Interview 24 - Ilan

Interviewer 1

All right. Uh, so I'll quickly go over the goal of the interview, then we will jump into the main content of the interview. Uh, so what we want to do in the end is to develop a catalog of quality issues in machine learning software system. So there's two word we need to understand there. It's what is a quality shoe?

Interviewer 1

What is a machine learning software system? Uh, by qual, by quality shoe. What, what we will need to understand is a word quality. And what is quality? Uh, to insert that, I will give you an example. Let's say we have two system that do the same thing, two software system that do the same thing. If you're able to see that one is better than another, then what you're usually referring to is quality, right?

Interviewer 1

So the one that is worse probably has some quality. . Um, and what is a machine learning software system? It's really simple. It's just, uh, any software system that has a machine learning component in it. So for example, you could think of a re recommended system or anything else, right? Mm-hmm. , um, is it clear for you and do you have any question?

Interviewee

No. So far is good. Like, I think, uh, everything's gonna just come up from the conversation.

Interviewer 1

Perfect. Perfect. And if you, if you have any question at any point, don't hesitate to ask me. Uh, so, uh, let's start off with some background information, a about you. So how many experience you have in machine learning or anything else.

Interviewer 1

Mm-hmm. .

Interviewee

Okay. So, um, I studied my, my Bachelor, uh, completed in 2016 in the University X. Uh, I, uh, the bachelor name was like information engineering, which is like this umbrella term for many things. Not necessarily machine learning related, but that's the first, you know, Start you get with the usual, so it, math and whatnot.

Interviewee

My first, uh, ex uh, direct experience with machine learning AI and that atmosphere was my master, uh, which I, uh, did in X. And there basically is like, I took a course in mathematical, uh, of course, like a degree in mathematical computation, uh, mathematical modeling and computation. Sorry.

Interviewee

Uh, and there basically, uh, the curriculum was enriched with deep learning, uh, machine learning, uh, image recognition and sort of stuff. Uh, now I have been working in, uh, Country X, uh, healthcare, uh, well not healthcare, but like, uh, Country X tech company develops software for healthcare, uh, since 2019. And, uh, sorry, as a junior now, I am a mid-level, uh, associate.

Interviewee

and uh, I am mostly responsible for developing such machine learning models. Uh, and now growing with seniority. I also start thinking about, uh, production, like how to develop those models, not only to achieve a central score, but how can they be used and uh, you know, so, uh, very interesting. I should have this conversation cause you know, it's day to day job.

Interviewee

Yeah. Thank you. At least I think so. . Yeah.

Interviewer 1

Uh, what is the size of your company? More.

Interviewee

So I think we have a head count of, uh, a hundred people, uh, shared across, uh, in c2 like main physical office in Copenhagen and remote, uh, scattered around the US mostly. Uh, we have also some people in the Region X and stuff like that.

Interviewer 1

Okay, thank you. That this is for demographic purpose,

Interviewee

so Yeah. Yeah. Again, uh, I'm not explicit with any information, but you can just find the LinkedIn, but. Okay. Yeah. Perfect. Just because I don't wanna say anything stupid. Mostly . Yeah. Yeah. That works.

Interviewer 1

Uh, so let's start with the general question. Uh, what are the main quality issues you've en encountered with your data model or system so far?

Interviewee

Oh boy. Uh, so, um, well first of all, uh, it's very difficult to translate any type of, uh, let's call it more academic metric into an actual quality standard, right? I will say, so it's very easy for me, uh, to understand like, oh, this model, I don't know, it's, uh, classifier and a bi a a binary classifier has, uh, a certain recall, a certain position, a certain F1 score, but depending on the purpose of the model, , it might be still unusable, right?

Interviewee

Like having an F1 of 90%, depending on what application you're translating it to, um, is inconsequential to some extent. So converting those numbers into an actual, uh, quality measure, uh, it's, the gap is not Zoom media, right? Uh, so that's, that's the first problem in overall design. Uh, another issue of, and to be honest, this is definitely the biggest one, is data.

Interviewee

Uh, Not only working with sensitive data, which of course poses some challenges in like to storing data, acquiring data and, you know, uh, data agreements and whatnot. But you are dealing with, uh, um, organizations that do not necessarily have standards of quality or technology to store this data or handl this data.

Interviewee

Uh, moreover, in terms of, um, uh, standards, like you usually have access to silver standard labels at attached. That is to say that those labels are not necessarily reliable. They are generated by a human, and there's no ultimate source of ground truth to definitely check if this is correct or not. For example, a medical record they might have, uh, um, I don't know, a diagnosis, right?

Interviewee
Medical diagnosis associated to it. You'll ne that medical diagnosis. It's still generated after somebody's opinion, and there's at best somebody else down the medical chain that can confirm or disprove this label, right? And more often than ever, the usual, sorry, you only have access to a certain type of label, not the entire chain of labels, right?

Interviewee
So stuff like that. , uh, those two are things that I, oh, well, okay. Another that I can think of is, uh, model monitoring. Uh, so that is to say, okay, let's say that you have, again, you're binary classifier and, uh, you have performed your classic machine learning infrastructure training, validation testing. Uh, you assemble your metrics and your convinced this is the best model we can offer to perform disciplinary classification.

Interviewee
Good. after that model goes live in a live scenario, it is not so immediate how you can have a live pulse of how the model performs. Right? Uh, because you don't necessarily, I mean, again, it could be a practical problem, like you don't have access to live data. Or it could be something that because this data is not representative, uh, of uh, what happens in reality, uh, it's kind of trivial to, or it's not trivial, um, useless to compare it against.

Interviewee
Because like, uh, if you don't have, if you only have human notations, but no other ground truth, that you're gonna compare against the human and it's gonna be like, okay, you just confirm that your model has trained or uh, um, has captured a certain behavior. You're not checking that you are better or worse than an actual human.

Interviewee

Does that make sense?

Interviewer 1

It's, it's not clear to me are, are you saying that you're comparing the performance on your, of your, you're using training data to evaluate your model or am I ? Oh, no, no, no, no, no, no.

Interviewee

Absolutely not. Sorry, sorry. . What I meant is like, okay, no worries. Imagine if you have like, uh, a model and you wanna say, okay, this model, how does this model perform to a human operator?

Interviewee

Right? Yeah. If the only label system that you have are labels produced by pre-human operators, there is no way to tell that you are better than human operator. Because the, like you have, don't have another set of labels that are better or more reliable than humans. So the only thing you can show is that you will co how good you are compared to a human.

Interviewee

But that, that's it. Right. I see,

Interviewer 1

I see. Or, or how much you, you, you, you are the same as a human basically. Exactly. But Exactly. Yeah. Precisely. I see. And how do you address this problem?

Interviewee

Uh, you have to come up with more unsupervised, uh, metrics if you can. Uh, That's a good question. . Uh, so okay. Unsupervised s nitric is one.

Interviewee

Trying to see if you have any other type of, uh, uh, external data you can rely upon or super expensive, but the most reliable will be to have another set of connotation by performed by experts. I should, usually it's not. . Okay, I

Interviewer 1

understand. So one issue you have is really like you're missing annotated data somewhat.

Interviewer 1

You're able to have noisy annotated data. So the the labelers.

Interviewee

For example, to Frank. Yeah. Okay. But, uh, okay, let me get you to an example. So, uh, let's say that you have an a SR model. Uh, so speech to text, right? And you train on the usual, like liberal speech, uh, you know, u usually publicly available data sets.

Interviewee

Then let's say that you wanna test it on your own voice, right? Uh, you. Um, you do not have an annotated transcript of your own voice live, right? So what you have to do is speaking to a microso, uh, microphone, record it, annotate it, and then you can evaluate it if you translate the same example in production, right?

Interviewee

So you have like an know, uh, radio conversation as a, well, I, any type of audio conversation really, I cannot come up with example right now. Uh, either you have a set of people. Continuously take this production data, transcribe it, annotate it, and then you can compare the model. Or there's no other, like, there's no other good way, right?

Interviewee

To, for example, computer four data. Right? Does

Interviewer 1

that make sense? Yeah, I see. I understand. Yes, it does make sense. Thank you. Um, you also mentioned that there was no standard in data. Would you like to go more in detail about that?

Interviewee

Uh, I mean, this is at least for the domain that I deal with mostly, which is healthcare.

Interviewee

Uh, Healthcare is an older type organization. Organization, right? You can imagine the same with, uh, legal, I guess, uh, or banking. Uh, some banks that are a bit more, uh, lagging behind in the dig digitalization race, they might not have historical records that they still deem relevant, uh, but. Uh, they, the, that they were not collected according to a standard, right?

Interviewee

So for example, no. You might have, um, oh, we started annotate, uh, annotating data with, uh, uh, the user that generated the sta the data from 2005 onwards, right? Or, um, In the case of the audio, let's say that you have like transcripts and audio, but you do not have a linking between these two with IDs because those IDs were not generated back then because we didn't store the audio or we didn't, uh, uniquely identify the audio.

Interviewee

So you have to link it with dates and just like very crappy, right? Um, and healthcare has these issues like, you know, some organization are old. Not, uh, it also very dependent on the geography, right? Uh, in Region X there are some countries which are way ahead in the digitalization, uh, compared to others. Um, and also maybe at the understanding of these type of problems, uh, varies a lot.

Interviewee

Not only comes to country, but like, you know, to customer, to customer. So in this case, uh, uh, hospital to hospital, uh, or yeah, something like, . I see. Thank you. Uh, uh, sorry. And I forgot, and of course there's no stand, uh, certainty that one, uh, data collection, uh, center has the same way of collecting data to the dollar one.

Interviewee

Right. So as the same what, sorry? Has the same, uh, the, the same, uh, uh, way of storing data compared to another one. Okay. Of storing data. Yeah. Thank you. So, for example, one co can com, uh, store, I dunno, uh, think about audio. One can have different encodings, right? Uh, one can be, oh, I store my audio raw wave.

Interviewee

Sure. I store my audio, encode it as a ECM 16. This is our mono channel versus dual channel, right? There is no standard that, that like restricts any, anyone really storing their own data in awareness in. .

Interviewer 1

Okay. And this is an issue because you have to have the same, that your data has being the same encrypt, encrypted format,

Interviewee

something like that.

Interviewee

I mean, this is a silly example because like, you know, you just keep processing, right? Yeah. Uh, but this becomes way more of a challenge for labeling. Um, so, okay, for example, let's say that the customer or whoever owns the data, makes its own transcriptions. Uh, transcription guidelines are not universal. So, for example, a customer, uh, they are data collected to say, um, I only transcribe audio, you know, with model channel.

Interviewee

And uh, uh, I use these type of tags to identify, for example, uh, overlapping speech, uh, or noise in the background, whatever. And you, you as a person making a description for me are allowed to use the tag overlapping speech only if there's this amount of speech overlapping. X and y and blah blah ba rules, rules.

Interviewee

Next data collection center that also does this has its own way, right? So maybe it, it doesn't, uh, it doesn't use overlap speech. Uh, you just discard the audio if, uh, it's a weapon speech, like examples, right? And every time you have new data coming in, that is like real life data. Uh, this is like a challenge, right?

Interviewee

Uh, it, you might have your data restrain. Cute and nice slippery speech or people talking into a microphone or whatever with like well recorded ways of doing this. And you have like, I don't know, calls, uh, from like, uh, one 12 or whatever. That is like an absolute disaster, right? So, .

Interviewer 1
Okay. See, so, ah, it's, it's a great point.

Interviewer 1
Thank you. The, uh, mm-hmm. . Yeah. Thank you. Uh, so the next question I'm gonna ask you is really close to what you just discussed. It's about data collection. Uh, so did you ever use a service of someone manually fetched data for you? Uh, like manual data collector?

Interviewee
Um, can you elaborate? What do you mean?

Interviewer 1
Uh, so for example, you, I mean, maybe not an example.

Interviewer 1
So ju just someone who, um, Simple data from some problem, and they may label it or they may just label it, for example, DC n image and they give it a label.

Interviewee
Okay. So we have a team that does, uh, for us in the company. Um, I know that we relied on, um, external, uh, uh, company or service for annotating, uh, audio.

Interviewee
But this was for sure until 2019. I'm not, I'm not so sure anymore. Like, I have to admit, I'm not so involved in, for example, getting, uh, data contract signed. Um, because also there's all the extra level of security, right? Uh, so for example, we could do, I, I think actually we, we cannot do this anymore because we cannot share data.

Interviewee

With people that are not in the contract. So not that outside organization. And even then in the organization, it's not like everybody can access it. Right? We ab we have to abide to very strict rules. Um, this is a very roundabout way to say, I don't know. Sorry, .

Interviewer 1

No worries, no worries. Um, but if you use that data, maybe you are aware of some quality issues within this data.

Interviewee

Well, um, Okay. I'm, I'm just trying to remember because we had this external annotation at a certain point. Right. Um, okay. I can't remember to be honest, but like, that's fine. Let's make a, we can make another example with publicly available data sets, right? Because they cannot Yeah. That's my next question.

Interviewer 1

Yeah, go

Interviewee

for it. Uh, because, uh, so for example, um, well I, I mentioned Liberty speech and and, uh, Uh, now I can't remember any other public available data set. Cause Yes. Um, those are pretty right and, uh, well, well documented in literature. So those are okay. Like in the sense that you can come up with, uh, an understandable way p processing and handling the like.

Interviewee

Okay, makes sense. Um, but for example, right now we are dealing with a dataset called Mimic, um, mimic Free, uh, which is about medical. and that data is not great. , uh, first and foremost, because documentation is somewhat lacking. And second of all the, the extent of my knowledge about, uh, uh, publications regarded don't invest really that much time into data quality analysis.

Interviewee

Um, but when we looked into it, the dataset is problematic. , um, Now I know there's like a colleague of mine that is, uh, doing some research on top of it, so I'm not gonna, if, if you're okay with this, I'm gonna try to avoid sharing too much insights because, you know, if he's working on, uh, publication, whatever, I don't wanna like spoil his results.

Interviewee

uh, yeah, it's fine. But basically the data set is like, you know, it is not great. Like, yeah. Uh, very simply put, you don't know where the data comes from, right? I mean, you know where it comes from, but like, there's. Clearly issues, uh, with it.

Interviewer 1

Okay. And, and can you give us, if it doesn't spoil anything from, uh, your, the person you're working with, can you tell us what are the issues you have in the data

Interviewee

for, for example, uh, in, in honesty, do not know It is, reflects in a quality issue of the model.

Interviewee

But, um, as I mentioned, is medical coding. Uh, the, the data set. Medical coding, uh, is done with medical, uh, coding standards. So, uh, you have stuff like ICD or c p t, uh, you, you, the data set is collected from us. I can't remember the years, but basically throughout those years, uh, the, uh, the ICD nine standard for medical coding of diagnosis and procedures changed and there's no flag or information in the data.

Interviewee

of samples that have been used, a previous version of IC D or a newer version of IC d uh, which might translate in quality issues or not, but

Interviewer 1

yeah. Okay. And, and what is I C D I?

Interviewee

Uh, international Classification of diagnosis and procedures.

Interviewer 1

Okay. That, that's perfect. I will search by myself after. No worries.

Interviewee

International Classification of diseases. There we go. . Thank you.

Interviewer 1

All right. Um, so moving on to data preparation. Have you ever measured the quality of your data and or tried to improve it?

Interviewee

Hmm. Um, there's like a, uh, more fundamental discussion here that we had that is if you, uh, sorry. How much do you clear the data before the data is so clean that it doesn't reflect reality anymore?

Interviewee

Um, so for example, in the case of audio, there are many ways, for example, to remove background noise, uh, to try to, for example, do speaker ization. If you want to, for example, from a model channel, go to a fake dual channel, stuff like that. Uh, but then you have to ask yourself, okay, if I train a model on this super clean data, when this model will face again the shitty data.

Interviewee

It's a different data, right? It's not recognized anymore. So either you make the same pipeline that you do, uh, to clean the data when you actually tackle the issue, or there's, you are not presenting the same domain in a sense. Um, now for analyzing quality of the data, uh, yes. Uh, explor, we usually start work on project performing exploratory data.

Interviewee

uh, and they can vary, right? Uh, in general, I don't think we try to improve on the data, uh, but we try to understand what are the limitations of the data and how can we eventually circumvent them with modeling ideas rather than data engineering. Um, we still do some, like data filtering. For example, in the case of, uh, facilitations, we might remove, uh, depending on the applications, we might decide to, uh, remove labels.

Interviewee

because like, you know, underrepresented labels, there's like, uh, 10 examples out of a million is like, okay, maybe we can remove them, right? Um, or in case of underrepresented data, we do something. We could do some data documentation or up or under sampling, but, uh, we do not necessarily dabble that much in data documentation, like creating new examples.

Interviewee

Um, it. Not so far .

Interviewer 1

Okay. And you mentioned two solution to the, the, the first problem you mentioned that when you do too much data cleaning your test data is not similar to what you train on in the end. Uh, and the first solution is you can just apply the data cleaning to the test data or the live data, let's say.

Interviewer 1

Mm-hmm. , right? Uh, why is not, is it like not, um, It seemed problematic for you. Am I mistaken?

Interviewee

Uh, no. I mean, you're right. Like, uh, so again, it might be extremely mistaken cause like, have no idea depending on what you're dealing with. Uh, for example, a say you have a real, real time application. If the data processing pipeline takes time, you're gonna break real time.

Interviewee

So very simple. That first problems super practical, right? Um, the other thing that is, uh, again, uh, more of, uh, just pragmatic purpose. Let's say that, um, well, no, sorry, is that depending on, uh, who, uh, on which data the model is exposed to, you might want, uh, and, and, sorry. And, and if you process the same,

pipelines are very complicated and custom made.

Interviewee

It gets cumbersome to keep track of, uh, uh, this data source requires this pre-process. This other requires this other processing and then maybe something happens at the source that you don't keep track of because it's maybe not your responsibilities like somebody else giving you this data that changes some standards and terms communicate to you and your model breaks.

Interviewee

Right. So maybe something like that.

Interviewer 1

I see. So basically data processing first may slow down your air prediction, which might be. , which might be an issue in live setting. And the other issue is it's really difficult to, uh, manage and to build, like there's a lot of, uh,

Interviewee

pathways in your, it, it, it becomes, uh, difficult to manage easily.

Interviewee

Um, yeah. Or quickly. Um, uh, what else do you want to No, no. Yeah. Okay. I, I'll send it up as that. Perfect. Perfect.

Interviewer 1

Thanks. Um, just looking at time. Perfect.

Interviewer 1

Sorry, I'm just, uh mm-hmm. , I need a new question. Alright. Um, is there any other data quality should we missed that you consider relevant?

Interviewee

Mm,

Interviewee

well, I mean, uh, in general there's always the fact that, uh, I mean coming up from, uh, uh, academic setting in which. Image recognition on, um, uh, how's it called? The number one, uh, dataset? Yeah, like the, the, the classic, uh, image recognition, uh, sample data set. Cfar, nist. Yeah, for ex example. Thank you ma'am.

Interviewee

Listen, cfar. Um, you have those classic examples of like, oh my God, my segmentation, my, uh, uh, digital commission algorithm is amazing. And then you get to like real data, which is like a shitty PDF scanned, like, you know, an office God knows where, and it's like, it's super, super messy compared to ACA academic results.

Interviewee

Right. Um, so in, in general that that, but that's a broad discussion, I guess. Uh,

Interviewer 1

And basically the issue you just mentioned is that the models that are trained on the clean data like this or so far, they're not applicable in real, in real life, or they're

Interviewee

not using, I mean, not directly, right? Um, like either they might be Okay. So either because they've been trained on very safe data?

Interviewee

No, like super, uh, okay. Yeah. In a safe environment, let's say . Uh, or it might be because they're super engineered to solve a specific task, right? Um, like I think there were some dialogue stake tracking methods that I looked into. There were ex like super, super tailored to the application that they were trying to solve.

Interviewee

And the first question is, if then I change the data, Or if I change any of the assumption that the data has, for example, um, that my, uh, part of speech tags are correct, that, uh, my, I, I don't know, like, let's say just that cause , uh, or maybe I don't have part of speech or you know, it's unstructured text, so I have to.

Interviewee

More or less guess where, uh, sorry. I could generate those texts with space, but space, like an automated and I don't know what I'm doing. Right. Cause it's like, again, proper speech, not a guy saying, I would like to go to a restaurant. I would like the restaurant to be in the center. Like that's not how people speak.

Interviewee

Right. Um, sorry, going back to what I was saying is if you break any of the data assumptions, all this super engineered machine, how resilient this. And, uh, it's not, uh, I mean also, okay, I might be saying crap because like I haven't read papers, uh, as often as I wanted to, but it's not super common to people to run, uh, error analysis and ablation studies, or at least not all the time, or at least not as often as I would like to

Interviewer 1

Error.

Interviewer 1

Analysis of the prediction of the model. Yes. Right.

Interviewee

Okay.

Interviewer 1

Perfect. Thank you. Thank you for this information. Um, oh, moving on to model evaluation. So you mentioned earlier on that the metric do not reflect, uh, like what, what you define as quality for remodel, right? Mm-hmm. . Uh, so my first question is how do we evaluate the quality of remodel?

Interviewer 1

And as a reminder, quality is not only defined by the performance, but you also have other aspects such as expandability, uh, robustness, scalability, you name it.

Interviewee

Definitely, uh, well, I mean, you mentioned like some excellent metrics, to be fair. Um, in general, they all, they depend a lot on the end user, right?

Interviewee

Um, so you will have, um, So if you cannot necessarily, uh, quantitatively the define how good a model it is, or you can only test it, uh, let's say like, uh, in, uh, during, during training, right? Uh, then the only thing, if you cannot get anything afterwards, then you need to set some user tests. Uh, I cannot remember the acronym for basically, you know, user uses it.

Interviewee

How much do you like it? Uh, very simple put, uh, you can also have like stuff like, again, um, how robust it is. Like basically you could try to set up some, uh, deviation over time. Still not supervised, but whatever. So for example, let's say that you have, uh, this binary classifier and, uh, it classifies. , uh, something that you expect to be like, I dunno, you, you expect, uh, a certain distribution of positives and negative classes because that's what you're trying to capture, right?

Interviewee

So that, uh, for example, let's say, uh, what can be a good example? Mm. Uh, common brain. Think about it. , keep your time.

Interviewee

For example, no, uh, you could classify if, uh, uh, a call is, I dunno, damnit, if

Interviewer 1

if you don't have an example, we can move on. There's

Interviewee

no worries. No. Okay. Sorry. Just like completely blanked out on this one. But like, let's say for example, you expect 5% of positives, right? you could track, okay, what is my distribution of negative versus positive time?

Interviewee

If your models start either never predicting or always predicting, then either something change that any models stay the same, probably something in the data underlying it changed, right? Something in the process changed. So if you cannot qu have quantitative results, like constantly monitoring the precision or L F1 or whatnot, you can monitor deviation from the expected behavior.

Interviewee
Um, yeah, again, user testing is definitely something that is vital, uh, especially for, uh, companies or any type that is like pro a product. Um,

Interviewee
scalability is not a metric, but like, again, depends on what you want to do with it, right? Explainability. Explainability is definitely interesting, but. . Again, it depends on what, what you're trying to do. Like a classifier would be. Uh, uh, again, explainability is a nice to have, a really nice to have, but my experience is that it's not, it's useful during the sign rather than, uh, and use, let's say, it's more useful for me as a developer to know what the model.

Interviewee
Rather than a user that only sees zero ones and only cares about zero ones. Right. Um, to some extent,

Interviewer 1
yeah, it does make sense and actually it collaborate corroborates, uh, something else I read in another paper about explainability, so, yeah. Oh, nice. Makes sense. Thank you. Um, so you mentioned two ways. Sorry, I think I, I don't remember anyway.

Interviewer 1
Mm-hmm. , you mentioned user tests, right? How do you conduct these user tests? And, or my question is more what are the challenge you have when you conduct these user tests?

Interviewee
Hmm. Well, defining the test in a way that the user understands, and like very, because like, you know, usually the user is not a tech person, right?

Interviewee

Uh, so you have to make the test, uh, uh, foolproof in a sense that you test what you. and not just, uh, a way to spread a series of things, I will say. So you have to be really, really clear on, uh, the instructions and also the environment, which set exercise is a test, right? Uh, the same is true for any type of like human quality evaluation.

Interviewee

Like for example, I mentioned, um, if you have like an annotators doing tasks, Annotating, uh, data and you want to check how good are these annotations? So some type of infra annotator, uh, test. You also have to define that bulletproof, uh, procedure, right? Um, then in all honesty, uh, I haven't really conducted that many, uh, end user tests.

Interviewee

Uh, also because usually I, um, Again, being a company that were more product oriented rather than machine learning oriented. Uh, so I mean, of course the machine learning powers the solution, but I find again, out that, uh, um, a lot of the machine learning quality can be, um, uh, sorry. Uh, a lot of the perceived quality of the machine learning solution depends, for example, from the front end or the way these things are presented to the user.

Interviewee

So, The same model in two, presented in two, in, in two different ways, uh, can be perceived as better or worse depending on how good or bad the, the interface is. Right?

Interviewer 1

I'm sorry. Yeah, I was noting it down because, uh, yeah. Someone else mentioned it and I made sure I, I add, I added noted, so I didn't want to miss it up, but thank you, . Sure. Yeah. Um, yeah, that's, uh, that's pretty much it. Thank you. I will move on to model deployment. I, I see we are short of on time. Yeah. Sorry.

Interviewer 1

No, no worries. Um, I will ask you, I will jump, I will ask you like maybe three or four questions. Yes sir. If it's all right with you and uh, yeah, no problem. We

Interviewee

can. Thank you. I'm fine. We're going a bit over time. Uh, if you're okay with like, I don't mind.

Interviewer 1

Okay. That's great. Thank you. But we won't go too much for a time.

Interviewer 1

Thank you. Sure. Um, yeah. Uh, so, alright, so what are the challenges you have encountered during the deployment machine learning software system?

Interviewee

Mm. Well, man, uh, I mean, uh, real time is a big one depending on the application you're trying to serve. Uh, Um, it's, it's nice and dandy to dev to develop like these super complicated models that do like a bunch of stuff with millions and millions of half parameters.

Interviewee

But, uh, if they're not usable because like, oh, uh, do this prediction and it takes one minute and a half, no. Great. Right. Uh, but aside from this one, again, it's a bit, um, more of a design issue in terms of deploying it themselves, uh, tracking conversion. It's, it gets out of hand really fast if you don't have an infrastructure.

Interviewee

The supports proper version in tracking. Like, uh, for example, let's say that I have, uh, three customers and the same, the same model train on, uh, actually, sorry, three customers or each one of 'em of a different language. So tracking and uh, therefore I have three models. Each one that has to be deployed in the right.

Interviewee

Developed in parallel track and I continuously improve upon these models, being able to perfectly track model one. So because also, you know, training is not, you train the first time, the second time is better, the third time is better. You train 10 times, number seven is the best one, and you train out 10 times number, whatever, uh, who knows, right?

Interviewee
So tracking already upon your pool of experiments, uh, which one is the one is the best. And propagating this information throughout all the stack is challenging or more than anything, it, it, you really have to find a system and some tools that help you to do that. Um, then again, uh, I haven't dealt with diplomas themselves.

Interviewee
Like I have helped developing infrastructure, uh, to develop to. But actually monitoring and tracking of deployments, like, you know, checking that like traffic is all right and whatnot. Uh, not really my field of expertise. So I'm speedballing a little bit here. Um, there's still a challenge of like ensuring that, uh, um, you have right capacity, right?

Interviewee
So like if you start assessing a simple infrastructure, And you don't immediately go for Kubernetes, for example, with auto scaling and whatnot, eventually you might be limited by that, right? Um, you might also have issue with, uh, more legal stuff. For example, you might have, uh, some customers that again, hey, my data doesn't leave Region X.

Interviewee
So even if you are on Amazon, on Amazon, uh, sorry, Ws, sorry, uh, Google Cloud and whatever, you have to be sure that the service you're using to do all of this is in the region that you actually, you know, , uh, respect your legal terms. So again, it, it creates extra, uh, tracking issues. Like you have to be sure that everything deployed.

Interviewee
It's like maybe this model two is in the right region, otherwise I get sued. Right.

Interviewer 1

Okay. I see. So, and your model, or tied to some region, is it the model or the data that is tied to some region?

,

Interviewee

if the model has to listen to live data, it's more or less the same. Right. Okay. In the sense that if the model has to receive data, the data means that has to go through whichever region the model has deployed.

Interviewee

Okay,

Interviewer 1

I see, I see, I see. It's not the training data. The issue, the, the, the issue is the, the data you receive to

Interviewee

make prediction. Okay. The training data depends, right? Like if the training, if the training data is customer data, then depends on what the customer wants, then you might have your db that is like an instance on, on Asia that is like hosted in Region X.

Interviewee

Sure. The new can only access through, uh, an API that has certificates, yada, yada, I don't know, whatever. Uh, but then, okay, uh, that is again, design issue. Like I just make se make sure that all the data is in the right place, that only the right people can access it using the right security measures. But then traffic, everything goes back and forth with the model cannot be public, right?

Interviewee

Like, and again, it. Difficult, just extra work. It's very easy to forget, oh, I have my classified or whatever, and then I, I'm done. And that there's definitely not, not issue the case. Right. .

Interviewer 1

Okay. I see. And the naive approach to, to make sure the data doesn't go outside the country will be to deploy, uh, Amma all in every region of the work.

Interviewer 1

Right. But I guess you're trained to avoid that.

Interviewee

Uh, I mean, if you have infinite money, it'd be my guess, but, uh, , yeah, , no, but like, it is just like, again, it's more process. Like, it's, it's not a challenge in terms of, uh, code or tooling or whatever. It just process.

Interviewer 1

I see, I see. Thank you. Uh, moving on to maintenance.

Interviewer 1

How do you ensure that the quality of a machine learning software system does not decrease over.

Interviewee

Uh, I've already mentioned this in the past, right? So like live monitoring is also one of the things that, at least in my experience, I struggle the most. Uh, it's not so straightforward to set up monitoring, but if you should, uh, if you can have access to new labels, uh, incoming, then you can just like basically try to replicate your metrics that you have in.

Interviewee

If you can't, it should come up with some type of unsupervised metrics to check model deviation. Uh, you can also check, uh, not necessarily model outputs, but also what comes the input of the model. So, uh, is my data any different from the usual using, I don't know, some metrics? Um, That's, that's it. I mean, the more stupid approach should be, the model gets worse if the end user complains about it, but, which is true.

Interviewee

Right. But, uh, it is also probably not the most viable option to pursue commercially speaking. Yeah.

Interviewer 1

Yeah, that's for sure. All right. Thank you. Um, finally, I will, there's seven.

Interviewer 1

I will ask you the question straight away. Okay. If you have a question, ask me. Uh, so did you ever add quality, uh, add issues with one of the following aspect? So fairness, robustness, and some of them, you already mentioned them. Mm-hmm. . So fairness, robustness, explainability, scalability, privacy, and security of your models.

Interviewee

Um, well, okay. So for, I think the only topic that we haven't really mentioned is fairness. Right? Uh, it is definitely true that, uh, there are different demographics representing today in each data set, right? Especially, especially in stuff with like healthcare. Um, so we haven't faced any blatant issues.

Interviewee

Like for example, the were, uh, models trained on, uh, web crawl. They were like, you know, contain the. Slurs or biases towards like ethnic ethnicities and genders and whatnot. This is not by saying that we do not necessarily face biases, same gender, uh, which again, in the health world, it is true that like, uh, I dunno, some diseases are more prominent in females in bi, bi, uh, biological females and biological men.

Interviewee

Right. Um, but still, right. Um, we haven't, as far as I'm. Uh, faced any issues like that. Uh, but at the same time we, uh, I think we performed some tests. We tried to slice the data, uh, depending on the recorded gender, and we had, I can't remember what we found. Um, but not nothing like blatant, nothing like awful, let's say

Interviewee

I see. It is definitely though something that, uh, should prepare more a. Especially like, uh, I mean for healthcare, sure. But anything that, that, the more sensible the data, the more this should be a point of attention. Right. Um, I think there was like, what was it like this, an algorithm used by official recognition algorithm used by the police.

Interviewee

Uh, don't remember where it was like extremely biased, stuff like that without a bias and analysis. Like it's, it's not fine. Right. . , I know what you're talking about. Um, for everything else, uh, I mean, we have found issues in the sense that we have dealt with delivering quality in terms of, you know, ensuring security of the data, daytime, the models, ensuring, uh, robustness and whatnot.

Interviewee
Um, but I think I mentioned like challenges that we faced and tackled in Bruce. Perfect. Thank

Interviewer 1
you. Um, in your opinion, what is the most pressing quality issue researchers should try to solve?

Interviewee
Oh boy. Uh,

Interviewer 1
it can be for you. Whatever helps you the most.

Interviewee
Well, okay. Uh, again, full disclosure, I have, uh, I, I am, I'm not a researcher myself, right?

Interviewee
So I am not so involved into the research research world. Uh, I, I, again, just because I've been exposed to a lot of papers regarding automatic medical coding, uh, they're the most, uh, like prominent issue was zero to non data analysis, right? So zero one. Zero to non data understanding and data analysis.

Interviewee
Like many of the papers just went, okay, the data is a multi classification problem now, but now to the modeling aspect. Right. And uh, you know, my birth model is amazing. Like, okay, great. Um, sure it is like, I think the majority of the, of papers want to prove their modeling idea is the best. And so there's no necessary space for the analysis, but sometimes I think like, you know, trying to understand that better could like definit.

Interviewee

Like, you know, short, your model achieves an amazing advance score, but you're not really, so like, you're just like throwing parameters at the issue. Right. Um, I think that's something a machine learning tends to get an issue. Just the adding parameters and solving the, the problem that you're trying to solve instead of like, okay, can I achieve similar result with our fully connected network and just do some data pre processing.

Interviewee
Okay.

Interviewer 1
Since. Yeah, so more work on the data part, basically

Interviewee
to some extent, like at least understanding what you're dealing with. Again, I speak from a very biased sample of, uh, 10 papers that like I was actively looking for. Did they say anything about the data? Like nobody mentioned this change in, uh, in medical coding in the years.

Interviewee
Right? And if you're trying to predict medical coding, that should be like, pretty relevant to mention, right? So, . And again, it's not hard to find, it's like .

Interviewer 1
No, no, it's, it's perfect. Your, your answer is perfect. It, it was meant to be subjective, not objective, so, no, no. Okay.

Interviewee
Okay. Awesome. Don't worry about this.

Interviewee
Yeah. Uh, I just want, I want like to, um, just insult the category, you know, speaking from no place of authority, .

Interviewer 1

Yeah, no worries. No worries. Yeah. Thank you. Um, and do you have any other comment about the quality of ml?

Interviewee
I'm, um, I mean, I spoke, um, mostly from like, from my own experience, right? So all the comments, I, I don't have any other example of like that thing I don't like, like I know, I know, uh, a lot of my own company.

Interviewee
I know a lot of, like the bigger companies, like, I don't know Tesla, I guess. Uh, but the major criticism I think is already out there. Like I don't have any, I don't have anything revolutionary too or too smart to add to the big players. Right. Um, it, it's kind of funny to some extent how, uh, a lot of the papers, like again, Bert.

Interviewee
who has the resource to retrain a bird from scratch? Right? I mean, from my understanding, like a lot is pre-trained birds, which mean fair, but at the same time, making a new bird requires Facebook or Google computational power, right? And, uh, again, from my limited world or person, that is not too much in, into that, in into research.

Interviewee
I wonder how much that is impeding, uh, senti competition to some extent, right? Like, I don't know. Was it, uh, has compute available by still a university against Google. Right. So even just in terms, if you want to reproduce their, pay their papers. Right. Yeah, that's a good

Interviewer 1
point. That's good. All right, well, uh, I guess that's it for us.

Interviewer 1
So we, we don't have any more question, but what you didn mention gave us, I think it'll be really important and interesting and important for our research. Awesome. So, yeah. So yeah, thank you. And. I guess. Have a good night. Yeah,

Interviewee

thank you. And have a nice day. To you, it's like what, like XXh to Exactly. Yeah.

Interviewee

Nice. Okay. Uh, well, let me know when I can have the transcript, just like to double check that I haven't like broken all. Yes, everybody . Don't

Interviewer 1

worry. All right. Bye. Thank you so much.

Interviewee

Thank care. You too. Bye. Okay.
