: 24:12 Yeah. And yeah, so in addition to HELM and the Hugging Face Open leaderboard, which I'll be sure to include in

the show notes you also briefly mentioned the Chatbot Arena, which, yeah, in some ways it collects more a valuable, more expensive data because instead of having evaluations be done on these benchmarks, there's head-to-head comparisons, and then human users select whether they like the output from Model A or Model B, and they can be blinded as to what those models are. And in the very next episode coming up episode number 707, we've got Professor Joey Gonzalez of Berkeley University, who is one of the key people behind that Chatbot Arena. So he's gonna go into a lot more detail and he'll also disclose for us why it isn't as perfect an evaluation as it seems. There's still some issues. Like, there's always, yeah, I guess we're, you know, I guess like many things in science and technology, we are making errors, but hopefully smaller errors all the time and moving in the direction of progress, which again, it's safe to say like, you know, all these kinds of criticisms that we can have of these particular evaluation benchmarks or leaderboards. Ultimately, we know qualitatively that this is a very fast-moving space, and it's crazy what these models are doing recently in the past year.

Caterina: 25:42 And, you know, what I mentioned towards the beginning of the podcast, having to do with end users and what do they actually think of as good performance? What does that even mean? And I think Chatbot Arena actually gets quite close to this idea with their system of incorporating these Elo ratings. So that's something I really enjoyed playing around with earlier today myself. So broadly speaking, this is an approach that's been adopted from chess. So in terms of what happens in larger tournaments, you might have two players opposing each other, and depending on who wins, they either get a boost in points, or if they lose, they actually get points deducted. And the same sort of approach is used on these LLMs. But just as a regular user, you might have some prompt your mind like please generate text as though

Elon Musk had written it, or something like that, or like the text of a tweet.

26:54 And I tried this earlier myself and to be, to be fair, both answers I got from the competing models were actually quite legit Musk-sounding, if you will. So yeah, that's, that's a lot of fun to to play around with. And it's definitely a highlight in terms of what Chatbot Arena contributes as opposed to say HELM. Although even in that case there is an attempt made to incorporate some human feedback into the loop as well, but I don't think it's anywhere near being the focus of that body of work.

Jon: 27:34 Nice. Yeah. But a, a good mention there of the kind of thing, this kind of human feedback as being a great way of moving forward and the Chatbot Arena, I think everything is made available. All the data are made available for people to use and make models better. So very cool space to be in. Very exciting times to be in AI in general, as I'm sure all of our listeners are already aware and maybe part of why they're listening to the show. Yeah. So Caterina, before I let you go, I ask our guests for a book recommendation. Do you have one for us?

Caterina: 28:11 I do. It's something that sprung to mind, although actually my first encounter with this book was a very, very long time ago when I was still doing my psychology degree. And I actually have it right here with me. It's The Illusion of Conscious Will by Daniel Wegner. And when I came across this, I was actually studying in France on an Erasmus grant, and I remember being stunned at this concept that conscious will can actually be manipulated experimentally. And it honestly, it's, it's a joy to read the level of intellectual ingeniousness in how these experiments are devised. So that people's subjective feeling of having wanted to do something ends up being manipulated is, is just to me at this point unique. So if anybody has any curiosity about this, I highly, highly

recommend it. And who knows, these notions of conscious will maybe will kind of come into the conversation and kind of have already with LLMs. So there you go.

Jon: 29:23 Yeah, that is certainly something, the, the relationship between conscious experience, artificial general intelligence, this is something that we dove into with Ben Goertzel in episode number 697. And it is something that as somebody with a neuroscience PhD, I'm really fascinated by. As I mentioned to you Caterina, before we started recording I had a full PhD scholarship to do a PhD in consciousness, so the neural correlates of consciousness, so trying to identify using brain scans or probably some, some of the kinds of experiments outlined in, in your book, in The Illusion of Conscious Will, where we use things like intracranial stimulation so you yeah, you've, you-

Caterina: 30:12 TMS.

Jon: 30:14 Transcranial magnetic stimulation. Exactly. That's what we're Yeah, TMS, thank you. Which allows you to have a magnetic signal. And you may remember from physics that magnetism and electricity are directly intertwined, and so you can send these magnetic signals through the skull and then impact the way that your brain cells work, which involves some electrical conductivity. And yeah, you can influence people's conscious perceptions like you're saying. And so there's this really, in some ways, it's kind of an obvious thing to say to probably scientifically minded people, like a lot of our listeners, that because we live in a system of cause and effect, you can't possibly have some little person in your brain that is separate from all that and somehow is making decisions in some, some way that's beyond just physical processes like you know, cause and effect collisions of molecules. Yet we very compellingly have this illusion of free will. And to

some extent, yeah, I mean, if you come to grips with that, if you really accept that free will is an illusion, then, I don't know, it can be tough. Life can start to feel tough, so-

Caterina: 31:32 It's a terrifying idea. Yeah.

Jon: 31:34 So, yeah, I didn't, I didn't end up taking up that PhD scholarship 'cause I was like, this might really do my head in. And got into machine learning instead.

Caterina: 31:45 Yeah. Well, I, I'm pleased you did because now here we are, luckily.

Jon: 31:49 Yeah. Well, anyway, thank you very much, Caterina. This was a really interesting episode, a really nice dive into evaluating large language models. Very last thing. If people want to follow you after this show, hear your latest thoughts, what's the best way to do that?

Caterina: 32:06 Probably on Twitter so we can find me at c__constantine.

Jon: 32:12 Nice. We'll be sure to include that in the show notes. Caterina, thank you so much and catch you again in a bit.

Caterina: 32:19 Awesome. Thank you. Bye.

Jon: 32:21 Super. What an informative discussion. In today's episode, Caterina covered how ordinary users of LMSs may have qualitative evaluations that diverge from benchmark evaluations. How evaluation dataset contamination is an enormous issue, given that the top-performing LLMs are often trained on all the publicly available data they can find, including benchmark evaluation data sets. And finally, she talked about the pros and cons of the top LLM leaderboards, namely HELM, Chatbot Arena, and the Hugging Face Open LLM

leaderboard. If you'd like today's episode, be sure to tune into the next one, number 707, when we have Professor Joey Gonzalez, a co-creator of the Chatbot Arena, as well as seminal open-source, LLMs, like the Vicuña and Gorilla. Yeah, he'll be on the show next week.

33:05 All right, that's it for today's episode. Support this show by sharing, reviewing, or subscribing, but most importantly, just keep listening. Until next time, keep on rocking out there, and I'm looking forward to enjoying another round of the Super Data Science Podcast with you very soon.