Interview 16 - PO

Interviewer 1

. All right. Uh, so we'll start with a brief description of what are the goal of the interview. In the end, we want to develop a catalog of quality issues in machine learning software system. The first thing to understand is what is a quality issue? There's two word, uh, and well quality. What is it? Um, imagine you have two system, two software system that accomplishes the same.

Interviewer 1

Well, if you can say that one is better than another, generally what you're referring to is quality. So quality issue is something that the system that is worse may have. Right? Um, and we are not just only interested about quality issue, about the machinery software system, but about you can be a bit more general.

Interviewer 1

Meaning if you have issue encountered anywhere, uh, you can tell us and maybe we will be able to trace it back to our, uh, topic. And a machine learning software system. It is really simple. It's a software system that has a machine learning component in it, right? Sure. Um, so if it's clear, uh, I will move on.

Interviewer 1

Uh, we have approximately 20 questions to ask you. Uh, and we have question for each phase of a machine learning workflow. So getting data, cleaning it until, like deploying it and everything. We do not expect you to have experie. everywhere. So, uh, if you do not have, there's no problem, we'll just skip some question on some section, right?

Interviewer 1

Sure. So if you're data, data scientist, maybe you have less experience in like maintaining model if you're a software engineer, maybe it's ML engineer, maybe it's the opposite. Mm-hmm. . All right. Uh, so that, that's all for me now why I will ask you some question. Um, so, um, , can you give us some background information about you?

Interviewer 1

Like what's your current position, how much experience you have, general or specific to machine learning?

Interviewee

Yeah, so I am a machine learning engineer at, uh, Company X and e-commerce company. Uh, so on, on my day-to-day, I'm not necessarily involved directly in the model training phase. Um, so more like on the pipeline production, I think serving it runtime.

Interviewee

Um, so that's probably where I might be able to offer more insight into. Perfect.

Interviewer 1

Thank you. How long has that,

Interviewee

has it been. . Uh, so I've been at Company X three years, but before that I've worked at two smaller startups. Okay, great. Thank

Interviewer 1

you. Um, so we'll start with the first general question. Uh, what are the main quality issues you have encountered with your data model or system so far?

Interviewee

Sure. Um, so you said data system and what was that again? Uh, data. Okay. Yeah. So for better or worse, like in our current org structure, like for our data, like, you know, like users interact with Company X search system and like beacons are sent. So, um, we have other teams that kind of deal with, like stitching that data.

Interviewee

So we don't necessarily, um, have, you know, like the. issues of like kind of instrumenting that whole pipeline. Uh, but we have had issues in our data where maybe there's a new feature that's being billed and you know, like the beacons that are sent, they're missing some values or something. So it's incomplete in that sense.

Interviewee

So we can't properly attribute a feature or, or maybe something is like there has been a bug in like, uh, an attribution code, so then, you know, We need to decide whether we need to backfill it for the whole year or so. So usually most of our models or features, we usually try to have at least like a year of data.

Interviewee
Um, because, you know, it's like, E-commerce, kind of like a cyclical kind of, uh, nature. So, um, so those are usually kind of the data issues that are there, but luckily it's usually another team that kind of manages that. But those are issues that I've definitely brought up to the team to kind of help fix, um, Yeah, so assuming like, you know, we have this nice or semi nice clean stitch data where, you know, this is the session id, this is the query, these are the things that a user has interacted with.

Interviewee
You know, we have standard pipelines to translate that to kind of label data for models for an e-commerce setting. Um, going specifically for like quality issues for models, um, , like I think we definitely do not do the best job with maybe things like model monitoring, like data drift or um, like definitely when there are big things like covid or maybe just like a change in user patterns.

Interviewee
Like I think in production we are not necessarily catching it really well unless it's surfaced by like other product owners or just. things that people have found. So definitely that is one area we are not the best at. But we do frequently train our models so like at least every, like, people are constantly iterating, so we do a lot of experimentation.

Interviewee
So, um, and so usually like every three months or so, like a model in one of the core stacks do get promoted. So, , um, if there are any issues they kind of get hidden away by the or? Yeah, by these frequent model iterations. Um, uh, system. Well, I guess like, um, I mean I already mentioned about like the not really doing model monitoring from a system piece.

Interviewee

You know, like since the model is just a piece in a larger service, you know, we do have pretty good. Good information about like, um, you know, the latency for a model, if it's like a realtime inference or like how much is this feature slightly missing for most of the data or whatnot. So like those runtime instrumentation, like we do have it published, but.

Interviewee
They're not necessarily maybe acting on it per se. Um, so I maybe said a lot in all these things, but maybe I could dive deeper into any specific point or

Interviewer 1
No, it's perfect. Thank you. That's the point of that question to go really broad in general, like to bring some ideas and then we go deeper as well.

Interviewer 1
Perfect. Yeah. Uh, so you mentioned the well data integration issues. So you said we need to stitch a data set. , could you go a bit deeper on, on that? Do you have any tool to help you stitch the data? Like are the issue you have encountered because of stitching data with the dataset you, you obtained?

Interviewee
Uh, yeah.

Interviewee
I mean, not necessarily any fancy tool, just spark, um, um, but I think like the stitching is like we have be, um, like we have three main platforms, web, android, and mobile, and. Because there's like a different user interface basically. Um, like from my current team, which is the search, like we, we basically care about like, like a useful event for us as like a user type.

Interviewee
This query, they saw these items, they interacted with this item, they applied so and so feature, like we basically would ideally like a record like this, but you know, The beacons that are basically sent is, okay, the user clicked this page, the user landed on this page. So like these individual beacons need to be stitched.

Interviewee

Um, and so maybe in some platform, you know, an event is not properly triggered or, you know, variation of the e um, event names. So those kind of data quality issues. But again, that's mainly handled by another team, which is the consumer of. .

Interviewer 1

Okay. And, and just to, to be clear, when you, you, when you're talking about stitching data, you mean fetching data?

Interviewer 1

For example, you, you had the, a user that clicked somewhere. Mm-hmm. , when you, when you're talking about stitching, you, you're talking about fetching other information that is complementary to the user as click,

Interviewee

right? Yeah, exactly. Like, so for, um, like for us, The main thing might be like a session ID or something, or a customer id.

Interviewee

So, um, ideally at one record, we would like all the information that is relevant for that session. These are all the queries that the user interacted. These are all the metadata for the query. This is, and in the end, Did the user actually complete their journey? What were the items they added to cart? So all those things across, you know, the 20 minute or 30 minute inter session, however, user spends, we would like to kind of stitch that in a holistic view.

Interviewee

Okay. And do

Interviewer 1

you know, how do these team, how does the, sorry, how the people in this team do to stitch data together? Do they manually filter data from one source? So it's a pipeline that merge everything. .

Interviewee

Yeah, there's several pipelines. Yeah. Perfect. Spark based pipelines. Yeah, I understand.

Interviewer 1

Do you know if there's a lot of issue with these pipelines?

Interviewer 1

Like the data changes and they always have to update?

Interviewee

Okay. Um, I mean definitely just like, maybe like they, like the team that we interact with, they're also consumers of like a larger Company X kind of instrumentation. So, um, there is a larger. A company level to kind of standardize some fields, but it's just like not, maybe all the themes are following it or Definitely we have had issues with like the three platforms, iOS, Android, web.

Interviewee

It's just like, um, it's not consistent, um, the events that are filed. Okay. I see. Okay. Thank you.

Interviewer 1

So you mentioned you use data generated by some system, like the user click somewhere and it generates an event. Mm-hmm. , which creates data. Uh, so you, you, you have ingested data created by another system. Now, I, I have a question. Do you ever use data generated by an external external source, like a dataset, a third party API or web script data?

Interviewer 1

Um,

Interviewee

So like we have used, like, because we are in a search, um, space, like we do submit data to like third party vendors to get, like, get the results like annotated. Um, so from that mention there, um, we do kind of, uh, kind of scrape like how our competitor returns results for certain queries. . Yeah. I mean those aspects there.

Interviewee

Yeah. Okay.

Interviewer 1

That's interesting. And yeah, go for it.

Interviewee

Uh, but I'm not aware of like any, or, or I guess maybe like, you know, if we are implementing like a new feature, like, um, uh, like, you know, like maybe Company X doesn't really support like Spanish or something. Right. So to get kind of labeled data for like a English query to a Spanish, you know, we may use.

Interviewee

One of the cloud providers or something like that to kind of get, um, some ground data. Yeah. Okay. I see.

Interviewer 1

A And when you use third parties, uh, to obtain data, do you have any issue with the data you retrieve from them?

Interviewee

Um, I think, um, I think that issue maybe is more like the third party might have like a model.

Interviewee

that is building, like, um, it's a more general model, but like for us, we are in an e-commerce setting. So, um, like the predictions that are returned are not completely quite right for us. So maybe like a simple thing that we tried is we tried like, uh, like bing spellcheck, but like, or Google Spellcheck, like, Because we are an e-commerce setting.

Interviewee

When someone types something, like the suggestion should be in an e-commerce kind of scenario, not like a general, like a spell spelling thing. So just like a out of domain kind of, um, context that's missing. I mean, yeah, domain is, domain context is missing from like sometimes these third party, um, services.

Interviewee

Yeah. Okay. I see. Thank you.

Interviewer 1

Moving on to data preparation. Uh, which data type have you worked with?

Interviewee

Um, I mean, . So I guess for most of us, most of it, or like our data is in a structured format like, uh, hive, where, you know, kind of the format that we mentioned, like query session information, all that stuff. Um, we also have like data that's related to the Company X catalog, so it has.

Interviewee

Product id, you know, item title, item attributes, image, url. Um, so yeah, those are the two main things that we mainly work with. Okay, cool. And also, I guess event, I mean, maybe this is related to first one, but like event logs. Um, but yeah. Yeah.

Interviewer 1

Perfect. Thanks. Uh, have you ever measured the quality of your data and or tried to improve it?

Interviewee

The quality of my. , um, yeah, how we measure the quality of our data. Um,

Interviewee

so we have definitely, um, the quality of our data. Hmm. So like we do have things, I mean, this is not maybe directly measuring the quality of the data itself, but like the output of our data. Like, you know, like the, if we say like, these are the top end queries, or these are the top things that users are interacting with for certain query, like we do send them for manual review.

Interviewee

Um, so things do get flagged there. . Um, and obviously when models are being billed, like the engineer who's building it, when they're doing the feature collecting stage, like they will see some issues there. Um, there are, like, again, since we, our team is not directly involved in the data quality piece, like there are sanity checks

about like certain row counts session count, uh, certain number of queries and all.

Interviewee

um, things like that. But yeah, again, since we are not directly involved at that stage, yeah. Okay. I see. Thank you.

Interviewer 1

Um, is there any other data quality should we missed that we, you consider relevant?

Interviewee

No, I think, yeah, yeah, you can, since we are not directly Yeah. Creators of that. . Thank you. Okay.

Interviewer 1

Uh, moving on to model evaluation. How do you evaluate the quality of your model? And as a reminder, quality is not only defined by ML performance. Uh, there's other, other, also other aspect such as explainability, robustness, uh, scalability,

Interviewee

and, yeah.

Interviewee

Uh, yeah. So, um, so the performance of our models like definite, I, I. Like, um, sorry for e-commerce, like we mainly use like three based models and then we have some deep learning models. So for the three based models, which is mainly where we get to experience the, I mean that's the, most of the three based models are the consumers of those, uh, kind of session based data.

Interviewee

So, For that, definitely. Uh, we, you know, training time, we do look at, um, explainability, uh, like for actually boosts, you know, like the, what are the top end features. Um, and, uh, we, the features are, uh, maybe a little bit complicated, but it, we do expect certain features to be on the top. Like, especially like if you're training like a engagement based model, like we.

Interviewee

Certain engagement based features to be on the top. So, um, so there is like that sanity check that's there. Um, for any feature also like to launch, to like production, like we do test it, we test it several ways. Like we, again, we send it, we generate like, um, predictions for the model and then send it for human annotators to evaluate.

Interviewee

Kind of, you know, the ranking that's returned by the model. Uh, we, uh, so that's like to get like N D C G kind of evaluation, um, at runtime, you know, there's certain business metrics that we care about. So we see how if one variant versus standard, like how the results would affect the business metrics. So, um, so that's, , we do it like for, we start with 1% of the traffic.

Interviewee

Um, and then when it kind of meets, when it shows a certain lift, then we do a proper AB test. Um, so that's how we test that. Um, regarding scalability, um, one of the reasons we do, uh, focus on tree based models is because it is, um, much faster, like in a online. , um, to meet, to meet like our latency, s l a. Um, so, um, yeah, we do have like deep learning based models recently.

Interviewee

Uh, but that's because we now have g serving at our runtime. But still, most of our models are three based models and simple, like linear classifiers.

Interviewer 1

I see. Great. Thank you for, for your answer. Um, , have you ever used existing be benchmark models for quality aspects to evaluate your model? So for example, you may maybe use baselines, uh, to evaluate some models

Interviewee

mm-hmm.

Interviewee

Um, yeah. So, um, like for the deep learning based models, like we have always started from a, like a pre-trained, um, weights and then fine tune it on our custom data set. So, , uh, we compare. So yeah, we compare the baseline as the pre-trained versus then our fine-tuned. So we definitely use that. Um, but, um, we, we do compare, like, again, I'm not directly involved in all the model training, but like we do compare like some of the approaches against like if there's like.

Interviewee
Like a K D D or like some research data set, like, um, like the people who are involved directly in the model training, they do test their approach versus like on that published data set to see. But in general, like we mainly focus on like our internal data sets. Okay. Okay. Perfect. And just to clarify also, like, um, in order to launch anything.

Interviewee
Like, it's really like those business metrics that really people put a lot more emphasis on. So, uh, we do need to show, like an AB test. There's a lift, so. Okay. I see. Perfect. Thank you.

Interviewer 1
Alright, and do you have any issue with the metrics? Like how do you choose, do you know how they selected which metric? Uh, that should be used to evaluate the model, like lift.

Interviewee
Um, yeah, I mean, like, since we have like a e-commerce setting, like, you know, things in general like add to carts and orders are, um, like the most obvious choices. Um, but then there are maybe other, um, things that, you know, like in a search page, it's not just like the search ranking there, like maybe other things like ads or like, um, like third party kind of.

Interviewee
Modules. So we have metrics to make sure if engagement on those are being affected. And obviously things like session abandonment rate and all those things track. So even if our, even if our, um, model is really affecting the ranking, but if somehow, It's causing the time to render a page to be slower or something like those other metrics would be affected and that would affect, um, our launch.

Interviewee

But in an offline setting, we do have like offline metrics, like N D C G, we do have like, um, offline simulated at cots and those kind of offline things. But again, the online metrics are what really affect the launch.

Interviewer 1

Okay, thanks. , um, have you ever assessed the quality of an ML model with the users of your system?

Interviewee

Um, so, um,

Interviewer 1

yeah. In your case it might be a bit more

Interviewee

difficult. Yeah. Yeah, I mean like, so hopefully like, uh, we itself are consumers of our own, uh, search systems. So, um, we do have like feedback channels where people do share, um, issues, uh, where the results don't make sense. So then, um, then we do have like several in internal debugging tools to kind of, um, Expose some, um, internal information.

Interviewee

Um, one problem though is like, there are like several models, there are several features. So, um, uh, being able to properly explain all those things, um, like where an issue might be there. Uh, it takes a lot of people to be involved. Um, definitely since we end the search team, like. There's certain type of features that most, like most people are aware about.

Interviewee

So, uh, because they've been there the longest and they, um, it's obvious like if maybe it's an issue is coming up because of that, like maybe a feature is missing or maybe we are filtering out because something is missing. Like, um, some features missing, like those things are clear. Um, but maybe some other things that go this ranking or results don't really make.

Interviewee

Um, it's a little bit difficult. We do have an internal tool that kind of exposes most of the scenarios, but it still needs some work. Okay, I see.

Interviewer 1

Thank you. Uh, and have you encountered any other quality issues during the evaluation of your model?

Interviewee

Um,

Interviewee

I am not sure if this would be considered a quality issue, but like we do have scenarios where like the offline metrics and online metrics sometimes don't quite match. Um, like we might see a lift when we did maybe an interleaving, but during an AB test, the lift is negative or neutral. Um, Maybe that implies an inherent quality issue, but, um, if we haven't necessarily dived deeper into that.

Interviewee

Um, the other thing is maybe it's a data quality issue. Like we do have like features that are updated every day. Um, so. , like, um, if there was an issue in someone's feature or pipeline or something, like, we don't necessarily have like a snapshot of that, uh, to kind of, to go back and, or this was the issue that day because of that.

Interviewee

Um, I mean, it doesn't make sense to necessarily store like a rolling snapshot, so, so if their issues it, like the system is kind of naturally evolving and. Most of the things are evolving at least a daily day frequency, but some are evolving like in a, every, um, every hour or so. So it's hard to, maybe if there's an issue, sometimes it might get silently fixed.

Interviewee

Um,

Interviewer 1

okay. And the issues you're talking about, is it like someone fix a data pipeline and you do not have any more diversion of the code, which had the bug or it's the dataset, like you don't have a history of the data?

Interviewee
Um, actually it's kind of bought, so like, um, so one, someone might have fixed an issue, so like now that feature has the correct value.

Interviewee
The other thing is like maybe, uh, that pipeline had an issue and it was generating an intermediate data set, but in the end of the, at the end of the pipeline, they only save the final data set. Um, so the older version. , um, it's kind of lost. And also like the catalog, Company X's catalog changes every day.

Interviewee
Some items becomes deal listed, some items come new. So, um, yeah, so like if you save these intermediate tables, like, uh, it's just too much data to save. Okay.

Interviewer 1
And, and why do you need the intermediate representation of the data set, uh, for your.

Interviewee
It's an honest question. Yeah, yeah. Like, so one thing, like for embeddings or something like we, um, like, like for text embedding or something, like, we have had issues where, uh, like the product title or product type or something like, it's just somehow, especially like third party, um, products, like the inputs are not quite right.

Interviewee
So, Like, like maybe keyword stuffing or they copy and paste it from something else or something. So like the embedding that we generated might be wrong, but then, you know, maybe they later updated and the next day it gets corrected. So, um, we don't really have a history of that. I mean, we do have maybe for a week at least for this specific scenario, but.

Interviewee

No one really goes back and looks at it unless someone escalates an issue.

Interviewer 1
Okay. That's really interesting. So what I understand is, uh, you receive the data, it, it gets transformed in the, in the pipeline and at some point you may have an issue and then you save the final version of the dataset on a feature store.

Interviewer 1
Mm-hmm. or something like that. And that final version of the data set is in a format that cannot. Understood or, or debug. It's, for example, Deb embedding everything is, is merged together. So you cannot see where's the issue in the embed embedding, is that right?

Interviewee
Yes. Yeah. Yeah. I think that's

Interviewer 1
good. Okay.

Interviewer 1
Really interesting. And how do you end this kind of issue? Uh, generally, uh, it's a difficult question, but how do you proceed? Yeah, go for it.

Interviewee
Yeah, I mean like, um, in some cases, like for this scenario, we do save like at least data for a week or so. So, um, if someone brings an, um, if someone brings an issue with an embedding, at least we can see that, okay, this was the intermediate data we fed into the model.

Interviewee
Um, so that's one way, but um, Again, sometimes these issues are brought by, uh, like our product ops teams or like other users inside Company X kind of searching. So it's not like directly noticing, but it's the, we only able to properly debug this is because we kind of save what was the input that was fed to the model, but we

don't always do that.

Interviewee

Okay. .

Interviewer 1

Right. Then how, how did you detect these issues? Like the embedding, it is problematic. How, how can you see that the embedding as a

Interviewee

problem? Um, yeah, so again, it's more like, um, so like, um, you know, someone enters a query, um, then, you know, we compute embedding, it goes to our a and n feature store, and, and then.

Interviewee

you know, the ranking that's returned from it, um, it looks kind of off. So like someone may search for something, they escalated and then we do the debugging to find out. But there's no natural way or like a systematic or automatic way where it's like, oh, are these qu Like, are we returning embarrassing results for these queries?

Interviewee

It's more like a, um, It's more like an ad hoc stuff. We, we did have like a, I think it was like a monthly pipeline where like for certain top queries, like where we ha like we crawled our results and we sent it to, um, like a third party human annotator to kind of evaluate and show results. So we did, we do have like a monthly.

Interviewee

Kind of average score of our system to evaluate how Val's doing. But again, that's at a higher aggregated level. Okay, perfect. Thanks a lot.

Interviewer 1

Um,

Interviewer 1

yeah, I'm sorry I lost track of where we were. Uh, have we touched upon, upon model deploy? No, we haven't touched, no, we haven't here. Right. Okay. Thanks. Thanks, Hasha. I see you unmuted yourself. All right. Um, so what are, what are, you mentioned earlier on, you mentioned some challenge with modern deployment.

Interviewer 1
You said, uh, the metrics online and offline were not, uh, the same. Right. And you mentioned if, uh, so what are the challenges you encountered when you're during the deployment of a machine learning

Interviewee
software? . Yeah, I mean like the, um, me, the metric mismatch is not necessarily an issue of like the model deployment.

Interviewee
Like, so we have like offline models and online models, um, like for generating the, some of the features that we use at runtime, like those thick, those can be offline features generated and then fetched, uh, from the feature store. Um, , I think I mentioned most of the models since we are latency sensitive data, like three based models or linear, uh, simple linear models or like if they're fancy models like neural nets, they're served on GP use.

Interviewee
So we have like infrastructure now to kind of deal with that. Um, but so, It definitely took us a while to get there, but I think model deployment, either offline and online, has been standardized a little bit. Um, but before we kind of standardized, you know, we had things where like, you know, maybe people were not like pinning versions of the, especially like the Transformers package, like it has gone through so many version upgrades.

Interviewee
Uh, and some, there were some issues with upgrading it. Um, , um, or there were some issues that were silently cost because it got upgraded. Um, and I think we do have like a couple of models in production where the versions are not properly pinned. Um, so, um, and also like the person has left, so, but those are like small models.

Interviewee

But for newer models, we do a better job of like pinning, keeping like the training data, keeping the. The training log. Um, and um, also having like a runtime test to make sure like the, for certain queries or something like the results are kind of what are expected. So, yeah. Okay, great. And,

Interviewer 1

uh, just I have two questions and the first one is, do you have any tool to automatically check if someone wrote the version of the transformer they want or is it something you have to manually check each?

Interviewee

Yeah, we don't have any like, um, um, automatic tooling for that. Uh, there is a slightly manual process and we do need to involve another team for promoting the models to a higher environment. So they are like our safeguard in that, but there's no like, automatic tooling for that. Okay. Perfect.

Interviewer 1

Thank you. And you mentioned online and offline?

Interviewer 1

Um, yeah. Can you clarify what is offline? Uh, in my head offline is when you, you just test your, your, uh, your model on a data set on your computer. Uh, but I, I think I understand that it's something different in your case,

Interviewee

offline. Uh, yeah, like, uh, so, um, you know, like for a three based model, like there may be like 50 features that are needed for the tree based model.

Interviewee

Um, but like, Most of the features, like whether it's a query level feature, an item level feature can be generated in an offline setting. So like maybe an item embedding like, um, you don't need to compute the item embedding in an online, like when a request comes in, you can compute like a. item embedding completely offline, and you can use like a fancy model and all that kind of stuff, but in an online setting, you can just fetch the embedding.

Interviewee

So, um, yeah, so like in an online, the models that are serving online traffic, uh, like I think we do a very good job of like pinning the versions, you know, unit tests, regression tests, all that, but models that are in an offline setting since. . It's like any developer can, has the flexibility to play or try anything.

Interviewee

Like, so not all the best coding practices are followed in that setting. Yeah.

Interviewer 1

Okay. Interesting. And you just mentioned unit test, regression test. How, uh, which test are

Interviewee

you running on the model? Yeah, so it's, it isn't necessarily like on the, um, model per se, but like I guess with the model there may. Some packaging on top of it.

Interviewee

So like, so it'll be like, okay, for these, um, like for these, um, like if it's like a classification model, like for these queries, um, I expect, you know, this to be the top prediction, this to be the score to be in a certain range. So, uh, and if there's any upgrade or whatnot, at least we know that something is off.

Interviewee

Okay. Yeah. Perfect. Thank you.

Interviewer 1

Uh, so we were already talking about model maintenance, and I have a few, few question about that. Uh, so how do you ensure that the quality of our machine machine learning software system does not decrease over time? Mm-hmm. ,

Interviewee

um, Yeah, I, I feel like this is an area that we are a little bit poor at.

Interviewee

So, um, we started recently doing like, uh, feature logging. So like for a tree based model, we do know, we do now know that, okay, for this request, these web, uh, online features that were computed. Um, and so we have all those things logged. So, So this could be a way to identify, okay, which features are not really adding value, which features are not really used cuz we have like so many features that are, but we haven't necessarily taken any proactive action on that.

Interviewee

Um, other than like finding out, okay, maybe these things are probably. A candidate for decommissioning, but we haven't necessarily looked at, okay, for this feature that all the scores are in this very small range. So maybe it's not normalized, maybe it's not adding much value, but we ha we, except for discussions, we haven't really taken any action on that.

Interviewee

Um, again, we do, so we have the features and then we also have, like for classification models, we do again log the. Confidence, um, scores and the label that we emitted, but there's again, no proactive action taken on that data yet. Okay.

Interviewer 1

I see. Thank you. And why do you want to, oh, sorry.

Interviewee

Sorry. One thing maybe I do want to mention, which maybe we are doing a little bit better about is like we do have like, because I mean, we have so many models and they may be interacting.

Interviewee

Together. Um, so from a system level, we do have things like, which queries like we do have like alerts on, like these are queries that we are getting a lot of traffic, but users are not converting. So, um, we like, so from a business perspective we do have those things, but, um, it could be a model issue or it could just be some other.

Interviewee

But yeah, we do have at least that alerts. Okay, I see.

Interviewer 1

Thank you. You mentioned, or earlier on that sometime you decom decommission some feature. Uh, why is, why is so, why is that?

Interviewee

Um, um, this is more like from a production thing, like, so, um, again, like latency is very important for us. So like, if. Feature doesn't need.

Interviewee

If a feature's being fetched but not really used at runtime, that's just like extra load or, yeah. Okay.

Interviewer 1

Perfect. Thank you. Um, I will go quickly because I see we are at the end of the meeting. I a very encountered issue with data. Uh, no, sorry, we already covered that. That's, that was not the whole sentence.

Interviewer 1

So, but yeah, we already. Uh, have you ever have you I'm trying to go fast and now I can

Interviewee

talk anymore. No, I, I can stay a little bit later too if needed. So yeah, feel free to take your time. Okay.

Interviewer 1

Five, five or 10 minutes and we're finished. Thank you. Um, have you had other issue regarding the maintenance of your model or system?

Interviewee

Um, Like, uh, I think like before we moved our models to Docker, uh, like we had all those version dependency issues. Um, we do have now like C I C D that like tells us like, okay, your, the base image has like a vulnerability or like, um, these packages need to be updated. This is a more recent addition. Um, we

also, Since we have kind of standardized most of our model process, like, um, so like we have a specific, you know, a model is the code and a model binary and there's a certain version to it.

Interviewee

Uh, but to upgrade, to release, like, uh, if there's like a patch fix or something, um, like we do, we do need to do. A little bit intense validation of like comparing the old model versus the old model, like seeing if there's any difference in the ranking or whatnot. So, um, yeah, so that's, that has been like the, so model maintenance has been a very pretty manual process and a slightly newer process in our old.

Interviewee

Okay, great. Yeah, go for it. The, the more frequent thing that happens to is like every three months or so. People are introducing new features and that indirectly replaces the model. So that's usually what happens. More than like upgrading a specific version of a model. Yeah. It's like, okay, a new embedding is trained, or like a new re ranked model is trained.

Interviewee

That's usually what happens more. Okay, I see.

Interviewer 1

And which dataset do you use when you compare to. ,

Interviewee

uh, yeah. So, um, when comparing, like, um, when comparing, um,

Interviewee

yeah, so I'm just thinking of how to say this. So we do have like golden data, you know, that either we crawl from our competitor or we crawl from like, I mean, we got like human annotated data, so, so we have that data set. So when someone is trying. A new approach. Um, they, they're comparing against that. But um, I mentioned like there's frequently like model retraining or model, like trying a new model like, um, that also occurs on like, you know, the last three months or the last year.

Interviewee

So data, so it's like a rolling window that kind of happens. So maybe it's a little unfair to compare a model that was trained on slightly older data. . Um, we do have, you know, metrics that are there, but again, all models, like the real, um, the real comparison happens in an online setting, uh, like with an interleaving or an AB test.

Interviewee
It's just that, uh, there's a lot more competition in getting your model tested there. That's why we go through like the offline process first. .

Interviewer 1
Okay. I see. And just quickly, you mentioned, um, data labelers. Did you have any issue with the, with these data, collect well with this process in the data or usually it's fine.

Interviewer 1
Um,

Interviewee
so we have done, um, this process for like, there's a certain type of model, like a ranking model, which we mainly engage with, uh, third party annotators. Um, so we have given them, Strict guidelines there. Um, so, and also like for every item that we ask them to label, we, we get, uh, we show it to three labelers.

Interviewee
So we get, uh, average from that. But, um, definitely like, you know, maybe if we are onboarding new labelers, like we have seen things where you. , like it's not consistent in labels that we are getting, um, sometimes, but that's like for the newer, newer teams that are onboarding, especially doing like the busy season where there's a lot of data that there's a lot of models that are trained and we need a lot of feedback.

Interviewee
Um, but usually the, um, service that we normally work with is, has been pretty standard or at least consistent, rather. Yeah. Yeah.

Interviewer 1

Okay. Perfect. Thank you. Last section is gonna be real quick. Uh, so it's about quality measure of ML model. Did you ever add issue with one of the following aspect? Uh, fairness, robustness, explainability, scalability.

Interviewer 1

We already covered it a little bit. And privacy. A any one of them? Yeah.

Interviewee

Um, sorry. Fairness. Robustness. Yes. Privacy and, uh, ,

Interviewer 1

maybe we can talk about privacy or, or security.

Interviewee

Yeah. Um, yeah, I mean, like, at least the models that in our org that we work at, um, like they, you know, most of the inputs are like a query or an item.

Interviewee

Um, we may have like session level info maybe, but we don't, like we get. Session ID And these are the items that you interacted with. Like we don't really get any consumer customer information, uh, directly. So at least, at least with the way that our current org is handled, there's no privacy issue there.

Interviewee

But, um, maybe like robustness might be something that we like, definitely, like, just like every other company like Covid, the spending. , um, like the type of quiz that people search for, right? Maybe when they search for like pants or something. Now it's like, not office pants, it's like, um, work from home kind of business casual ish kind of pants.

Interviewee

Um, so that intent or like that use of behavior is different. Um, definitely during the holidays. Um, like when someone searches for something like gifts or something. , that's a different intent maybe. So like, um, so our models are not necessarily robust in that piece, but we do have like, um, like features that are updated daily.

Interviewee

And we also have, we define our features so they not only just consider the last end days, they consider like the same time period last year. So that's why we kind of used like a last year or whole year's worth of data to, um, Kind of adjust for the cyclic nature. Um, yeah, so I think robustness is probably there.

Interviewee

And I think one thing is like, um, like misspelling, typos, uh, or like, um, new, new queries like, um, so maybe there's like a new product type or new new item that's released then. So there's certain words that are used in. Um, I can't think of a good example, but let's say like Star Wars were something that was completely new, so maybe if it's something new to our system, we probably would not like tokenize or understand the intent correctly.

Interviewee

So, um, . Yeah. So for some of these things we do have like manual overrides. Um, so where like for certain query we can pin certain items and we just kind of ignore what the model is returning, but yeah. Okay. Really interesting.

Interviewer 1

Thank you. Uh, so two small question. In your opinion, what is the most pressing quality issue?

Interviewer 1

Researchers such as, uh, us, uh, should

Interviewee

try to.

Interviewee

um, model, say what is the most pressing quality issues? Yeah. So, um,

Interviewer 1

any issue Yeah. You and Encount Daily that you think, uh, this is something we should try to solve and it'll make my life much easier?

Interviewee
Um, yeah. I mean, I guess, um,

Interviewee
I mean, this might be a little bit difficult, but like, um, like one thing that's surprising is that the end result is the combination of like many models, many systems that are interacting. So, um, and also like we are frequently seen, like we may improve like the query understanding piece, but in the end, Like even if you improve it by 10%, you know, if it's like a very popular query, like other pieces in the system might already be improving it.

Interviewee
So I think like if we have some way of like talking about like in a more, like in a ML system piece, like how to think. Like how, how much, how much? So like in a information retrieval thing, there's like the retrieval ranking and maybe other pieces, like how do all the, how to consider like improvements in certain areas, like how it might improve or if there's a gap in certain area, how to like, kind of quantify that thing or at least identify with issues there.

Interviewee
Um, so maybe just like some way of thinking about how to address that. . I guess maybe this is a, I mean, maybe this is a more simpler thing about like the model monitoring, I guess. Um, I mean, I'm sure I, I think that like some good companies and, um, days of thinking, but if there's more accessible literature I guess on that, or especially like for the big companies, um, if they do share it, I think it'll be helpful for me, I guess.

Interviewee
But, .

Interviewer 1
Okay. I misunderstood the part. When you mentioned one model that is aggregating

Interviewee

other model. Oh no, maybe not necessarily a model that's aggregating, but like, I mean, if you just take search, right? There's things like, like type ahead, uh, there's like query understanding, item understanding. So all these things together are interacting.

Interviewee

Okay. And. , like, you know, like if I look at the offline results for, or even like semi online, like if I look at like the offline results for like a query understanding thing, like I might see a big lift there, but when it enter ab test, um, because there are so many other pieces interacting together, like, like the results were neutral.

Interviewee

Um, . So like how do you think about, yeah, how do you compare? Um, like in that kind of scenario? Yeah, yeah.

Interviewer 1

Understand. So how to attribute some improvement. Is it this system that we can attribute the improvement or another system basically when there's a lot of updates to different sub master learning software.

Interviewer 1

and you got a, an an improvement on some metric. What you dunno, is it because of this system or this system or another one? Is that more or less?

Interviewee

Um, I mean, I guess one thing I do need to mention, like during the AB test, like things are kept constant. Um, or what I mean is like we only testing one variation, so.

Interviewee

assuming every, with a large number of samples, like everything else should be consistent. But like, I guess more the point I was trying to mention is like that um, like there may be multiple pieces in a system and just the contribution in one layer may not translate to the and end result in any significant way.

Interviewee

So, um, you just that kind, I mean, if there was a way to kind of. Assess that in an offline setting, that would be useful, but that might be a little bit harder task. And it might also be company specific, but, okay. Um,

Interviewer 1

sorry, LA last question I will ask you regarding the, this, this, uh, so what you're saying is you would like to, for the offline environment to be more similar to the online environment or to, to the production.

Interviewee

Yeah. Yeah, I think that's a good way of saying it. Um, okay. And that's because like the running ab tests are expensive. So, um, but the other thing is like if you just look at off, if you just look at model independent, I mean an individual model metrics, um, like you may see like a lift. When you integrated with the whole whole stack, the results may not, you may not really see that lift.

Interviewee

Okay. Perfect.

Interviewer 1

Well, thank you. That's all for us. So thank you. Thanks all for your time. I think it was really interesting and we learn a lot of new things and thank you. We spent 15 minute more, so your time is really valuable to us, so Okay. We really appreciate. Yeah.

Interviewee

Yeah. Cool. Yeah. Thanks PIerre. Thanks.

Interviewee

Appreciate for your time also. Yeah. Uh, have a good rest of day. Same

Interviewer 1

for you. Bye. Thank

Interviewee

you.