Python and AI Power-Up Program Offline Class-20250923_113928-Meeting Recording

September 23, 2025, 6:09AM

1h 37m 53s

Ajay Patel started transcription

Tarun Jain 0:04

When it comes to node, node is nothing but a simple function, right? So whatever logic you want to define, that is nothing but your function and that logical function is nothing but your node. The only thing what you have to worry about is the edge on how you need to create the edge.

For that, this is the reference.

And where will you usually encounter this? In most of the cases, if it is sequential order, it is very simple. But when it comes to human in a loop when you need to have, I mean it generates a final response if you want human to intervene whether it is right or wrong.

Hardip Patel 0:38

Yes.

Tarun Jain 0:50

If that particular logic needs to come within your application, during that time you can have this.

Example.

Workflows.

One is your money in a loop and then you have.

Parallel node execution.

And then also routing systems. Does anyone remember the example of a routing system?

So let's let's suppose you're building a customer service adbot. You have customer service adbot whenever you have a query. So let's suppose this query is related to HR.

Tirth 1:38

No, yes, yes. The routing system where we decide LLM, we use LLM to decide which nodes to call or which another LLM to call, yeah.

Tarun Jain 1:38

And.

Correct.

So if you want to create that architecture as well, if you noticed there will be one complete. Let's suppose this is the line. You will have one complete line for one of the node, but for the remaining other two nodes it will be dotted line. So let me convert this to.

So this will be complete line then for the other two nodes that you have.

Wait, let me draw this properly.

It will be like this. So you have a customer service chatbot, you have query, it is related to HR. So the first block is related to HR then remaining to it can be for Tech and then remaining it will be let's suppose this is.

Hardip Patel 2:42

OK.

It.

Tarun Jain 2:56

Product.

Hardip Patel 2:56

OK.

Tarun Jain 2:58

Is this clear? And now if you notice this should be dotted line, it should not be complete line. So this is how the routing system will look like even if you look at our workflow. So this is straight line, but if you look at parallel node this is dotted lines because both is happening at the same time and then you have the join nodes.

Tirth 3:06

Yeah.

Hardip Patel 3:15

Sure.

Tarun Jain 3:16

So this dotted lines is very important that says how the workflow will look like. If you have straight line, that means both is running at the same time and that will be executed, which is wrong. So let's suppose if this is a straight line, all this straight line, that means your query is being checked in all the three vector database which is.

Hardip Patel 3:17

Yeah.

Tarun Jain 3:35

Sequential. You're not trying to do routing so that you have to ensure whenever you're creating routing system and if you're using land graph you should have dotted.

Oh, is this clear?

Hardip Patel 3:49

Yes, so when I'm running this, it is asking for this grand RF like yeah yeah but I installed and still.

Tirth 3:49

Yes.

Tarun Jain 3:51

OK, so let's, uh, yeah.

Oh yeah, yeah, yeah, sorry, sorry. Yeah. So you need to install this.

Uh, did you restart the system again?

Hardip Patel 4:08

No, no, no, no.

OK, I'll do that.

TJ **Tarun Jain** 4:18

Now you can rerun this.

Hardip Patel 4:35

I'll start my time.

TJ **Tarun Jain** 4:42

Yeah, let me know if you got the response. You should have answer relevance. Then you should have response of. You need to have context relevance. Total. You have 5 variables now.

Hardip Patel 4:42

No.

Hmm.

TJ **Tarun Jain** 4:56

Then you have response of answer.

Hardip Patel 4:56

OK.

OK.

TJ **Tarun Jain** 5:07

You can crosscheck this.

So when I have get get answer relevance, I'm using answer relevance which is a key dictionary and then grade dot content is my value. Same goes for.

Context relevance.

I will remove this.

Let me know once you have the response.

Hardip Patel 5:52

Yeah.

So Tarun, this is just for the the particular keys we have to return is just for the just the constrain for the parallel, right?

Tarun Jain 6:17

Uh, can you repeat?

Hardip Patel 6:19

I mean like we only for the parallel nodes we we cannot return the complete state.

Tarun Jain 6:26

Yeah, if in case you have multiple nodes to the state property.

Hardip Patel 6:29

Yeah, yeah.

That was.

.

Tarun Jain 6:41

It's also follow this practice only. It's better to follow this practice because of their documentation.

Hardip Patel 6:41

Yeah.

Yeah.

Tirth 6:49

Mhm.

Tarun Jain 6:49

Because everywhere they are using the dictionary itself instead of returning state.

I don't know from where, but I guess it should be some cookbook. Uh, cookbook line chain.

Bancroft.

Hardip Patel 7:09

So the the I think I've read blog post or something like that.

Tarun Jain 7:15

Damn, this folks changed it.

So typically now it is. Let's change it. It's better. Let's change it.

Hardip Patel 7:19

Yeah, so.

No. So Tarun, I think it is because even if you have the query as a in both same same key in both the parallel node, it will fail because we cannot have the same key.

When we return the parallel nodes, I think.

Tarun Jain 7:46

Oh, where is the same key? We don't have same key, right? Answer relevance is.

Hardip Patel 7:49

Yeah, yeah, that's right. But when we are returning state we we have query same and we have context same so that.

Tarun Jain 7:56

Oh, oh, OK, yeah, yeah, yeah, yeah. You know, yeah, that that's also a possible reason because when you're returning state, you already have these values.

Hardip Patel 8:04

Yeah, because I was getting two separate errors based on their parallel. Sometime I was getting query mismatch and sometime I was getting context mismatch.

Tarun Jain 8:17

Got it. But let's just update this instead of returning state. I guess even in their documentation of it is dictionary.

Hardip Patel 8:18

Yeah.

Tarun Jain 8:29

So if you see they are returning context, insert context they have retrieve documents and for answer they have response dot content.

Hardip Patel 8:43
OK.

Tarun Jain 8:55
So guys, instead of returning state, let's just change it into answer which is dictionary.

Hardip Patel 9:19
Punch check.

Tarun Jain 9:19
Uh, is it done?

Hardip Patel 9:22
Three years later.

Tarun Jain 9:25
OK, so.

Tirth 9:26
Just a minute. Yeah. No, go ahead. You can continue. I'll just.

Tarun Jain 9:30
OK, so far what we have done is for evals we looked at first approach was rubrics based.
So rubrics based is nothing, but you will define it by your end itself, like how much marks needs to be given for particular subbedding and you have your own input data for that. So this is basically dependent on prompt, right? So you're just giving a prompt and you're using LLM to get a Jason.
So E Jason will have a key and it will have a score. So Rubric's based in the sense this is ranking.
Which is rank based. You give some values between zero to five and now when it comes to the second thing which is LLM as a Z.

We have something called as rack rayad.
So in this rack thread we had context relevance, we had answer relevance and we have groundedness.
And this RAG triad again is basically based on prompts. So you're writing prompt and you're using LLM to generate the final response. Now there is third approach. The third approach is metrics based which is statistics.

Hardip Patel 10:35
So.

Tarun Jain 10:45
And mathematic based. So you have certain mathematical formula based on which you generate specific metrics. So this metrics is nothing but you have.
Context precision. So basically just called precision and then you have recall.

Hardip Patel 11:04
But.

Tarun Jain 11:07
How many of you have heard these keywords before, precision and ripple?

Hardip Patel 11:13
I haven't.

Tirth 11:13
In stats and mathematics. No, I haven't.

Tarun Jain 11:17
OK, so you guys remember there is a story of a boy and a tiger.
So basically a boy usually goes back to his village and it says that hey, tiger has come. So let's take a possibility that.

Hardip Patel 11:30
Um.

Tirth 11:31
Yeah.

Tarun Jain 11:35
Boys.
Prediction.

Hardip Patel 11:39
OK.

Tarun Jain 11:41
And here I will tell actual prediction.

Hardip Patel 11:42
Yes.

Tarun Jain 11:48
So actual prediction in the sense whether tiger.

Hardip Patel 11:52
Chill again.

Tarun Jain 11:53
Was actually came or not? OK, so you have how much possibilities either it is true or false, zero or one and same for action. So what you will usually do is you will define this as 10 if one that means tiger appeared.

Tirth 12:01
Yeah.

Tarun Jain 12:10
If 0 means tiger, didn't appear. Same goes for boys prediction.

Hardip Patel 12:13

Um.

Tarun Jain 12:16

Instead of prediction, I'll make it voice slide.

Hardip Patel 12:20

Mhm.

Tarun Jain 12:23

Or not like judgment.

Hardip Patel 12:25

Mhm.

Tarun Jain 12:27

OK. And then you have 0.

So let's suppose boy's judgment is 1 and even actual tiger appeared. This means it's called true positive.

Hardip Patel 12:32

Oh.

Mm.

Tarun Jain 12:40

And if in case boy said there was no tiger but actually tiger appeared, this is called false negative.

Hardip Patel 12:49

Hmm.

Tarun Jain 12:53

And boy's judgment is said, hey, actually there was a tiger, but in reality there was no tiger. So this is called false positive.

Hardip Patel 12:54

M.

Mm.

Tarun Jain 13:03

So why is it false positive in terms of judgment? So this is what your actual prediction is, right? When you're building a model, you're only focusing on this judgment. You're not actually worried about the actual thing. So basically most of the time what happens is if you're building model.

Hardip Patel 13:05

Yeah, yeah.

Mm-hmm.

Tarun Jain 13:20

There are right chances you will not have actual information on the new data. So your only judgment is this particular prediction which is 10 by boy judgment. So if boy is telling one but in the actual it was zero, that means it's called false positive. I hope these keywords are clear.

Hardip Patel 13:37

Yeah.

Tarun Jain 13:39

And if boy said no, and even if tiger says no, I mean if tiger didn't appear, it's called true negative.

Hardip Patel 13:40

Yes.

Tirth 13:40

Yes.

Tarun Jain 13:50

So this is just one example, but if I have to take an actual example, this is mainly used for spam prediction.

Tirth 13:59
We actually use just false positive when there is a bug or when there is, you know, like someone said it's a bug on production and it was not. So we say that you know it's false positive. Don't worry about it.

Tarun Jain 14:10
Uh, firstly you do, correct?

Tirth 14:13
I can use many terms. True positive, false negative, true negative. You confuse people with this.

Hardip Patel 14:14
OK.

Tarun Jain 14:15
Right.
But this time sometimes this will sometimes be very confusing. So we overlap like this.

Tirth 14:22
I know.

Hardip Patel 14:27
Hmm.

Tarun Jain 14:27
The reason why I kept this diagram here is because I could have confused it again.

Hardip Patel 14:32
Mm.

Tarun Jain 14:33

This I've done multiple times but still I get confused. So one of the best example is spam prediction. So what happens in spam prediction if it is spam but actually it was also spam then it is too positive. So instead of voice judgment and tiger up here. So this use case is nothing but.

Hardip Patel 14:36

M.

Mm.

Hmm.

Tirth 14:51

Oh.

Tarun Jain 14:52

Did Tiger appear?

Hardip Patel 14:52

Hmm.

Even the example on the right side is very good.

Tarun Jain 14:58

Which time?

Hardip Patel 14:59

This pieces, yeah.

Tirth 15:00

Other species, species, other species.

Tarun Jain 15:02

Yeah, but this one is the best, like you actually have emails predicted and actual.

Hardip Patel 15:07

Um.

Tarun Jain 15:10

OK, so now based on this you can arrive at three different metrics. One is accuracy. One is precision.

Hardip Patel 15:20

Hello.

Tarun Jain 15:22

And one more is recall. And the most important when it comes to neural networks or simple machine learning, it will always be accuracy because everything is related to numbers. But when it comes to LLMS, since you don't actually have numbers, you have text.

Precision and recall is widely used. So how does this interpret? So in precision, let's suppose of all the items that you have predicted positive, how many were actually correct? So this is your precision, which is.

Whenever you have predicted the positive, how many were you actually correct? And recall is nothing but whenever you have actual positive items, how many did you successfully find? Which is nothing but this.

Hardip Patel 16:05

Hmm.

First negative.

Tarun Jain 16:12

This particular part when you have 0.

Hardip Patel 16:16

Sorry.

Tirth 16:16

Can you can you go back to the definition again? What is the definition?

Tarun Jain 16:16

So this is.

Tirth 16:21

So precision is of all the items I predicted as positive. How many? Yeah, yeah, yeah. So that's.

Tarun Jain 16:26

AI, in the sense the model.

Tirth 16:31

That's the first column, right? That's the positive one. Predicted is positive.

Hardip Patel 16:38

It could be positive or negative, right?

Tarun Jain 16:42

No. So if you look at here, what is it positive or negative?

Hardip Patel 16:46

You.

So for the confusion metrics, negative is better, right? As in like we want to predict spam, so that's why we are having.

Tirth 16:58

It is negative, right?

Tarun Jain 16:59

So this is 0 here, this is 0.

Ajay Patel 17:01

But but whom should we consider having a, you know, major chunks, false or negative?

Tarun Jain 17:11

Uh, can you repeat?

Ajay Patel 17:14

Example boys or tiger.

Tarun Jain 17:18

Huh.

Tirth 17:28

What is good is true positive is good for accuracy. True negative is also good for accuracy.

Ajay Patel 17:32

Mm-hmm. Mm.

Tarun Jain 17:34

This one. This one.

Tirth 17:38

No, no. True positive and true negative. They both are accurate. I mean, voice judgment was right.

Ajay Patel 17:38

So cross is good.

Tarun Jain 17:38

No.

No.

Tirth 17:46

In true positive and true negative both.

Tarun Jain 17:53
Oh, can you repeat?

Tirth 17:54
So voice judgment was right in true positive as well as true negative.

Tarun Jain 18:03
You mean this blog? Let me undo this.

Tirth 18:05
Yeah, this. No, no, no, no, no, no, no, no, no. Just the first one. The cross one like the diagonal. True positive and true negative.

Tarun Jain 18:07
This one.

Ajay Patel 18:10
Cross, cross, cross.
I can.

Tarun Jain 18:14
Uh.
But diagonal will always. This is likely to be an option, so not every time you'll have diagonal, right? So most of the time what will happen is if you have confusion metrics, your diagonal will have more marks, so four and three. But if you see your thing will also have certain values which are.

Tirth 18:30
Hmm.

Tarun Jain 18:33
Very less so if in case you have 1000 emails.

Tirth 18:38

Mhm.

Tarun Jain 18:38

Your true positive can be around 488 or 499 something and this will be around 4:50. Then remaining will be your what you call false prediction that you made, but the method it will be.

Tirth 18:44

Mhm.

Yeah, false positive and false negative.

Tarun Jain 18:56

But when it comes to which one to pick, usually we use accuracy. So in accuracy if you see it, everything will be covered. So you're considering TP plus TN, which is nothing but your TP and TN, and then you're taking all the options.

Hardip Patel 19:01

M.

Tirth 19:09

20 in right?

Tarun Jain 19:12

So accuracy is preferred in that scenarios, but if you're actually focused on.

Tirth 19:16

So accuracy means TP plus TN divided by the mean or that is the total.

Tarun Jain 19:21

Total. So now what will happen? We have 1000. So 1000 is nothing but it will be distributed across all this 499. Suppose this is 1.

Tirth 19:24

So we are getting OK.
499, yeah, 499 450 / 1000, yeah.

Ajay Patel 19:30
OK.

Hardip Patel 19:31
No.

Tarun Jain 19:34
Let's suppose this is one, this is 50. So 50 times what happened is my model predicted it as not a spam, but actually it was a spam.

Tirth 19:43
M.

Tarun Jain 19:45
It is 50. So now if you want to predict accuracy, it will be 499 450 / 1000. So how much is this? 500 four 5949.

Ajay Patel 19:45
Mm.

Tirth 19:50
That's 5450 divided by 1000, yes.

Hardip Patel 20:02
Fair 950 -, 949 by 1000.

Tarun Jain 20:03
200.
So this will be 94.9% which is your accuracy.

Tirth 20:12
20%, right?

Hardip Patel 20:13

OK, yeah.

Tirth 20:18

OK.

OK.

Tarun Jain 20:19

Precision. So how will you calculate precision now? Which is your TPTP is nothing but this one.

Hardip Patel 20:26

Hmm.

Tirth 20:26

The decision is true positive only. Decision is with true positive only, not false positive.

Tarun Jain 20:31

Huh.

Hardip Patel 20:32

49.9%.

Tarun Jain 20:34

True positive divided by PP this one. So you have.

Tirth 20:38

True positive plus false positive, so with only positives ones that is precision.

Hardip Patel 20:40

M.

So 99 by 1000.

Tirth 20:46

No.

Tarun Jain 20:47

No, it's.

Hardip Patel 20:47

Mhm.

Tarun Jain 20:50

Sorry, sorry, 499 divided by 500.

Tirth 20:53

499 / 500, yes.

Hardip Patel 20:57

Well, fine there, OK.

Tarun Jain 20:58

So if you see most of the time you just made one small mistake, that's it. If you clear when it was supposed to be 0, which was not a spam, you made it as a spam. So you made only one mistake which is related to positive.

Hardip Patel 21:05

Hmm.

Tirth 21:05

Yeah.

Hardip Patel 21:11

Mm.

Tarun Jain 21:16

So your precision is almost 99%.

Hardip Patel 21:16

Mhm.

Tirth 21:18

99.899.8%, yes.

Hardip Patel 21:19

M.

Mm-hmm.

Tarun Jain 21:21

So this is what if you want to know how much positive you are creating, this accuracy will be used. Here also if you see precision, the only thing is we used 10 in this diagram we're using 01.

Hardip Patel 21:33

Mm.

Mm.

Tarun Jain 21:38

So here actually it is 0 and this is one. That means if it is 1 means it is a spam which is red colour. So now if you look at the diagram this is also zero, this is 1. If it is 11 it is true positive which is true positive.

Hardip Patel 21:39

Mm.

Tirth 21:40

Yeah.

Hardip Patel 21:41

M.

Tirth 21:46

OK.

Hardip Patel 21:47

Mhm.

Tirth 21:56

No, no, one verse. Oh, OK, yeah, true positive, right?

Tarun Jain 21:56

If it is 0.

21:58

Hmm.

Tarun Jain 21:58

11 is positive. Here it is 1.

Tirth 22:01

Super.

Tarun Jain 22:02

Here also it is one and it is 0.

So now if you see 00, it is true negative. 00 is true negative. Then this is 0 and this is one. It is false positive, zero and one false positive. The only thing is it is reverse. That's it.

Tirth 22:06

Right.

OK.

Tarun Jain 22:22

But when it comes to precision, you're only taking the positive.

Tirth 22:27

OK, what should we call?

Ajay Patel 22:28

OK, yeah, this reminds me of all my college base Karnup map K map.

Tarun Jain 22:34

Correct K map. Usually K map is also used for this only.

Ajay Patel 22:35

Hmm.

OK.

Tirth 22:41

Yeah.

Tarun Jain 22:41

Of bullions.

Hardip Patel 22:43

This I will forget very easily. It's.

Ajay Patel 22:43

Yeah.

Tarun Jain 22:47

I'll forget this.

I don't know how many times I've given this talk, but still, but in talks it's like if I don't have the reference, gone.

Tirth 23:04

What about the recall?

Tarun Jain 23:07

I recall it is this line.

How many times it was actually correct? Which is your actual prediction?

True positive then true positive EP plus FN. So how much this was? This was 499 and this was 50.

Ajay Patel 23:21

Yes.

Tirth 23:28

So what is? What does that stat give? Like what is the importance of free call?

Tarun Jain 23:30

Right.

Ajay Patel 23:33

Hmm.

Tarun Jain 23:34

So let's suppose 499. How much this was below 499 plus 50549. So how much is this for?

Tirth 23:38

That is 499 + 50.

Yeah.

949 is approximately 90.8 percent, 90.89% or 90.9%.

Tarun Jain 23:57

OK, so I'll take 91% roughly. So this is recall. Now what this will tell is in terms of your actual prediction.

Actual prediction that you have. How many times were you actually correct?

Actual prediction how many times the model was positively correct?

Hardip Patel 24:21

Hmm.

Tarun Jain 24:23

This word is important positively.

Because you are only judging based on the actual prediction. Actual prediction is 1 which is your positively correct.

Tirth 24:39

Yeah.

Tarun Jain 24:42

Wait, let me get this simplified meaning somewhere I'd cooperate it.

Tirth 24:48

The actual prediction was one. The boy said no, but it was one, so it is false negative.

Tarun Jain 24:59

So precision of all the emails marked spam. How many were really spam?

Now recall is nothing but of all these spam emails, how many got mark spam?

Tirth 25:09

Of all the spam emails, how many?

At sure. So of all the spam emails, how many were actually marked spam by the LMM or whatever we are using?

Hardip Patel 25:22

Play.

Tirth 25:27

Understood. Understood. No, this this makes sense. OK.

Tarun Jain 25:27

What?

Tirth 25:32

I'll just repeat it in Gujarati, you know, so I'll I'll make it comfortable for myself.

Tarun Jain 25:33

So this is.

Hardip Patel 25:49

Accuracy.

Ajay Patel 25:53

OK.

Tirth 25:54

Was very precise.

Tarun Jain 26:14

Yes.

Tirth 26:16

Oh.

Ajay Patel 26:17

OK.

Tirth 26:20

Spam mark 90%. You're 90% precise.

Ajay Patel 26:24

In.

Hmm.

Mm-hmm.

Mm-hmm.

Tirth 26:34

But my location so marked.

Mitesh Rathod 26:34

Me too.

Ajay Patel 26:42

OK, Undu.

Tirth 26:44

OK, now I have really don't do hot.

Ajay Patel 26:45

Hmm.

Tirth 26:56

Yeah.

Mitesh Rathod 26:59

Today it's kind of we can say.

Tirth 26:59

OK.

Ajay Patel 27:02

Sambono again.

Mitesh Rathod 27:04

So.

Ajay Patel 27:07

Probability.

Mitesh Rathod 27:08

OK, probability, OK.

I'm not. I didn't. I didn't like it.

Ajay Patel 27:17

Statistics. This is a pure statistics.

Tarun Jain 27:22

Uh, it's purely stats, the mathematicalist.

Cool. So I've attached this screenshot. If you got the result, what we'll do is we'll proceed with the metrics based. So first we have faithfulness. So faithfulness, it will range between zero to 1, which will just tell whatever response you have, if it is related to the retrieved context or not.

Tirth 27:32

OK.

Mhm.

Tarun Jain 27:44

So if you look at the groundedness, So what was the symbol for groundedness?

So if you see for answer relevance.

It is a slash Q.

Then you had context relevance.

This was C.

Tirth 28:08

Groundedness means uh.

Tarun Jain 28:09

Slave crash queue. Then you have groundedness.

Tirth 28:12

Hmm.

It is with U slash A.

Tarun Jain 28:16

Or it's also called faithfulness.

This will be. No, this should be.

Tirth 28:21

Useless, useless.

Tarun Jain 28:24

Uh.

Hardip Patel 28:26

Hmm.

Tarun Jain 28:26

There are only six.

Tirth 28:28

But.

Hardip Patel 28:31

Gradian.

No answer and.

Tarun Jain 28:36

Yes.

So if you see response is nothing but A and retrieve context is nothing but C This is C and this is A.

Tirth 28:40

Um.

Hardip Patel 28:47

OK. OK. Yeah.

Tirth 28:48

But.

OK, the relevance with question.

Hardip Patel 28:50

But uh, the grounding is dependent on the extraction right of the context.

Tarun Jain 28:58

Oh, which one?

Hardip Patel 28:59

Groundedness where like we have the context as a second, but it is dependent on if everything is available in context, right? So I don't know if I'm clear or not.

Tarun Jain 29:14

I didn't get it. Can you repeat?

Hardip Patel 29:18

OK, So what I'm saying is like in the context we have to give the information right then and only then we can have the extraction right. So like even here can like we'll ground this with the right factor.

Tarun Jain 29:35

So basically for groundedness.

Tirth 29:35

Groundness will be the right factor I guess. So OK, let me answer this and Tarun you can correct me. So you know even I have the answer. So we have the answer which is you know was used by some context. OK.

Tarun Jain 29:42

Yeah, yeah.

Hardip Patel 29:46

Mhm.

Tirth 29:50

Now what we are trying to check over here is whatever context that we received from our vector database, is that more in line with the answer that LLM has given?

Hardip Patel 29:50

Mhm.

Yes, exactly, but.

Tirth 30:05

So we are, so we are checking with the context.

Hey.

Hardip Patel 30:22

But also we need to have context where it mentions that Jalan is the sister, but if it does not have it under the context, wouldn't it be the problem like even like then can I trust?

I think I don't know if I make sense.

Tirth 30:42

But that is what we are testing over here, man.

Tarun Jain 30:43

No, you can't. So here basically what will happen is let's suppose you have whatever teeth said, right? So whatever response you have, you're judging that against the context. So far we never judge the answer with the context directly. So if you see here it is answer and question.

Hardip Patel 30:47

M.

Oh.

Mhm.

OK, you're not expecting.

Tarun Jain 31:00

Then it is context and question. Now let's suppose you have some query like who is

Virat Kohli. So this question itself is not there, but when you look at the context technically, what should your context be?

Hardip Patel 31:05
Hmm.
M.

Tirth 31:14
Context would be nothing blank.

Tarun Jain 31:16
So it should be. It should be done.

Hardip Patel 31:16
Can we get a?

Tirth 31:18
It will be blank.

Tarun Jain 31:20
It should be null, but it won't be null. So during that time what you will do, you will have groundedness which is for this edge cases.

Hardip Patel 31:22
OK, OK.

Tirth 31:29
OK.

Tarun Jain 31:31
But you can take this. What you can do is here as this is Virat Kohli.

Hardip Patel 31:37
Hmm.

Tarun Jain 31:42
So when I do context, this should be empty, but it won't be empty.

Hardip Patel 31:47
No, it it should be empty now.

Tarun Jain 31:51
It won't be empty. So what are you trying to do here? You're trying to do search. So when you do search, even this will have its own embedding.

Hardip Patel 31:55
OK.

Tirth 31:59
Who is also will have an embedding and it will try to find something with who is.

Tarun Jain 32:05
But the thing here is your scores will be less. When you try to pin the scores, your scores will be less, but context will be there. But if you take the length of this, I guess it is not 3.

Hardip Patel 32:06
Yeah.
Yeah.
OK.

Tirth 32:14
OK.

Hardip Patel 32:14
OK.

Tirth 32:18
We got it. We changed it to seven.

Tarun Jain 32:21

OK, it's three only, but now if you look at the answer, it will be I don't know.

Tirth 32:22

It's a tape.

Hardip Patel 32:25

Oh, obviously, yeah, that is the right answer to have for groundedness, right?
OK, so I had wrong expectations. I understood. Thank you.

Tarun Jain 32:33

Correct.
So now we'll be using a library called Ragas, which is widely used in most of the frameworks. Let us import that. And if you remember I told you always we evaluate on 2 basis, one is on single turn.
And 2nd is multi turn.

Hardip Patel 32:54

M.

Tarun Jain 32:56

And just like llama index, you have a default LLM in ragas and the default LLM is open EID.
So the default LLM is Open AI. So the LLM that we are using is Gemini. So what we need to do is we need to create the Lan Chain wrapper so that we can use the Lan Chain LLM directly in Ragas and then we have to choose anything from this. We'll be using Singleton.
So there are total 3 import statement. One import statement is for Singleton, the second import statement is for LLM and 3rd import statement is for faithfulness. So from Ragas.
You have data set schema.
Import.
You have the Singleton sample, then from ragas you need LLMs.
Because the default is open AI and we don't need open AI and I want to use the

langchain wrapper since the LLM that we are using. So if you look at this LLM that we have, this is from langchain.
Chat Google generative AI. So if you want to reuse the same LLM, what you can do is you can define.

Hardip Patel 34:16
Mm.

Tarun Jain 34:23
From ragas dot LLM import LLM wrapper and then we need to import this faithfulness. So this is a metric. So whatever metric you need, you just have to do define from ragas import metrics.
Import faithfulness.
And apart from that, we also need precision.
You have context precision. I wanted to show one more thing. So here let's now that you understood the precision, right? How does this work in terms of flag setup? So.

Hardip Patel 34:53
Yes.

Tarun Jain 35:04
How many K values do we have?

Hardip Patel 35:06
3.

Tarun Jain 35:07
Case 3. So precision can be done in two approach. One is precision for all which is nothing but you will consider all the 3K values and then you will have precision for one. This means for K equals to one. What is the precision?

Hardip Patel 35:10
OK.
Yeah.
M.

Tarun Jain 35:27

In most of the cases here you will have very high score, you will have 90% or you will have 99% then precision too. Also you will have very high score the lower you go. Let's suppose you go to the last chunk which is your.

Hardip Patel 35:37

Yeah.

Yes.

Yeah.

Tarun Jain 35:43

Precision of three, you might also find 0. So when I add a prompt like what is the e-mail to contact?

Atyantik So this contact, whatever I have, it is present in only K equals to one and K equals to two, but K equals to three. There is no information in it, right? So what should be the precision technically for the last K value? It should be 0.

Hardip Patel 35:56

No.

Yeah.

Tarun Jain 36:10

So you can do precision on each K value as well.

Tirth 36:17

OK.

Tarun Jain 36:18

So here instead of you have LLM context.

Hardip Patel 36:22

Hello.

Tarun Jain 36:22

LLM context, we don't have the reference. Let's suppose you have the reference. For example, whatever question you're asking for that question, you already have an answer and then what are you trying to do is you want to compare it with LLM, but we don't have the actual response. So what will I do?

Hardip Patel 36:26
Yeah.

Tarun Jain 36:38
I will import from LM context precision without reference because there is no reference answer.
So if you see here, here I've defined precision of K Precision of K is nothing but you're checking for every single chunk. Either you can do it on the entire level. If you want to check only on the chunk level, like what is the precision on chunk one, chunk 2, chunk 3, you can also do that as well.
And Ragas, it's the documentation is very clean if I open Ragas.

Hardip Patel 37:14
Mm.

Tarun Jain 37:16
You can directly get the code. So you have Singleton example, you have user input, then you have response, then you have retrieve context. So this is nothing but AQ and C which we already have. So this is without reference.
If you look at with reference, you have user quotient and then you have reference and then you have retrieved context. So if you look at this reference, this is the output which is already existing. So there are two variables, one is response and one more is reference.

Hardip Patel 37:47
Hmm.

Tarun Jain 37:47
And in 99% of whatever we built, we will never have the reference because you don't

know what data you are using. So we'll go with without reference and just like that you have faithfulness, you have recall.

Tirth 37:52

Um.

Hardip Patel 37:52

Mm-hmm.

Um.

OK.

Mm.

Tarun Jain 38:03

And.

So NVIDIA metrics is nothing but the answer relevance.

Answer context and groundedness, which is nothing but your ragtriad.

So you can import these three lines.

Hardip Patel 38:19

Hmm.

Tarun Jain 38:24

And now what we have to do is we have to define our LLM, the eval LM that we need to use. Whatever line chain wrapper is there, copy this.

And inside this we have Gemini LLM and the variable is just LLM.

So now this is the LLM from Ragas.

Tirth 38:43

Yes.

Hardip Patel 38:45

Mm.

So default is open AI. That means that it will require open AI key if I'm not specifying LLM is it?

Tarun Jain 39:01

So let's suppose if you don't define this and if you're directly defining what you call faithfulness. So you define faithfulness equals to faithfulness. That's it. So if you define like this, it's using open AI.

Hardip Patel 39:04

Hmm.

Hello. Yeah, OK.

OK.

Tarun Jain 39:18

If you Scroll down below in the same code here if you notice I'm using Open AI because whenever it comes to data set creation Gemini after creating 3 questions it will hit the limit. So usually I use Open AI for data set creation.

Hardip Patel 39:22

Hmm.

Mm.

Hmm.

No.

Tarun Jain 39:34

If you Scroll down below.

Hardip Patel 39:36

Hmm.

Tarun Jain 39:36

You will find metrics. So in this metrics if you see I'm directly defining faithfulness, factual correctness, LLM context. Why? Because I already have open EIP.

Hardip Patel 39:40

Hmm.

OK, right, right, right.

Tarun Jain 39:50

But here I don't have open AI, so I need to define eval LLM and once I have eval LLM here you have a parameter called LM.

Hardip Patel 39:59

Yes.

Tarun Jain 40:02

Uh, where is the parameter?

Hardip Patel 40:05

Uh, sorry for another question, but is it OK to use Open AI right now or?

I mean like OK and we have used the Gemini in the upper one and this is a completely different. We are using this as a judge so we can have a different LLM right?

Tarun Jain 40:13

Here if you want to use, you can use.

You can have any different LLM here, so let's suppose.

Hardip Patel 40:33

OK.

Yeah.

Tarun Jain 40:35

Where is a Tuba article?

So here if you see in the Uber article, where did we complete? We created this data set creation, then we save it in Vector DB and once you have Vector DB we usually perform search. So when we do search, are you comfortable with these two keywords now? Now what kind of search is this?

Hardip Patel 40:50

Mm.

Yeah.

Uh, Dance and Spa Hybrid.

Tarun Jain 41:02

This is hybrid search. Now if you look at the site here you have LLM as a judge and then you have their own data set. When it comes to post processor, can you see the difference here for answers in this you have LLM large.

Hardip Patel 41:04

Yes.

Um.

Mm-hmm.

Tirth 41:17

Small.

Hardip Patel 41:18

Mhm.

Tarun Jain 41:18

Then for 4th processor you have LLM small.

Hardip Patel 41:21

OK, so this is for you.

Tirth 41:21

Yeah.

Tarun Jain 41:22

So post processor, what is happening? Whatever data you have here, they're trying to evaluate for that evaluation are using a small LLM.

Hardip Patel 41:28

Hmm.

OK.

Right.

Tarun Jain 41:33

So this LLM can be your Mistral, it can be llama, it can be anything. So that doesn't matter, right? So you can have two LLMS in your entire code base. One it is for actual LLM and one more it is post processing is for evals.

Hardip Patel 41:37

Hmm.

Mhm.

Mm.

This this will be on Prem right?

Tarun Jain 41:53

Huh.

So if you look at the Uh evals, it's the same thing whatever we are doing now.

Hardip Patel 41:59

Mhm.

Tarun Jain 42:00

So you have metrics based. So for metrics based they're clever. They didn't mention the accuracy and all.

Hardip Patel 42:04

Well.

Tarun Jain 42:09

But if they're using Langraph, I'm pretty much sure what metrics they're using. They're using faithfulness, recall or precision. Apart from that, there is no other metrics.

Hardip Patel 42:14

Yes.

Mm.

Mm.

Tarun Jain 42:23

They didn't mention it, but I'm pretty much sure it is the same thing. When they say LLM as a judge, LLM as a judge, there are only 6 evals. Answer relevance, context relevance, groundedness, faithfulness, and sometimes you have correctness and other things, but usually ragrad should be sufficient.

Hardip Patel 42:41

Hmm.

Tarun Jain 42:43

So make sure whenever you're defining this faithfulness, open AI is default. If you hover on this you will see LLM and then you will see base ragas LLM.

Hardip Patel 42:45

OK.

Tarun Jain 42:55

Can you see this? This is not what you call open AI LLM base ragas LLM. So I want to change this into line chain.

Tirth 42:57

Yes.

Hardip Patel 42:58

Hmm.

Tarun Jain 43:13

Till here is it done? We are just defining the LLM and we have defined the metrics.

Hardip Patel 43:15

OK.

Tirth 43:20

Mhm.

Tarun Jain 43:22

Now what we need to do is define a sample which is singleton sample equals to single.
Run sample and inside this you have user input. So what is our user input? It is result of query.
And then you have uh response. Response is nothing but result.
Of answer. Then you have retrieved context.
Equals to result dot context. There are two other parameters. So far whatever we have seen is we saw user input, response and retrieve context. We have these three things, but when you are creating your data set you will find 2 new variables.
So those two new variables are reference.
Which is nothing but your actual answer and then you will have reference context.
Which is your actual context or chunk?
So this two will be generated when you have your own data. I mean when you have your own data, which is nothing but this one.
Golden test data and this code is available. Once we do this, if you scroll below, you have that code available. We'll come to that part.
So for time being you have 5 input, user input, response, retrieve context, reference and reference context. Why is it showing yellow?

Hardip Patel 44:49

Hmm.

Tirth 44:49

OK.

Hardip Patel 44:53

OK.
Response. We are using response.

Tarun Jain 45:04

So now you can just run this. It should be 0. Why 0?
Can you tell me why it should be less?

Hardip Patel 45:11
Um.

Tarun Jain 45:15
What is the question response of?

Hardip Patel 45:18
OK.

Tirth 45:18
Is it with Virat Kohli?

Tarun Jain 45:21
This is who is Virat Kohli. So now we can generate 4 faithfulness 4 equals to. OK, we have to run the update because.

Tirth 45:22
Yeah, Virat Kohli.

Hardip Patel 45:32
Oh.
OK.

Tarun Jain 45:35
Single turn, whatever the the function is there, it is asynchronous.
So if I do faithful.

Hardip Patel 45:43
OK.

Tarun Jain 45:45

dot.

Singleton score. So if you see this EA is nothing but what is this called?

Hardip Patel 45:51

Hmm. There's in for us.

Tarun Jain 45:54

Asynchronous programming. So what variable should I use?

It is not variable. I mean what keyword?

Hardip Patel 45:59

Mm.

Ishan Chavda 46:02

Listen.

Tirth 46:03

Basically.

Tarun Jain 46:05

No, async is used when we are using function. Here we have to use await.

Hardip Patel 46:05

OK.

Ishan Chavda 46:07

OK.

Hardip Patel 46:07

OK.

Tarun Jain 46:10

And make sure you run those tool code cell with this.

Hardip Patel 46:14

OK.

Tarun Jain 46:16

Where is that apply nested?

OK, uh, we have to import.

Import Nest Asinshio.

Nest atensio dot apply.

Hardip Patel 46:37

Mm-hmm.

Tarun Jain 46:40

And once you do that, here you can just pass your Singleton sample.

Hardip Patel 46:51

Thank you.

Tarun Jain 46:57

It is 0.

Now what you can do is you can pass any other question. What will I do here? I'll just ask what is the mail to contact.

Yeah.

Tirth 47:15

Can you scroll down what? What are we using? Uh.

Hardip Patel 47:18

Yes.

Tarun Jain 47:19

Yeah.

Tirth 47:20

Yeah, eight full dot Singleton score, OK.

Tarun Jain 47:25

So make sure if you're using VS code, you don't have to do this. This is only for those who are using collab.

Hardip Patel 47:35

OK.

Tarun Jain 47:43

Now it is 1.0.

Hardip Patel 47:44

Yeah.

Tarun Jain 47:57

So this is the same for precision also. So what we will do is let's try to put this little bit above.

Hardip Patel 47:58

Yes.
Mhm, mhm.

Tarun Jain 48:06

And then whatever faithfulness is there, I will bring this below.
This is metrics based. This is same for both precision and other metrics as well.
Yeah, now you can copy it.
So for context precision, what we will do is we'll use OPIC because I want to show how to use feedback scores, what do you call feedback score feature in OPIC.

Hardip Patel 48:57

Yeah.
No, you done.

Tarun Jain 49:19

Uh, everyone also.

Tirth 49:20

Yes, works for me. Yes, works for me as well.

Tarun Jain 49:23

So let's open Opic.

Hardip Patel 49:25

Only thing is for LLM I had to change it to Gemini. The the base Rep one does not work for me.

Tarun Jain 49:36

Did you add this thing? Import OS OS dot environment open AI API key? Then you have to press the key.

Hardip Patel 49:43

On that, yeah, I forgot.

OK, OK, got it.

OK.

Tirth 49:56

When I tried the same with Virat Kohli, I got faithfulness of 0.5 rather than 0.

Hardip Patel 49:58

Right.

Tarun Jain 50:05

No, that's fine. So the range is 0 to 1.

Tirth 50:08

And we should only consider greater than 0.6 or something.

Tarun Jain 50:13

Now if it is less than three or four, that means there was not a right response.

Tirth 50:19

OK.

Tarun Jain 50:19

So usually you'll not have 100% or 0%. Your scores will actually be like this.

Hardip Patel 50:20

OK.

Tarun Jain 50:26

66 percent, 75 percent, 69%. Here we are just playing with one.

Tirth 50:29

But what should be the cutoff? What should be the cutoff for faithfulness?

Hardip Patel 50:31

Yeah.

Tarun Jain 50:34

If you are just working with one or two queries, the cutoff is two or three.

Tirth 50:39

OK.

Tarun Jain 50:40

But if you are working in bunch, let's suppose here I'm taking around 20 questions. So during that time, even if you have more than 60%, your app is very good.

Tirth 50:46

Mhm.

Hardip Patel 50:47

But.

Yeah.

You.

Tirth 50:51

OK.

Tarun Jain 50:52

Not, not very good. I mean, it's good. If you have more than 80%, that means it's well performed.

And the 80 more than 80 is very rare. If you're doing too much of fine tune your application, then probably you might achieve 80%.

Tirth 50:59

OK.

Hardip Patel 51:00

Yeah.

OK.

Tirth 51:02

Mhm.

Mhm.

Mhm.

Tarun Jain 51:11

So most of the RAG applications that we have developed, it is somewhere around 78 to 84% and I don't think we have achieved more than 84% yet. There are these folks, Yamini.

Oh, sorry, it's Lamini.

Hardip Patel 51:29

Uh.

TJ **Tarun Jain** 51:33

So these folks have 90% plus they're using something called as memory tuning, which they're not shared it to anyone.

Hardip Patel 51:39

OK.

Tirth 51:41

Mhm.

TJ **Tarun Jain** 51:41

And they do have SDK. It's in case someone wants to become their customer, then only they'll share it. But they're not exposed what this memory tuning is. So with this memory tuning, they're saying they have more than 95% accuracy for RAG, but not sure how true it is.

Hardip Patel 51:46

OK.

1.

TJ **Tarun Jain** 52:02

They do have memory, I mean.

Hi, you want many people on what we call Reddit also does like how they achieve 95 because it's very rare seeing any application more than 95.

Ajay Patel 52:18

M.

TJ **Tarun Jain** 52:20

But they have 95 plus.

So is this done?

Hardip Patel 52:29

Yeah.

Tirth 52:29

Yes.

OK.

Tarun Jain 52:30

OK, so coming to comment, what we did last time is we just had all this example.

Hardip Patel 52:35

Yeah.

Tarun Jain 52:42

So if you see you have input and output, you have what input you use, what prompt you use. So. So this thing is good for latency, input, output and prompt. So the other two parameters are still pending, which is feedback scores and metadata.

Hardip Patel 52:47

Yeah.

Tarun Jain 52:57

So metadata, when will we use? If we are using memory, what is saved inside memory? If you want to check that, we'll be using metadata. Second for function calling. Let's suppose you have 5 tools. When I ask a question, which tool was used? What was the input that went into the tool and what was the output that a tool generated, not the LLM, the tool. So that input and output of tool will be inside metadata. So you have memory, you have function calling and then you have something called as MCP.

Hardip Patel 53:20

Mhm, mhm.

Tarun Jain 53:28

So MCP, there are lots of concerns regarding the security concerns. So most of the time what happens is you might have prompt injection or tools injection which we will discuss when we talk about MCP. So MCP is just like function calling or function calling.

So you have one server. Inside that one server you have 10 tools. So what in tool was used? What was the input of it and what was the output of it? If you want to track that, we'll be using metadata. What we will do today is we'll be using feedback score. If you are working with any evals and if you have scores and if you want to track that, you can use the feedback score. So instead of writing it in the single lines, let's define a function.

Def get context precision.

But before that, let's import OK.

From optic import track.

Uh.

OPIC context. So last time how did we use OPIC? So we define OPIC as a tracer. Then LLM has something called as callback inside callback you are giving. So here I'll show one simple way to do it. So let's suppose you have one LLM response.

Hardip Patel 54:47

Mhm.

Yeah.

Tarun Jain 54:58

Right, you have LLM response. You have a query.

And inside this what will you do? You'll just do response equals to LLM. Don't copy this.

Query then you will return response dot content. So if I want to track this in simple ways, what will I do? I'll just use a decorator track. That's it.

Tirth 55:22

Hmm.

Hardip Patel 55:23

Mm.

Tarun Jain 55:24

So let me see if I have the import statements.

OK, it's not there. So can anyone tell me what are the three import statements we need here? I mean environment variables for OPIC.

Tirth 55:38
I'll be.

Hardip Patel 55:38
OK, URL, username and project name.

Tirth 55:41
Open.

Ajay Patel 55:43
E.

Tarun Jain 55:43
But what is that exact thing?

Hardip Patel 55:48
Uh.

Tarun Jain 55:48
It's clear you have OPKPI key then.

Hardip Patel 55:51
Hmm, yeah.
You are in box, OK like uh.

RamKrishna Bhatt 55:56
Project name thing.

Ronak Makwana 55:56
Or picking order.

Tarun Jain 55:59

Correct project name.

Hardip Patel 56:01

Right.

Tarun Jain 56:03

And one more.

Hardip Patel 56:04

Workspace.

Tarun Jain 56:08

Yeah, it's workspace.

Open workspace.

So Opic API is already there inside the environment variable.

Hardip Patel 56:21

Stop.

Tarun Jain 56:22

I mean the secrets.

I just do.

Hardip Patel 56:26

Yes.

Tarun Jain 56:29

User data dot get OP KPI key.

So the workspace is nothing but the username, which is Tarun Arvind.

And the project name is nothing but.

Yeah.

And this should be OPIC project name, not just project name.

Hardip Patel 56:50

Yeah.

Yes.

Tarun Jain 56:55

So OPQPIT, opaque workspace, everything should be in capital.

And opic project name. Yeah, these are the three variables. And now if I just.

Run this track and if I call this LLM response.

Who is the main character of Demon Slayer? So how many of you watch Demon Slayer? All of you, yeah.

Hardip Patel 57:24

Yeah.

Tirth 57:31

Yeah, no, not all of that. But you watch it then.

Hardip Patel 57:34

Mhm.

You started with.

Tirth 57:41

Yesterday was the Margi. Margi's not an animal fan. Margi, you don't know. I don't think so.

Hardip Patel 57:46

Mm.

Tarun Jain 57:48

OK, sorry, this should be dot invoke. I forgot invoke and this is dot content.

Hardip Patel 57:50

OK.

Mm-hmm.

Tirth 58:00

She'll come to me.

Tarun Jain 58:00

So now if I come back to OK and if I refresh.

So if you see you have an entry.

But if you see we didn't use any, uh, what do you call?

Callbacks or anything. The simple thing is you have tracked. That's it.

Tirth 58:15

For context, so in call back we have to give.

Yeah.

Tarun Jain 58:22

So now what we'll do is let's define track.

I'll delete the cell. I just wanted to show the functionality of track.

Hardip Patel 58:38

OK then.

Tarun Jain 58:42

What happened?

Hardip Patel 58:44

Nothing. Nothing. We were on.

Tarun Jain 58:46

OK.

So you have get context precision, then you have response.

Hardip Patel 58:53

Mhm.

Tarun Jain 58:55

So response has three input variables, one is context, one more is answer and one more is query. And inside this I'll first define single term.

Hardip Patel 58:56
Yes.
OK.

Tarun Jain 59:05
Singleton is nothing but Singleton user input response, retrieve context. You have query answer and response. Then you have context precision, LLM context precision without reference. Just define LLM equals to eval LLM.

Hardip Patel 59:14
OK.

Tarun Jain 59:27
Now you just have to get this scores. So let me see if I define scores.
It should be numpy.
What is the type of this?
OK, just float.
So scores equals to just copy this range.
So instead of faithful, what we'll do is I'll copy it as context precision, then dot Singleton a score and the Singleton is nothing but this one.
And then after you have this, this define Opic context.
dot.
Update current trees. Inside this you have two input variables. If you want to define metadata, you can define metadata. But what do we need? We need feedback score.
I will just copy this feedback scores.

Hardip Patel 1:00:21
OK.
Yeah.

Ajay Patel 1:00:39

List.

Dictionary of school.

Tarun Jain 1:00:46

One is list and inside this I'll define certain name. So this name can be context precision.

Hardip Patel 1:00:52

But.

Tarun Jain 1:00:56

And then I can have scores.

Sorry, this should be value. Value is nothing but a scores. That's it.

And you can also return scores.

So this is the only new line we wrote remaining the same. Your single turn is same.

Then instead of faithfulness I changed it into LLM context precision. So by default it uses open AI again. So what I need to do, I need to replace it with Gemini.

Then scores equals to context precision, get the score, give the sample and then what you need to do is if you want to update. So all this is called traces. So if you see this keyword trace you want to update this trace.

Hardip Patel 1:01:41

OK.

Tarun Jain 1:01:45

And when you update the trace, when you check there are only two input variables.

Either you can define metadata, not either. You can define metadata and feedback scores together as well. But the data type of feedback score is list.

Hardip Patel 1:01:55

Yeah.

But.

Tarun Jain 1:02:00

Inside list you need to have a dictionary and this dictionary is nothing but context

precision. Value is nothing but the score. That's it.
Sorry, this should be async. If I'm using asynchronous programming inside this function, which is await, the function should also be async.
This syntax, everyone are comfortable, right? Async and await.

Tirth 1:02:29
Yeah, yes.

Ajay Patel 1:02:32
Yeah, yeah.

Hardip Patel 1:02:32
Mm.

Tarun Jain 1:02:33
So now what I'll do is I will just call this function get context precision.
OK, I made the mistake here. This is precision.
Spelling mistake and just pass the response.

Hardip Patel 1:02:50
Hmm.

Tarun Jain 1:02:57
Uh, now if I come back here.
And if I refresh.
Where is the entry?
Print scores.

Hardip Patel 1:03:19
Yes.

Tirth 1:03:26
You have to await that as it is a nice thing.

Hardip Patel 1:03:30

Yeah.

I'm getting invalid syntax.

Tarun Jain 1:03:39

Correct. Now it's working.

So now if I open this and if I click on.

Hardip Patel 1:03:48

OK.

Tarun Jain 1:03:49

Feedback score. You have context regression and this score. There is also something called as reason. If in case you need any reason or justification, you can add it, but I don't think we have any reason. If not, what I'll do is I'll just add.

Hardip Patel 1:03:58

Right.

Tarun Jain 1:04:05

Reason as the output whatever I got response of answer.

Hardip Patel 1:04:08

It.

Tarun Jain 1:04:14

Oh, what are you getting, Ardip?

Hardip Patel 1:04:15

Yeah.

I'm getting syntax error, but I think it's my fault. No problem, I'll I'll check.

Tarun Jain 1:04:26

And now if you see you have a reason, reason is nothing but whatever output I got.

Hardip Patel 1:04:33

OK, I got it.

Tarun Jain 1:04:40

That's it. One is we use track and apart from track we also have OP context where you update the trace.

Now this precision is over all the context that you have. If you want to try precision for first K So what you can do is we usually define like this precision at K1.

Hardip Patel 1:05:04

Yeah.

Tarun Jain 1:05:07

Then precision at K2. So now what will happen when you're defining the single turn? Instead of giving the entire context, what will be your syntax? You will make it as list.

Hardip Patel 1:05:14

Yes.

Tarun Jain 1:05:22

And then you will get the zeroth index. So now what is this function?

So if I want to replace this function, I'll just make it as get context precision at K is.

At K1, is this clear?

Tirth 1:05:46

8.

Tarun Jain 1:05:47

So what is the data type of response context?

Hardip Patel 1:05:49

M.

Tirth 1:05:51

List of list of thing.

Tarun Jain 1:05:54

This is a list, right? So if I do 0 index, what is happening now?

Tirth 1:05:56

Yeah.

It's a string, just one.

Tarun Jain 1:05:59

What I did, I added one more thing. That's it because retrieve context data type expected to be list.

Tirth 1:06:03

OK.

Ajay Patel 1:06:09

List.

Tarun Jain 1:06:10

So now this function is for get context precision K1.

So I'll add name context precision.

Hardip Patel 1:06:15

OK.

Tarun Jain 1:06:19

K1 and most of the time what you can do is you can define name also for this. I'll define this as precision.

Hardip Patel 1:06:30

No.

Tarun Jain 1:06:35

Our response is same.

Hardip Patel 1:06:37

Mhm.

Tarun Jain 1:06:39

So this should be 90% if you see this is 90%.

So now you have a name and now you can use the same name and upend everything.

So name precision at K is 2. Now this will be one.

Hardip Patel 1:06:58

OK.

Tarun Jain 1:07:06

And I'm using the same tracking.

This is 0. So now what is happening here? In my second context there is nothing related to the given user query.

Tirth 1:07:16

Yeah.

Hardip Patel 1:07:18

Yeah.

OK.

Tarun Jain 1:07:31

Is this clear? What precision at case and what precision as this?

Tirth 1:07:36

Yes, yes, but what do we do with this like?

Hardip Patel 1:07:39

OK.

Tarun Jain 1:07:42

So let's afford to.

Tirth 1:07:42

We we get it like we get it at LMLMS. So then we are getting the position of the context and everything. But how would that help?

Tarun Jain 1:07:51

So obviously for one user query it won't help. What you need to do is you have to run it in the batches. So let's suppose you're working in the report generation, right? So for report generation, let's suppose you took on experimentation one.

Tirth 1:08:01

Mhm.
Mhm.

Tarun Jain 1:08:08

So in this experimentation one, you just used hybrid search.

Hardip Patel 1:08:13

Yes.

Tarun Jain 1:08:15

And you used re-ranking. Now what will you do? You will run this thing. You have your own data set. Once you have your own data set, you will define the same metrics.

Tirth 1:08:15

M.
Mhm.

Hardip Patel 1:08:24

OK.

OK.

Tarun Jain 1:08:28

Let's suppose you have faithfulness. You will define context precision, you will get certain value. So if this is more than 75%, let's suppose for your experimentation one you got faithfulness as.

Tirth 1:08:32

OK.

Tarun Jain 1:08:43

75%.

Hardip Patel 1:08:44

M.

Tarun Jain 1:08:47

And Prithishanas.

Hardip Patel 1:08:48

Yeah.

Tarun Jain 1:08:51

79%. Now you want to improvise this and let's suppose you got a different experimentation. In this experimentation you are doing.

Tirth 1:08:52

Mhm.

Mhm.

Yes.

Hardip Patel 1:09:02
Yes.

Tarun Jain 1:09:04
What was the approach? You have metadata filtering, right? I told you yesterday to use summary.

Tirth 1:09:10
Right.

Tarun Jain 1:09:10
So you're doing filtering technique. You're adding summary with the experimentation one.

Tirth 1:09:15
Mhm.

Hardip Patel 1:09:15
Hmm.

Tarun Jain 1:09:17
Now your score has increased to 78% and our precision is 82%. So this whatever scores you're getting right, this will tell you which experimentation was working fine.

Tirth 1:09:31
Hmm.

Tarun Jain 1:09:33
And once that particular approach that you have taken has more score, that is what you will deploy in your main application.

Tirth 1:09:33
Mhm.

Hardip Patel 1:09:37

Yeah.

Tirth 1:09:41

I see. I see.

Hardip Patel 1:09:41

OK.

Tarun Jain 1:09:43

So now let's suppose I took some of the research experimentation. In most of the time we use self reg, but self reg doesn't actually help us, right? You have scores like faithfulness as 76%.

Hardip Patel 1:09:44

Mhm.

Tirth 1:09:47

Um.

Hardip Patel 1:09:56

Yeah.

Tarun Jain 1:10:00

And precision as let's suppose I have 85% even though I will not go with this approach. Why? Then I will check the latency and I'm pretty much sure the latency of self flag will be more than 15 seconds.

Tirth 1:10:10

E.

OK.

Tarun Jain 1:10:18

So we usually get the particular, we export this as CSV and then we check experimentation, how much score was there, the precision, what was the faithfulness and we also have rag. So rag rad is basically true or false.

Hardip Patel 1:10:18
Yes.

Tarun Jain 1:10:35
So here we'll just have two our polls.

Tirth 1:10:39
So this whole thing is for deploying a strategy to get better output and then we do not deploy this, we deploy the strategy on production.

Hardip Patel 1:10:42
OK.

Tarun Jain 1:10:49
Correct. Because when it comes to machine learning, you can't always bet on one approach, right? In some cases, even if you use traditional drag, traditional drag in the sense no hybrid search, no re-ranking, whatever we did just now.

Tirth 1:10:53
M.
M.
Hmm.

Tarun Jain 1:11:05
OK, this is actually hybrid search what we did here. But whatever we did first, right, the first web page, if that traditional drag has more than 80% for both your faithfulness and precision, you don't even have to worry about hybrid search.

Tirth 1:11:11
M.

Hardip Patel 1:11:12

Yeah.

Mm.

Tirth 1:11:18

Hmm.

Tarun Jain 1:11:25

So this is this all depends on for your data which experimentation is working fine that you will deploy and it's not like you build your code once and that is your final code 101% you will try to improvise. So what is the number? How do you know that you have improved?

Hardip Patel 1:11:26

OK.

Tirth 1:11:29

M.

Tarun Jain 1:11:45

So you can't judge that on just one prompt or just five or six prompts. So you need to have at least 40 to 50 prompts.

Tirth 1:11:49

True.

Hardip Patel 1:11:51

Yeah.

Tarun Jain 1:11:52

So this is just for which approach works done.

Tirth 1:11:57

Understood.

Tarun Jain 1:11:58

So now let's just quickly look at the batches. I have the code ready. Is it visible in your case? Are you able to? OK, so when it comes to generating synthetic data set, I will always prefer to use Open AI. So if in case you have Azure Open AI.

Hardip Patel 1:12:04
Yeah.

Tirth 1:12:05
Indispensible, yes.

Hardip Patel 1:12:07
OK.
OK.

Tarun Jain 1:12:16
I'll refer if you can use Azure Open AI because that has some free credits and you can use it. If you're using any free elements right like Gemini, it will break in between. Whether it is Gemini or Samba Nova or Grok, they won't generate all the.

Hardip Patel 1:12:20
OK.

Tirth 1:12:22
Yeah.

Tarun Jain 1:12:32
Questions at once, but if in case you have a version of Gemini 2.5 Pro then you can feel free to use that. So again I'm starting with the same stuff. I'm starting with web-based loader. Why? Because when I'm generating synthetic data set, obviously I need my main data.

Tirth 1:12:34
Yeah.
OK.

Tarun Jain 1:12:50

In the previous code we just did inference. So in inference we have search context. Search context is only taking user query, it's not taking the data. So now what I need to do is when I create synthetic data set, I need my actual data.

Where is it? So I'm starting with the Excel data set and then you have documents. So what will be the length of documents?

Tirth 1:13:15

So.

Hardip Patel 1:13:15

So.

Tarun Jain 1:13:16

So this will be too. So till here it's the same code. Now we are starting with the main data set creation. So you have ragas dot test set. You're importing test generator. This has LLM within it.

Tirth 1:13:22

M.

Tarun Jain 1:13:33

So you're using LLM to generate the data set. So what are the?

Tirth 1:13:34

OK.

Hardip Patel 1:13:35

Yeah.

Tarun Jain 1:13:39

Three rules of generating data set.

If in case you are writing your own prompt.

First one is diverse, second is different scenarios.

Ronak Makwana 1:13:52
Hey.

Hardip Patel 1:13:55
Hmm.

Tarun Jain 1:13:58
And 3rd is different persona. So what is the example of different scenario? You have singleton.

Tirth 1:14:02
OK.

Tarun Jain 1:14:08
You have multi turn, you have no answer itself and then you have invalid choices and you can also add some negative edge cases. So negative edge cases. What is the possible scenarios one can encounter?

Hardip Patel 1:14:12
Hmm.
Mm.

Tarun Jain 1:14:24
And then you have different personas. So different persona is nothing but are you targeting technical folks or are you targeting non-technical folks then expert level. So persona you understood, so you're defining persona.

Tirth 1:14:24
Yeah.

Hardip Patel 1:14:31
Um.
OK.
But.

Tarun Jain 1:14:39

Then you have single op. Single op is nothing but single turn itself. So you have two things itself. One is single op or you have multi op.

Hardip Patel 1:14:50

Hmm.

Tarun Jain 1:14:51

So multi op will have two portions within one portion. Single op is nothing but just one portion. So these are the new import statements. So this is same since we are using line chain. I'm using it from line chain the LLM.

Tirth 1:15:02

Oh.

Hardip Patel 1:15:05

Yes.

Tarun Jain 1:15:07

And then embedding also I'm using open AI and this is the language in open AI. Instead of Google generative AI, I'm using chat open AI. So is this clear the import statements?

Hardip Patel 1:15:19

Yes.

Ronak Makwana 1:15:19

Yes.

Tarun Jain 1:15:20

OK then I'm just saving my open API key inside environment variable and you just have to define your LLM. So LLM is nothing but open AI GPD photo and for embedding directly you're using open AI.
Functionality which is nothing but your client and here if you see open AI

embeddings whatever Ragas is there, you just have to pass the client. So what is the input variables in client? Just API key.

Ronak Makwana 1:15:49

OK.

Tarun Jain 1:15:49

And that API key saved inside environment variable.

Tirth 1:15:54

OK.

Tarun Jain 1:15:54

So you have embedding, you have LLM. So this LLM currently is in line chain.

So what you need to do is you have to. OK, I don't think we need this thing.

Because by default you have Open AI LLM itself, so you don't have to convert this.

Hardip Patel 1:16:13

Mm-hmm. For me, showing.

Tarun Jain 1:16:15

So if you see, I mean if you look at a set generator, here the LLM is nothing but your base ragas LLM, so you don't specifically need.

Chat open AI. By default it is open AI itself, but if in case you want to change the model name to GPT photo then you can change it. So if you want to use mini you can use mini also.

Tirth 1:16:43

OK.

Tarun Jain 1:16:43

Is this clear?

Till here is it clear? We are just defining the embedding model and LLM and then we are defining the persona. So if you want to define multiple persona, just copy this, paste it here. So now what is the persona that I've defined?

Tirth 1:16:50

Yes, yes.

Tarun Jain 1:17:02

The data set what I have is related to website and what is the best suited role. I thought the best suited role is expert software assistant who has solid foundation in SAS and customer service assistant.

Just like that, if you think there are different personas you can create, you can create those personas.

Ronak Makwana 1:17:26

OK.

Tarun Jain 1:17:27

So this is nothing but again your parenting.

Hardip Patel 1:17:30

OK.

Tarun Jain 1:17:31

So this whatever class you have, if you look at their code base, it is nothing but a pydantic class where you're defining a name and you're defining a role description, but it should be inside a list.

Tirth 1:17:31

M.

Tarun Jain 1:17:42

And inside this list you can define multiple personas and once this is done, this is only one additional functionality that you are adding. So basically what you're trying to do is do you know what is NER? We did discuss NER in NLP.

Hardip Patel 1:17:46

OK.

Tirth 1:18:01

We had BP. We had

Hardip Patel 1:18:01

Forgot.

Tarun Jain 1:18:06

This was before percent similarity.

So it's called a named entity recognition.

So named entity recognition totally has four plus. One is name, then you have organization.

Then you had geopolitical. Now you remember geopolitical locate.

Tirth 1:18:32

Yes, yes, yes.

Hardip Patel 1:18:34

Yeah.

Tarun Jain 1:18:35

What was one more name, organization, geopolitical location and?

Tirth 1:18:39

There was organization and geopolitical location was confusing. We had one more.

Tarun Jain 1:18:44

Oh.

Tirth 1:18:46

Um.

Tarun Jain 1:18:47

Athiyan Tech sessions.

Yes.

Name OK, location is different and geopolitical entity. So these are the four things only name, organization, geopolitical.

Tirth 1:19:08

Location.

Tarun Jain 1:19:11

And then location. So what are you trying to do is there are certain NLP techniques if you want to apply like stop words you can use if in case you want to use headline splitter. So this is again one of the NLP technique and if you want to extract the NER, NER is nothing but name organization. So these are some of the NLP.
Pre-processing techniques. If you want to apply those, you you can apply. If not, you can also skip this. But these are the two important things as per the documentation. One is headline splitter and one more is NER extractor. Even if you don't define this NER extractor, this is by default.
NER extracted will happen by default.
So that's it. So we define LL, we define embedding, we define persona, and once you define persona, you're also defining transform, which is optional. And once this is done, you can directly start with your generator. So generator will take your embedding model.

Hardip Patel 1:20:11

In

Tarun Jain 1:20:13

And a persona. So once you have the generator, what is the distribution you need? So distribution is nothing but we needed a single op. So either you can define single op, if not you can define multi op and the input variable is nothing but LLM.
And you just have to define 1.0. So 1.0 is nothing but all the combinations that you're trying. It should just be single op. If in case you have multi up here, let's suppose you have distribution, you're passing single op and you're also passing multi op. So what will be the ratio? You can keep it.
1.0 comma.
Sorry it won't be 1.0, it will be 0.7 comma 0.3. So now what is this ratio? 0.7 is for single op, then remaining 30% of the question is for multi op. This is similar to what

we saw in MMR.

So in MMR we define 0.5. That means 50% is diverse.

And 50% is relevance.

Same when you define multiple things, 0.7 is nothing but single lock.

Then 0.3 is for your multi-op, but you have to import that.

If in case anyone needs that example, you can check out the documentation. If not, you can let me know. I will just change this distribution. This is again optional step.

When you are generating the final documents, you're just adding it here.

Here is it clear we are just defining the we are just instantiating it LLM then personas embedding model transform and once you define your generator you just have to use generator dot generate with line chain documents. Why? Because the documents that you have it is in line chain.

Hardip Patel 1:21:53

The.

Yeah.

Yeah.

Tarun Jain 1:22:11

So what is this variable? This variable is nothing but web-based loader which is in form of documents.

You have document inside that you have page content and then you have metadata.

So if you want to generate in this particular format, you have a function called generate with langchain docs and then if you see you're applying the headline splitter, then applying NER, then creating scenarios and persona is already defined and this is where you have all the details.

Persona scenarios and diversity and once that is done you can just use data set to pandas.

So now if you look at your variables, you have user input, you have reference context, you will not have response. This is something that I was testing. So you will have user input, you will have reference context and you will have reference.

Ronak Makwana 1:22:51

Because.

Hardip Patel 1:22:54

Yes.

Yeah.

So.

Tarun Jain 1:23:07

So let me just write it so when you create the data set.

You will have reference context.

You will have user input.

And you have reference.

Ronak Makwana 1:23:26

OK.

Tarun Jain 1:23:26

So what are the other two variables which are pending here?

Hardip Patel 1:23:29

Yeah.

Ronak Makwana 1:23:30

OK.

Tarun Jain 1:23:34

So I told 5 variables, right? Which are the two pending here?

Hardip Patel 1:23:36

Hmm.

Ronak Makwana 1:23:39

Mhm.

Hardip Patel 1:23:39

OK.

Ronak Makwana 1:23:44

One minute.

Tirth 1:23:46

The response one is pending now. We don't have the response one.

Tarun Jain 1:23:47

Uh, hello.

Our response is pending and one more.

Tirth 1:23:53

Um.

Delete the score.

Ronak Makwana 1:23:59

Yes.

Tarun Jain 1:24:02

So in Singleton example, if you look at the Singleton, you have user input, you have response and you have retrieved context.

Hardip Patel 1:24:02

OK.

OK.

Tirth 1:24:11

Hmm.

Tarun Jain 1:24:11

And here if you see you have reference context, you have user input, you have reference, you have response. What is missing? Retrieved context.

Hardip Patel 1:24:20

OK.

Tirth 1:24:23

OK.

Tarun Jain 1:24:24

So this is your actual context.

Hardip Patel 1:24:27

But.

Tarun Jain 1:24:27

Actual context. This is the actual response.

And this is the.

Prompt or user queries?

Hardip Patel 1:24:39

With.

Tarun Jain 1:24:40

Generated by LLM.

Now this is generated by what you call the data set creation. Now what you're supposed to do is if you want to run evals, first thing is how will you run evals?

Hardip Patel 1:24:48

Yeah.

Yes.

Yeah, yeah.

Tarun Jain 1:24:56

You have to run the wells.

Hardip Patel 1:24:59

Yeah.

Tarun Jain 1:25:00

Against the model answer and the data set. So what you have as of now is you only have the data set response which is your reference context and reference. Now what do you need? You need the model answer.
So this model answer is nothing but response and retrieve context. So what will you do? You will just run your rack pipeline and ask all these questions.
Is this clear? So whatever pipeline you have created from land graph, what you will do is you will run a loop. In that loop you will ask all these questions.

Tirth 1:25:30
Mhm.

Tarun Jain 1:25:39
And once you run all these questions, what will you get? You will get the response and you will get the retrieved context.

Tirth 1:25:45
Hmm.

Tarun Jain 1:25:47
So if you see here retrieve context is there but it is none.
Because you didn't run it against your input, so this is what we need to update. So run a loop. When you run a loop, retrieve context will be overwritten by your Langraf code and if you scroll below you will also find response.
So if you see a response here it is answer. But technically when you print this right data set dot samples sample eval sample then inside eval sample you have retrieve context which is none. Then response is none. Now why it shows answer here?

Tirth 1:26:15
Mhm.

Tarun Jain 1:26:28
When I was testing this, I edited like it like this.
So now what will happen? In my first index it will make it answer. Now if I do 0 index, I mean if I do the second index and if I upend answer this output will be answered.
So this is what I was trying to do. Now what will you do? Just print the eval data set.

Tirth 1:26:37

Yeah.

Hardip Patel 1:26:40

Yeah.

Tarun Jain 1:26:50

So when you print the eval data set, what do you have? You have eval sample. Inside this you have user input. Retrieve context is none. So just ask all these questions. Run your Langra code once you run the Langra code.

Tirth 1:27:00

8.

Tarun Jain 1:27:07

You have the query which is inside eval sample so you can just cross check it here. So when you run the for loop right this for loop as eval sample from eval sample get the user input as it in your graph.

Hardip Patel 1:27:15

Mm.

Yes.

Tarun Jain 1:27:23

Graph dot invoke as the user query. Once you ask the user query, what does this response contain? For this response we'll have response dot query.

Then response.

Hardip Patel 1:27:38

Yes.

Tarun Jain 1:27:39

dot answer and then response dot context. So context is nothing but your retrieve context. Now response is nothing but your response answer. Once you append it now if you print your data set to pandas you will see this to field.

Is this clear?

So once what you can do is you can check the code. The reason why I didn't run this code live is because I have to generate the data set again, which is bit of time consuming.

Hardip Patel 1:28:06

So.

TJ **Tarun Jain** 1:28:11

Is this clear? This will take at least around 1:00 to 2:00 minutes. Hardly 2 minutes.

Hardip Patel 1:28:12

Yes.

Tirth 1:28:13

E.

Hardip Patel 1:28:17

OK.

Yeah.

Tirth 1:28:20

But.

TJ **Tarun Jain** 1:28:20

And once this is done, this particular code cell is there, right? If you have 20 questions, how much time does it take for one question? 2 seconds. So if you have 20 questions, this cell will take 40 seconds.

Hardip Patel 1:28:26

OK.

TJ **Tarun Jain** 1:28:36

Why? Because you are appending your retrieve context and response.

Hardip Patel 1:28:38

OK.

Tirth 1:28:42

Mhm.

Tarun Jain 1:28:42

So just keep this in mind. Retrieve context and response. It is coming from LLM generated response which is your blangraph user input reference context and reference. It is coming from blagas test case generation.

Hardip Patel 1:28:45

It.

That.

Tarun Jain 1:28:57

Is this clear? You need all these five things to execute your faithfulness, factual correctness. You can take all the metrics which is available in Ragas and once this is done you just have to run evaluate. Evaluate is a function from Ragas itself.

Tirth 1:28:59

Mhm.

Tarun Jain 1:29:15

Pass your data set, which is nothing but data set to evaluation data set. Then whatever metrics you have, define faithfulness, factual correctness, LLM context, recall. If you want to add precision, you can also add precision. Then if there is any LLM that you want to use, which is your base LLM, just add that LLM and once. Once you execute this cell, you will have your final results which is faithfulness, factual correctness and context recall.

Hardip Patel 1:29:47

Hmm.

Tarun Jain 1:29:49

So you can try this once. If in case anyone is facing issues, we can take up the question tomorrow. But with this we are completed with RAG and as well as evals. The only pending thing is agents and MCP, agent, MCP and memory.

Hardip Patel 1:29:59

Yes.

OK.

Tirth 1:30:05

Yes.

We need to go through this.

Tarun Jain 1:30:09

So this is something that will so try to run this. The only thing is I didn't want to run this live just because there are two cell where it is time consuming. One is this part and the another part is.

Tirth 1:30:12

OK.

Yeah.

Hardip Patel 1:30:17

Mm.

Mhm.

Tirth 1:30:21

E.

Tarun Jain 1:30:25

This for loop.

And if you notice here, I'm again creating the workflow because I don't need context relevance and answer relevance. If I add that, it will take 5 seconds extra. So that's the reason why I removed that. The only two things we need is context, context and answer. There is no need for.

Hardip Patel 1:30:31
Oh.

Tarun Jain 1:30:45
Context relevance and answer relevance here.

Hardip Patel 1:30:59
Yes.

Tarun Jain 1:31:00
Food recipe. Try to run the eval and see how good your program is.

Tirth 1:31:07
OK.

Hardip Patel 1:31:08
She.
OK.

Tarun Jain 1:31:15
Yeah, that's it. So any questions?

Hardip Patel 1:31:15
Yeah.

Tarun Jain 1:31:19
One more thing what you can try is don't define LLM here because open AI is by default. I was testing with Gemini but Gemini crashed multiple times. So that's the reason why you are seeing LLM defined. If not, if you are using open AI, don't define this one.

Hardip Patel 1:31:28

Hmm.

Tirth 1:31:29

Mm.

Tarun Jain 1:31:35

And don't even define any LLM inside this parameters. If you see here, I'm not defining LLM here.

If not, all these things requires an LLL.

So you're also don't have to define LLM.

Hardip Patel 1:31:47

OK.

Tarun Jain 1:31:51

So this is one part and.

Here also don't define the LLM which is our test case generator.

OK.

Hardip Patel 1:32:03

Yeah.

Tarun Jain 1:32:06

And I'll probably refer if you can use Azure Open AI. If in case you have Open AI credits, feel free to use Open AI.

Hardip Patel 1:32:11

OK.

Tirth 1:32:19

This was a lot today. I'll have to go through it.

Hardip Patel 1:32:19

Mm-hmm. Uh, then.

Yes.

And.

Tarun Jain 1:32:27

Yeah, eval is like usually eval you run only once. It's like last thing that you have to run.

Tirth 1:32:33

Mm.

Hardip Patel 1:32:33

OK.

Tirth 1:32:37

That's good.

Hardip Patel 1:32:37

Tarun for the the assignment thing like if you want to run it on a stream late right? If we are doing the hybrid search like we have to load the BM25 and Gina models.

So if we are doing that then stream later is giving up because of the I guess out of memory or something like that. Is there a way we can? So what I have done currently is like I am using similarity search from.

Directly from the quadrant. Is it OK to do or?

Tarun Jain 1:33:18

That is fine. I stream it in the sense when you deploy it is crashing or.

Hardip Patel 1:33:23

Yeah, when when we deploy it, it runs for a minute or or something, but for some of us like it isn't even running.

Um, it.

Tarun Jain 1:33:33

No. So when you deploy, you only have 1GB of memory at Max and whatever models you're using, right? Zena and BM25, that is sufficient.

Hardip Patel 1:33:38
Yeah, and.
Yeah, it should be sufficient, but for some of us like it, it just gives up. OK, so it could it be we need to.

Tarun Jain 1:33:52
So once can you share with this logs? So every stream lit tab right? Every stream lit tab has logs so.

Hardip Patel 1:33:57
So.
Yeah.

Tarun Jain 1:34:04
So what you can do is let's open any stream data app when you click on this right. So whatever is there, download that logs and you can share me that logs.

Hardip Patel 1:34:13
Yeah, on the manage one, yeah.

Tarun Jain 1:34:17
Just share me this TXV file. I'll check it once.

Hardip Patel 1:34:18
OK, fine.

Tarun Jain 1:34:22
And also what you can do is you can share your code base. I'll check that once.

Hardip Patel 1:34:23
Right.

OK, OK, I think the the the force force one also had the problem right? The same problem Ronak.

Tarun Jain 1:34:30

But.

No, it was working. No, it's working on screen.

Ronak Makwana 1:34:36

Yes, and.

Yeah, ask.

Yes, after that I get an option that reset the memory. So I do that and it is working.

Tarun Jain 1:34:51

No, this I tested on stream that it was working. The only thing what I wanted to test was the filtering part because filter was used in the code.

Ronak Makwana 1:35:00

Yeah, I guess I have removed that, but yeah, you can test it.

Hardip Patel 1:35:06

Right.

Tarun Jain 1:35:08

OK, that I was testing it locally. I mean I created the I cloned your repo on stream it fine. Stream it. It was working fine.

Hardip Patel 1:35:09

Yes.

Ronak Makwana 1:35:13

Mm-hmm.

Hardip Patel 1:35:14

Yeah, exactly.

Yeah, I mean like stream lit is working fine, but when we deploy it.

Tarun Jain 1:35:23

OK, what you can do is you can send me this locks. I'll test it once.

Ronak Makwana 1:35:28

August, yeah.

Hardip Patel 1:35:28

OK.

Tarun Jain 1:35:28

So you see this TXT folder, just download and share that.

Hardip Patel 1:35:35

OK, fine. All right. Thank you.

Tarun Jain 1:35:37

Yeah. Anyone else completed anything like streamlet or even the influence?

Hardip Patel 1:35:39

Yeah.

Tirth 1:35:44

Still working on now I'm looking at this all emails and everything and I'm so much tempted to 1st implement this. So I'll I'll experiment with this and share the code with you.

Tarun Jain 1:35:55

Well, So what you can do is you can take some time in getting the. So if you have more than 65%, that means whatever app you have built, it is something that you can share or demo it to anyone. But if it is less than 65%, we'll have to see ways to improvise it.

Tirth 1:36:07

OK.

Understood. Understood. OK.

Hardip Patel 1:36:15

I didn't use the high speed search, I just use the similarities.

Tarun Jain 1:36:16

OK.

OK, you're using. I'll, I'll test this once and I'll let you know.

Hardip Patel 1:36:24

Yeah, but in local hugging face thing works, yeah.

Tarun Jain 1:36:29

Sentence and so much, right?

Hardip Patel 1:36:31

Yeah.

Tarun Jain 1:36:32

OK, I'll I'll check it and share feedback.

Hardip Patel 1:36:35

I'm using the Gemma 1B.

Tarun Jain 1:36:41

Yeah, OK, LLM is also you're using bugging phase.

Hardip Patel 1:36:46

Hmm.

Tarun Jain 1:36:47

You're using LLM for mugging face.

Hardip Patel 1:36:49

OK.

Yeah, I mean, like, I'm just borrowing Gemma from Hugging Face.

Tarun Jain 1:36:55
OK, I'll, I'll see and I'll let you know. I'll see the memo utilization of the code also.

Hardip Patel 1:36:56
OK.

OK, OK.

Tarun Jain 1:37:01
Yeah, because the maximum is 1 GB if you use only embedding.

Hardip Patel 1:37:07
No, it's working fine because I'm using Azure for deployed version.

Tarun Jain 1:37:12
OK, then fine.

Hardip Patel 1:37:14
This is working fine because I haven't used the hybrid search directly. I'm using this quadrant search, the MMR and what do we say the similarity search.

Tarun Jain 1:37:30
Yeah, OK.

Hardip Patel 1:37:32
Yeah, OK.

Tarun Jain 1:37:34
I'll check it and I'll let you know.

Hardip Patel 1:37:36
All right. Thank you.

Tarun Jain 1:37:37

Yeah. Thanks.

Tirth 1:37:39

Thank you. Thank you everyone.

Mitesh Rathod 1:37:40

OK.

Tarun Jain 1:37:41

Thank you.

Tirth 1:37:42

Thank you.

◉ **Margi Varmora** stopped transcription