Interview 43 - Amin

Interviewer 1

All right. Uh, so to start off, can you give us a bit of information about yourself once again, and how much years of experience you have in machine learning

Interviewee

or in general? Um, so I'm a senior product manager at an AI company and at my company we create softwares that make it easy for people to create neural search software.

Interviewee

And neural search is nothing but, uh, using, um, Using deep learning embeddings for doing search instead of, uh, using T F I D F or something like that for the search like a solar system does, right? Uh, a solar search system does. That's what I do right now. Um, I have been working in the field of machine learning since 2012, first as part of, uh, my studies then as part of research, and then as part of, uh, uh, uh, industry.

Interviewee

So you can say about 11. Great.

Interviewer 1

Super. Well, thanks to, we're glad ah, we are glad to have you here today. Yeah. Um, so, so to start off, what are the main quality issues you have encountered with your data model or system

Interviewee

so far? Um, I don't know if I can list them all because that's also a very g generic question because I, uh, don't know what, uh, you would consider a quality issue or not.

Interviewee

But generally speaking, most of the issues, generally speaking, just uh, issues with software come out of issues from data itself. So, uh, this means that the data is either if, uh, if you're working with label data that you're sourcing from somewhere, then the labels might be dirty. The labels might not exist at all, which means you might have missing labels.

Data-missing-value   Data-noisy-features   Data-noisy-labels

Interviewee

Uh, you might have, uh, variables that are dirty, so there might be noise in the data, and you might have variables that are not, uh, not available in some cases. So if you build a model, assuming that variable X, uh, would be available. In real time, variable X is not available. That's a, that's a quality issue in the data that is eventually also going to have, uh, an effect on the quality issue of your model and the quality issue of your software.

Data-missing-value   Data-noisy-features   Data-noisy-labels

Interviewee

So, uh, in my experience, most of the problems in machine learning software has, uh, come from, uh, problems in. Perfect.

Interviewer 1

Thanks. And so you give us an example of mislabel data. Uh, so I have a question for you. How do you address this problem usually when you have mislabeled data? Or how do you detect that you have mislabeled

Interviewee

data?

Interviewee

We can't detect that. Uh, we, we simply cannot because there, uh, I mean, if we could, so suppose you had a function which could tell you if a label is, uh, mislabeled, then why wouldn't you always use this? Which means that you will never have mislabeled data, so you actually cannot detect it. Uh, you can only detect it manually by looking at the data.

Interviewee

And obviously if you're working with millions of uh, datasets, even tens of thousands of datasets, you can't manually check all the labels. Right. You can make some hypothesis about what the data contains. You can make some, um, priors about what the distribution of the labels are and then, uh, find out if this prior actually holds on the data set or not.

Interviewee

Uh, but this will only give you an aggregate amount of information about the data. You will not know, uh, for each of the data rose if you have the correct label or not. So that's in short the answer. But of course you can, uh, you can do some ad hoc checks for new data. You can check, uh, uh, you can check the labels by using a proxy.

Interviewee

To say that, uh, the label that is predicted by the proxy model is the same as a label that is coming in from the data, but this proxy model is obviously not the model that you are training, uh, for your machine learning system. So this could be one way to detect it. I have to say that we never use proxy models in production before.

Interviewee

Uh, it's a theoretical concept that might work, but I have no, uh, experience in whether this does work. . Perfect.

Interviewer 1

Thank you. Uh, so I'm gonna ask you a couple of questions regarding data collection. Yeah. And basically I'm gonna ask you if you ever use data that was collected in some way, and if yes, what are some of the issues you have with that, with the data collected from that data collection process?

Interviewer 1

Right.

Interviewee

Okay. Yeah. Uh,

Interviewer 1

so did you ever use data that was manually collected

Data-collection-manual

Interviewee

or labeled? Collected or labeled?

Data-collection-manual

Interviewer 1

When there is manual work

Data-collection-manual

Interviewee

in it, . Sure. Yeah.

Data-collection-manual

Interviewer 1

Yeah. And what, what were the main issues you,

Interviewee

you had, first of all, manually labeled or collected data is never enough, or it's very hard to have enough data, uh, to train a model because manual collection just takes a lot of time.

Data-collection-manual   Data-low-quantity

Interviewee

Uh, so does manual annotation. If you have manually collected data and the next step is amputation, that also takes a lot of time. So the main problem with manually collected data, There's not enough of it. Uh, and the other problem is, of course, that if you are labeling the data, uh, uh, which, uh, by, by a human, then you don't know if, uh, the understanding of the labeling task is the same that the labeler has as what the designer of the experiment has, which means the labels might be just.

Data-collection-manual   Data-low-quantity

Interviewee

Uh, and the third problem could be that if you are using a group of people for labeling a data, then there might also be discrepancies between what one person labels the data versus what other person labels the data. So that's for me the third problem. Perfect.

Data-collection-manual

Interviewer 1

Thank you. Uh, have you ever used external data, so public data set, third party, p i or web, web scrape data?

Data-collection-API

Interviewer 1

Yep. Okay. And once again, what, what were some of the main issues.

Data-collection-API

Interviewee

Um, I also have to say that the publicly available data or web scrape data or, uh, um, how do you call it? Like source the data from an external source. Uh, this can also be, uh, categorized into multiple. Kinds of data. So one could be scraping, which is publicly available, and you just get it from there.

Data-collection-API   Data-collection-external   Data-collection-web-scrape

Interviewee

And the other one is a feed data, right? So you actually, let's say, sign a contract with the data, uh, provider. And this data provider sends you a data every week or every month, what have you, right? And both of these, uh, methods have their own problems. Scraping the data means that you, um, are reliant on the, uh, quality of the software that is used for scrap.

Data-collection-API   Data-collection-external   Data-collection-web-scrape

Interviewee

Right. So that, that, that can be one, uh, that can be one, uh, kind of issue in your supply chain. Uh, the, uh, getting feeds from external sources might also have the problem that the feeds might not be as often as you, uh, expect it to be. The feeds might not always be of the highest quality, and the main problem could be that the feeds could actually change formats in.

Data-collection-API   Data-collection-external   Data-collection-web-scrape

Interviewee

So suppose you're expecting a column date, uh, and this date column used to supply you, uh, data in, uh, British date format, but it suddenly then changed into Unix date format. And this, this kind of, uh, column mismatch for feed data is also a potential problem in collecting data. Does

Data-collection-API | Data-collection-external | Data-collection-web-scrape

Interviewer 1

it happen often in your experience?

Interviewer 1

And if yes, how do you address the problem?

Interviewee

Um, Depends on what you mean by often, I guess. Uh, but it's not, uh, I mean, this is the main problem with data quality checks that you can put in some hypothesis or some assumptions about what the data would be. And if you don't do this, then you will always train the model with, uh, with dirty data.

Interviewee

So in my opinion, uh, when you. When you source data from an external feed, it you, you should always assume that you will have errors in the data frequently. Right. So in my opinion, this is a frequent problem. Right. I see. Uh, how do you address it? Um, well there are multiple, multiple ways to address it. One is that you simply ignore the data, which has problem.

Data-low-quality-fix

Interviewee

and you work with the data, which doesn't have problems. Mm. But this, uh, this method fails if the, uh, if a large chunk of the data has a problem, right? If a significant part of the data has problems, then you can't ignore it. You need to solve it. Uh, and the way to solve it would be basically to change your assumptions to change.

Data-low-quality-fix

Interviewee

Here, uh, to change your feature engineering to, uh, kind of expect this change in the model and then go ahead with the data that'll supply to you, or you completely get rid of the source and train with the old data while you're still looking for a new source of data. I see.

Data-low-quality-fix

Interviewer 1

Thank you for, for this information.

Interviewer 1

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

Interviewee

Both. Okay. And, oh, if yes, I'll. Um, have I tried to measure the quality of my data? I mean, everything that I just said about putting assumptions, uh, in place, that's how you measure the quality of the data.

Interviewee

You define what is important to you in the data, uh, by writing assumptions about it. Like this column should contain vintage values between 101,000. , uh, and then you see if this assumption holds. Uh, so basically just writing down these assumptions and applying them on the entire data set is one way to measure quality of the data.

Interviewee

Uh, the other way to measure quality of the data is how fresh it is. So are you working with the latest, uh, uh, latest version of the data or, uh, an older version of the data? This is another aspect which might count as a quality aspect, uh, and, um, . Um, how, what was the second, second question you, you asked for?

Interviewer 1

Uh, essentially where I was to is, uh, if you have any tools on or f framework to help you, uh, with. The process of cleaning and transforming your

Interviewee

data? Um, well, uh, I mean, transforming the data is a completely different topic than quality of the data, to be fair. Uh, but, uh, yes, we do use tools sometimes. I mean, uh, we use tools like Datadog or um, uh, or similar visualization tools to kind of apply some, uh, assumptions and then check if they're true or not.

Data-low-quality-fix

Interviewee

But, uh, most of the times what we do is write these assumptions by hand in a. It could be a pass script or a, usually it's a, it's a Python script, uh, to check, uh, if our assumption options is true, are true. And then if not, then we kind of raise certain alarms and maybe it could be a Slack hook, uh, that'll send alarm to a slack, uh, channel.

Data-low-quality-fix

Interviewee

It could just record in a power BI dashboard, any other kind of dashboard or whatnot. I see. Thank you. Uh,

Data-low-quality-fix

Interviewer 1

and finally, is there any other data quality should be missed that you consider?

Interviewee

Um, well actually to be honest, you, you didn't tell me what you consider data quality issues, so there is not anything that you could have missed.

Interviewee

I told you what data quality issues I consider. So, uh, this is just also my point of view, right? Yeah, I think I, I think, I think I have it all covered. Perfect. Thank you.

Interviewer 1

Um, how do you evaluate the quality of your model? And as a reminder, quality is not only defined by the ML performance, so influence, score, accuracy, et cetera.

Interviewer 1

But there are, there is also other aspect such as explainability, like I said before, explainability, robustness, scalability, efficiency, et

Interviewee

cetera. How do I evaluate all of this? Uh, well, explainability for, uh, for example, is not always important to us, even though it should be. It's not always. Um, but essentially, uh, one way that we think is useful for evaluating a model quality is that if we find in production that, uh, the model is doing wrong, predictions.

Interviewee

So if we have an input data for which we know what the label should be, and the model is giving us wrong labels for this, how easy would be would it be for us to determine why this label is not correct? Right. So how easily can I explain the reasons why this model is, uh, giving me value X and value y, right?

Interviewee

This is essentially the way to, uh, evaluate whether the model, uh, is high quality model or not. Um, and. The more, um, if you use non-linear models more than the explainability, uh, can go down. It's not necessary, but it can go down, right? But if you use more linear models, uh, then you are also likely to have models which are more explainable.

QM-explainability-issues

Interviewee

That's just the case, right? Uh, which means that, uh, if you use more deep learning kind of models, which have billions of parameters of, uh, in most cases, tens of millions, but it can also have billions of parameters there. The explainability of the model goes down, right? But also, uh, these are the cases where.

QM-explainability-issues

Interviewee

Where your performance, just in terms of accuracy and recall, inference code, as you mentioned, is usually much higher than, uh, non deep learning models. So explainability in the way that I explained to you might not even be that relevant cause uh, we are pretty confident of the results, but, That's essentially, uh, one way to measure explainability.

QM-explainability-issues

Interviewee

Uh, the other way is to how do we evaluate performance and efficiency? Well, that also depends on the task. Some, for some tasks, it's not important whether a model takes, uh, one millisecond or, uh, or one second for, uh, for inference, uh, it might be more important to us to get a correct prediction. The time.

Interviewee

Uh, and in that case is actually, the efficiency of the model itself is not that important. Uh, but there are other cases, for example, uh, if you're using a, uh, real time, uh, Computer vision application, like you have a smartphone on the, on your smartphone, you have a, uh, object segmentation, uh, app, and you need this to work in real time to segment.

Interviewee

And their efficiencies of course, important. It matters a lot. Uh, if there is a latency issue because of which the labels are coming in late or. Uh, or everything is fast enough in this case, we evaluate it differently. But, uh, in my experience has been only limited to consumer applications or APIs or web applications.

Interviewee

Uh, and I've not that much exp that experienced in, uh, realtime applications or embedded devices. So for me, the efficiency issues, uh, of models have mostly been less important than the other, uh, issue. I see. Thank you. Thank you for your

Interviewer 1

exhaustive answer. Yeah. Um, have you ever access a quality of model prediction with the user of

Interviewee

the system?

Interviewee

You mean ask the user whether the predictions are right or if the model is working Well, yes, we do. And in fact, this is, uh, this is something that almost all production systems that we develop have in place. We have something like a feedback loop from the user, so the predictions that we present with the user, um, and we want them to also tell us if these predictions are right or.

feedback-loop

Interviewee

So think about something like a search system. If they search for a keyword and we give them 10 results, then we want to ask them if the result on the first place is important to them, or the second place or the third place. And this is how we get feedback. What is not always the case is that the user is willing to give feedback.

feedback-loop

Interviewee

Sometimes they, most of the times the users don't give you feedback because they're, uh, if a result on third rank is, Is useful to them. They don't really mind so much if that it's not in the first rank. They only want to see it, uh, given back to them. So they're not willing to share so much feedback. But, uh, as long as we receive feedback, we want to use it to repay in our models.

feedback-loop

Interviewee

I

feedback-loop

Interviewer 1

see. And how do you ask for feedback? Is it something direct, like explicit you? How, how was this prediction? Or it's something implicit you observe how, how the, the person

feedback-loop

Interviewee

pa clicks or something like this? Yeah, yeah. Both actually. So we, uh, one way to do this is to, uh, to kind of, uh, calculate, uh, something called click through rate.

feedback-loop

Interviewee

So if it's a search system, then how many of the results get click through? Uh, this is of course an implicit way in your definition, how we collect feedback, but we also collect feedback explicitly by asking the users, Hey, can you just. Uh, check off this, uh, tick box by saying that this is the right answer and this is a wrong answer.

feedback-loop

Interviewee

So we do both. Combination of both.

feedback-loop

Interviewer 1

Okay. See, and what were some of the issues that were generally reported by, by the user if, if there is a wrong re strong

feedback-loop

Interviewee

predictions? Not just wrong predictions. Right. So the most common issue there is reported by the user is that the right result was either not in the list of, uh, results or the prediction was completely.

feedback-loop

Interviewee

Uh, the, the, these are the two main issues, but in like, even more important than which issues users report to us is the fact that, uh, it's very hard to get feedback from the user about it. So you have to make guesses or you have to include more of these implicit measures of selecting feedback rather than explicit ones.

feedback-loop

Interviewee

I see. Perfect. Thank you.

feedback-loop

Interviewer 1

Um, what are the challenges you have? Well, first I'd like to ask you how and where are your models deploy?

,

Interviewee

how are they deployed? They're usually deployed on a, uh, deployed as an api, uh, in, um, in a kind of app server, if you may. Uh, And, uh, you said how and where, where is the web server deployed?

Interviewee

I mean, some teams, sometimes these web servers are deployed, uh, as like standalone, uh, applications. But most of the times they're part of a, a Kubernetes cluster or a Docker file that is running on the server. On a virtual machine. Yeah. Is it on the cloud? Cloud. Yeah. Okay. Cloud.

Interviewer 1

Um, what are some of the challenge challenges you have encountered during the deployment of a machine learning software system?

Interviewee

Um, well, deploying machine learning systems is hard in the sense that there is, uh, so usually these deploy. Deployments are much bigger than other typical microservice deployments because the models can be very big. Uh, so this is a challenge, uh, in deployment. Um, but other than that, deploying of machine learning systems in itself, like conceptually speaking, is not different from deploying any other systems.

Interviewee

So I don't know if I will be able to tell you any specific ways in which deploying machine learning systems is harder. Um, Yeah, no.

Interviewer 1

Um, it's alright. Yeah. If you, if you haven't Yeah, we'll, we'll, but move on to the next question. Yeah.

Interviewee

Um, so,

Interviewer 1

um, have you, okay. How do you ensure that the quality of machine learning software system does not decrease over time?

Interviewer 1

So I'm referring to maintenance now. Yeah, yeah. Uh,

Interviewee

fair enough. I mean, uh, first is to ask ourselves why would the, uh, uh, why would the quality go down over time? Uh, and this could happen because of something like data drift or, uh, model drift, uh, which is nothing but your. The distribution of the inference time data being different from the distribution of the training time data, this can happen, right?

MM-drift

Interviewee

Uh, and the way to ensure that this doesn't happen is to do active learning. All the time. So we collect as much data from the users as they're willing to give us, uh, in terms of this feedback that I just mentioned, and use it to continuously retrain our models. This is one way to, this is basically the main way to ensure that, uh, the performance of the machine learning system doesn't deteriorate.

MM-drift

Interviewee

Uh, but, uh, in addition to this, uh, what is also important is a non machine learning aspect of it. So we also want to make sure that the, uh, servers and microservices that we've deployed do not have any vulnerabilities in the software. And if there are vulnerabilities and we update those dependencies, these dependencies don't have to do anything with the machine learning software.

MM-drift

Interviewee

They can just be system level dependencies. Um, and this. Uh, this is more often, more often than the machine learning part. This is the way to make sure that your machine learning systems are not deteriorating. I see. Perfect. Thank you. Um.

MM-drift

Interviewee

Uh,

Interviewer 1

yes. I mean, you already mentioned it a bit, but I'm gonna ask you if there's anything more to say, just in case,

um, ag issue with data during the maintenance of machine learning software system. Um,

Interviewee

Yeah, no, as I said, I mean, uh, uh, what you can notice is this shift in distribution of the data, essentially,

right?

Interviewee

And this is one way to see if the data that is, uh, being used to make predictions is completely different, uh,

from the data which we trained it from, uh, which means that our initial assumptions are no longer correct, and

then we should train a new model from scratch rather than continue to use your.

Interviewee

Okay.

Interviewer 1

Perfect. Thank you. And is there any other issue regarding the deployment or the monitoring of machine

learning software SY system that uh, we missed and you'll

Interviewee

consider relevant? Um

Interviewee

hmm. Uh, what did we talk about? We talked about the data quality issues, we talked about the deployment

issues. Um, I mean, one thing that we did not, uh, talk about are, uh, kind of continuous integration issues.

So, uh, how do you make sure that one version of the data set is, um, uh, is. Still working as expected

compared to the other version of the dataset which you trained with, or which was, uh, uh, which was, um,

kind of in place before.

Interviewee

So, and, uh, One problem that is connected with this is that, of course, your continuous integration data set cannot be, uh, as big as your actual validation set or your in, or your, uh, test set is because then your continuous integration will simply take too long. Uh, and one of the, uh, one of the issues here is to, um, uh, is to determine which tests need to be included in continuous integration, which, uh, you can call something like, um, minimum viable.

ME-testing-model-challenging   ME-testing-model-difficult-fix

Interviewee

Software system, right? So, uh, for example, if you are, if you're deploying a classification system, uh, then you should in continuous integration, have some tests which tell you that, hey, at least for this very obvious looking picture of a dog and a cat, you should always get the right prediction. Uh, In other than these, some more complex images might have back prediction, but at least for these three kind of images, if the prediction is wrong, then we do not deploy or we say that the continuous indication is failed.

ME-testing-model-difficult-fix

Interviewee

Right? And figuring out what these tests are and what classes of input are kind of crucial for the application. This is a big part or, or a big set of challenge that is associated with, uh, developing a machine learning software system, uh, and deploy. That's

ME-testing-model-difficult-fix

Interviewer 1

interesting. And how, so I'm curious, how, how do you decide which image, like the image to, to choose in your validation?

ME-testing-model-difficult-fix

Interviewee

It depends. It depends completely on your business use case. Uh, you should, uh, this was just an example, but even sticking to this example, uh, if you are, uh, if you, again, uh, if I use the example of a mobile phone application that uses machine learning to kind of, uh, segment photos, uh, this mobile phone application.

ME-testing-model-difficult-fix

Interviewee

Should not say that if I'm pointing to a window during daytime, then this window, uh, is, um, how do I put it?
So if I'm pointing to a window during daytime, then my application should not say that this part has a night
scene in it. It should say that it's a day scene in it because it's so obvious, right?

ME-testing-model-difficult-fix

Interviewee

It's just one of the obvious things that it should pass. And this just depends completely on your, uh, on your
use case, why you're deploying machine learning system. If you're having a recommendation system, for
example, uh, then for an item which is a chair, uh, another recommendation should. It contain at least one
other chair.

ME-testing-model-difficult-fix

Interviewee

It should not always recommend you something that is completely unrelated to the chair, maybe, but at least it
should recommend another chair. So this kind of business use case is what drives, which examples do you
add to it? But there's, they, they can't be a general rule for this. This is why also, it's hard actually.

ME-testing-model-difficult-fix

Interviewee

I see, I see. Thank you. Yeah.

ME-testing-model-difficult-fix

Interviewer 1

Um, so I'm gonna list a few quality aspects and, uh, basically if you read an issue with one of these in your
experience, uh, please tell me and we'll dig in, dig into the, into that story. Uh, so did you ever add issue with
one of the flowing quality aspect, fairness, robustness, explainability, scalability, data privacy, and model
security?

Interviewee

Um, well, I don't know the definition of these terms, so I wouldn't know what to say. Sure, sure. For example,
what do you, what do you mean by modern security? Uh,

Interviewer 1

so there's people that are able to, um, infer what is your model based on the prediction it makes on some

data, right? So they try to steal your model, uh, this way?

Interviewee

Yeah. Basically it's called like, uh, proxy model training, something like this, right? Yeah.

Interviewer 1

You, you're just trying basically everything that when you try to steal the model, this is what I, I'm referring to,

uh, by

Interviewee

model security. Okay. Yeah. Yeah. This is not experience actually. Perfect. Do you want

Interviewer 1

me to go back on the first one?

Interviewer 1

Yeah, I can. Okay. Yeah. Uh, so furnace is when you have group, uh, sensitive groups, let's say men and

women, and you want your model to be fair to both groups. Uh, robustness, it's out, it end all, um, like outliers

or, or points out that are, um, like difficult to. , uh, explainability. You know what it is? Uh, scalability.

Interviewer 1

We're looking for a system that scales. So anything, uh, more data or more system or any, anything regarding

scale and data privacy. Well, uh, uh, data purpose,

Interviewee

right? Okay. Yeah. Um, , I mean, scalability, scalability can be a problem depending on your use case. Again,

so as I said, usually for the use case that I was working on, efficiency was not that big of a concern because

they were all web based applications or test stop applications.

Interviewee

Um, and, uh, fairness of the data could be a problem, but we never worked with, uh, or none of our use cases were related to personal information. They were all related to, uh, either. Demographic groups are actually non-personal information. So information about things rather than people. So this was also not a problem that I've faced.

Interviewee

Uh, privacy, uh, was probably an issue for certain models that we trained because, um, uh, you, because there are certain. There are certain features that you can get from the data that are important for prediction, but if you use them to train a model, it also reveals about somebody as a person. So you are not actually allowed to use these without their explicit consent.

QM-data-privacy-issue-fix   QM-data-privacy-issues

Interviewee

This is, uh, uh, so I worked very closely with the GDPR e law in eu, and you need their express consent if you wanna train a model on this data. So this, this was. Often a problem. And to solve this, what we did was we created synthetic data, or we created, uh, some kind of, um, uh, key anonymization concept on top of the data, which could kind of anonymize the identifying part of the information and then train a model with the anonymized data.

QM-data-privacy-issue-fix   QM-data-privacy-issues

Interviewee

This is another way to do it, but uh, yeah, this could be a problem. . Okay. And, and

QM-data-privacy-issue-fix

Interviewer 1

what happened if someone tells you that they don't want their data to be in their database anymore? So you have to retrain

QM-data-privacy-issue-fix

Interviewee

No, no. To be, to, to make it clear, uh, we, we do not have the capacity to ask everyone for consent if we have 10 thousands of data.

QM-data-privacy-issue-fix

Interviewee

So we don't even train with that data. What we do is we anonymize the data, so all the I then personal identifying information that we have in the data is no longer personal identifying. And then we train with the privacy preserved data set. Okay. I see.

QM-data-privacy-issue-fix

Interviewer 1

And, and, um, that, that's synthetic dataset, um, is it not generated by the real data?

QM-data-privacy-issue-fix

Interviewer 1

So,

QM-data-privacy-issue-fix

Interviewee

So, no, um, uh, no, not completely. It's fair. It's ok. Yes, you're right. If there is, if there is a function that can convert, uh, uh, that can generate a synthetic data, then there should be an inverse of this function, which can generate like real data, theoretically speaking. But this inverse of the function, uh, is, um, is very hard to create then.

QM-data-privacy-issue-fix

Interviewee

Right. It's, uh, it's not tractable. You're talking about inver. , very large matrix. That's not easy. Uh, so, uh, the, these, um, not anonymization, sorry. This synthetic data generation is usually used by looking at distribution of certain data, uh, in another dataset, but creating something synthetic that is not that dataset and then training using it.

QM-data-privacy-issue-fix

Interviewee

But, uh, as you can imagine, this could also mean that the machine learning model that you generate using synthetic data is actually not good on the real world data. It's a, it's a big.

QM-data-privacy-issue-fix

Interviewer 1

I see. And, and can you tell me some of the tool you, you, you or your team have used for

QM-data-privacy-issue-fix

Interviewee

that purpose? No, uh, we basically just use some libraries in Python and Art to generate the synthetic data, but that we can't use any out of the box tools per se.

QM-data-privacy-issue-fix

Interviewee

Okay. Okay. Nice. Perfect. Well, thanks.

QM-data-privacy-issue-fix

Interviewer 1

Uh, I have two questions for you and then we're finished. Uh, in your opinion, what is the most person quality issue researchers should try to solve or any issue you'd like to seek? Go away.

QM-data-privacy-issue-fix

Interviewee

Basically. Um, I mean, I would like to obviously see all of these issues go away.

Interviewee

The question is what is the most pressing one? And the most pressing one is the data quality one. I don't, uh, I don't think it's possible. It's impossible, and nobody should work on removing this problem because it's, it is impossible. You cannot, uh, get rid of data quality issues per se. What you can do is generate visibility on data quality.

Interviewee

So if, uh, if there are generic ways to convert, uh, business use case problems to something that can be visualized or something that can raise alarms, this is the most pressing need because, uh, right now all business use cases and all real world use cases create these assumptions by the. And then they kind of generate some kind of metrics on it.

Interviewee

They generate some kind of alarms on it. But the most important thing right now is to create visibility on what the data quality is, not get rid of data quality. This is a futile problem, uh Right. Uh, and if this is some direction that, uh, researchers can go into, this would be my pick. Interesting.

last-question-answers

Interviewer 1

Thank you.

Interviewer 1

Yeah. And do you have any other comment about the quality of machine learning software?

Interviewee

No, I think, uh, I think we should think more about writing tests that are specifically for machine learning software systems. There's so much literature and so much practice out there that is related to testing of traditional software systems, but not enough machine learning software systems because this, uh, this is not something that software engineers can do, but they need help from the business colleagues to write this and this.

Interviewee

Some software engineers do not like to talk to people. They like to work in isolation, and this is one of the main reasons why the data quality tests and machine learning system tests are not as easy to find out there than traditional software tests. So this is, this would be like my closing comment if you may.

Interviewee

Super.

Interviewer 1

Well, thank you for being there. I, I think you said a lot of interesting information that will be useful for our

study. Uh, so yeah, I wish you

Interviewee

a good day and appreciate, thanks very much. Thanks for your time as well and all the best with the research.

Yeah. Uh, I will send this sign sheet, uh, maybe next week if you can please remind me again to send it.

Interviewee

That will be great. Yeah, it's perfect. Uh,

Interviewer 1

and just before, uh, for demographic purposes, can you tell me the size of your startup or the company? Uh,

between 50 and

Interviewee

hundred. Okay.

Interviewer 1

Okay, perfect. And well, can you remind me, you are doing an API for,

Interviewee

uh, your, uh, well, we are building several APIs and not only just APIs, but it's all related to something called

neural search.

Interviewee

This means that you want to create search systems using machine learning embeddings rather than just

back, which, just words and gift. Okay. I

Interviewer 1

see. Perfect. Uh, neural like, uh, neural research.

Interviewee

Exactly. Alright. Alright. Nice day all.

Created with the Delve Qualitative Analysis Tool (http://www.delvetool.com)