Python and AI Power-Up Program Offline Class-20250825_113919-Meeting Recording

August 25, 2025, 6:09AM

2h 12m 6s

Mitesh Rathod started transcription

Tarun Jain 0:04

Yeah, now put up the and you can just mail it out for and.
So this over the idea, what are these?
Now he has to run a hotel department.
A4 to Soundmore.
So immediately there are two kind of sampling patterns. So I'll just open example. So the first sampling balance there will be sampling. I'll give you an example. Let's suppose.
So there are two kinds of samples. One is you have talking, then you have talking, right? And each sampling. Let's take an example. Of course there is a.
I need to have.
Cheeseburger, right? After a particular query, how many times was it repeated when you train the data, right? I need to have. I need to have free of shop.
Then if you want to predict the next word, what is the probability of the next word that you can predict? So what is the sampling parameter you use, right? So I need to have K then you can have 3 scales. So this is the ideal next word you are supposed to predict. So this is what is decided by top P and top K.
So top top which are the most preferable occurred tokens that can be used for our next word right other minute top K could be a one that means.
Once you give any sentence, how is it predicting the next row depends on the K value. It will do the sampling between the ohh once of it. So I need to have, but I'll have total 40 samples.
So this is what top and same top is a probability, right? 0.6. So usually how is this calculated?
Once you train any data, talking at your other, immediately probably. So let's suppose that we have.
Cake is burger.
And XYZ, right? So take a probability 0.98, whatever is the next word dependent. You

also try to calculate the three state values.
So this is your probability, right? Probability will be bunch of sentence. If you calculate the probability of the sentence, it should be equal to it will be one.
In one sentence.
The probability.
He is always equals to 1.
It's like example of that dice. Dice is the best example. Let's suppose where dice rolls around. What are the total possibilities? I want to roll a dice. I want to roll the probability.
3 by even number I have. So how many even number do we have? 3 right? 3 by 6 there is a 50% probability. Now what if it is not even?
No.
Which is odd, which is also 50% other. If you calculate both of them, it is 100, which is 1. So probability distribution of any given sentence is equal to one.
So again I have a sentence called. I need to have a K if the probabilities of this token is already generated, right? Because once you ask any question until full stop, it will take it as a one sentence tokenizer.
Our one sentence tokenizer for probability will be 100, right? So that is how you usually calculate tau P So tau P is just cumulative probability, right? Instead of the sum of one, it will be one and.
If you want more randomness, of course other sampling, then you can have somewhere around 0.95. This is the ideal value, right? Other sampling, you just don't want to be creative enough. You only want to stick with the data that you have.
Then you can have your top t to be 0.1.
Is this the top key?
Top is most occurred frequency based. This is probability based 0.1270.1 It's basically sampling.
I need to have cheesecakes. I'll only get few. I won't get in huge amount, right? So if you want to be creative enough, you want to see more data, then you can try to become creative. But creativeness is denying me. Hey, don't.
This is the data that you have and stick to it right? So the stick policy that you give is where you have T to be 0.1 or.
Top page one just from my perspective where we financial data at 4.1 means more because you are passing your own data and you don't want to adapt to the style of any other.

Top here by default it is 15 all the default. I'll write default again.
Some frameworks are 50, some frameworks it will be 40, but it is travel there best practice. Then you have 0.950.95 is default default is fine. 0.95 will not come your performing.
But of course as the sampling thermine again, then you can have top feed that is don't be creative enough with temperature. So there are combinations.
We'll just take the first value that we get from top P and then we can keep the top P01, even that works, that combination is different top P1 and top P.
Top K is 1 and top K is 0.1 and then the most critical and the top K frequency based frequency of what? So what I so what I.
So we will take first three minutes, first three minutes, but we only take the first three. Manage possibilities. So now let's suppose I will have the so since all the elements are automated.
My goal is to create next token, the token. So now vocabulary head. This is not user's. This is LLM specific. OK, so user has no there is no control on this. You can only control the sampling.
OK, which is random sampling. OK, although words are good, that words is so further right? LLM vocabulary. OK, so in the A vocabulary, there is no trouble there.
There is no flexibility for user to customize it. OK, although for customized connects like you need to build your own model. Here since LLM most of them are not.
They're not configurable or they're not getting it from scratch. You can only just play around with sampling, OK? Sampling is like, OK, OK, this is was my first set of results for this combination. The new combination can be.
Now for top key, top kit, I will have the and then you have the option of match, couch, bed, chair, car, bike, bucket. Now if I give top kit to be three, we could only choose between match, couch and bed. These values are mainly calculated based on probability ratio.
This will not be sorted. I will talk to you later. I will talk to you later. So when you select right during selection time, it will use the sorting. OK, it will not be like randomly. There is fourth for every probability.
So when you generate you'll have move OK, but when you pass it between models, let's suppose I define specifically, the talking will take it for you. So I need this 10 during the time it will start and then we then give that workout.
But this is not so also we should not compute.
He's only embeddings, embeddings is quotient in that. No pay by calculation. OK,

that is not probability, right? Probability is different.
Distance measurement is different. LLM will never do distance measurement, right? Which may see probability and statistics here.
And the next parameter is temperature, though is I mean temperature is only one parameter that many people can perform in terms of I open Google if you see you have temperature one.
If you see over on this creativity allowed in the responses, that means suppose we had the usage, we have a PDF which is in chemistry.
So you will need your model to be very creative enough to come up with some basic examples like, hey, age this to 10 year old kid, age this to five year old kid. This is not simplified, but it's better, right? For that you need creativeness. Now for finance related example, you don't need wrong response.
More of a specific stones I get. So during that time, what will you do? Your temperature will be as low as possible. Zero, 0.1, but never be more than 0.2 if you don't want any creativeness. And if you need creative, for example.
So, so the formula is almost same of what temperature does, but temperature is logic bias. There is another parameter which is always one.
So basically temperature is mainly on randomness. OK, if you keep 0.1, it will not randomly generate. You can remember that as well. So it's like machine learning. Once you training the model, same data, same model you have different vectors. So usually you don't want the vectors to change. OK, if I train my model or if I that vectors.
Should never change. If I keep zero, it should stay same. If I keep one or any other experiment, give me different samples. So the goal here is to test what kind of sample in the output.
sampling here, sampling here. So that is the that is the main agenda of sampling. So the difference between top P and temperature is temperature is on randomness, whereas top P is on probability.
There is no probability in temperature, temperature rate. There is a formula called sound. So either you can keep it -1 or one.
But probability the range is always between zero to one or either we have therefore they have more than one.
So probably probability will never increase more than one. So temperature will randomly. Some people will use 4242 doesn't even increase more constant value area, but most of the frameworks are using just in one.

OK, so in short, for example like education, you want to personalize learning, exercise regenerate which are very user friendly, diverse. So during that time the temperature is very high.
We want to come up with some ideas for marketing, for strategies. During that time, we'll want to think out-of-the-box, right? So during that time we'll have temperature right? Hey, think out-of-the-box.
But if you want your model to think within the box, hey, this is your context and this is the tone, don't add to it, or never give wrong weight. That is 0.0, zero is good.
If it doesn't support 0, 0.0, you have to get note of that and it doesn't support 0.05. And if you get error in that then 0.1, 0.1 will work. So you just have to see is it possible to have more than 0.1.
So these are common three parameters. By default it is 0.5 in the framework, but if you don't want any creativeness.
Your temperature will be 0.0.
Or you can be 0105 just because they don't support exact 0.
And now it is a new open source.
That we will be using today.
Some of the opens are primary. If you notice there are garbage users, there are garbage tokens. Let's suppose as what is the capital of India, New Delhi is the capital of India and Kuba there are some randomly generated tokens.
Like uh, New Delhi that we which we there are related to New Delhi because not required right? So if you want to have some kind of penalty, hey don't get any random problems, you have a uh.
Parameter called repetition.
Canal T equal to 1.1. If one is in there, it can be somewhere between 1.1 to 1.5.
He got all the names.
So all these things are better, which are keyword organs.
So when you do model dot invoke model dot generate. So in that time we give this as a keyword argument. When I say keyword argument it's called data type all it will be dictionary.
So only reputation, penalty and one more is return.
Text to be false. So yeah, some libraries not there, but if you're using and if you're defining any elements, you can define this parameter. And when I say you're given an initial egram, it's only applicable to open source. So if you're using Jeminar, open AI.
This is not required so return full text. Most of the times prompt is also added in your

response. So let's suppose I print response. So let's see some prompt and then you are getting the response. So if you don't want the user to see.
Prompt. See some prompt and prompt. Then you can directly return full text to the point. However you prove that means you will have prompt in your response. This is default value which I have to keep it false.
These two are again optional parameters only when you observe it. During that time you can do repetition penalty 1.1 says other OK return full text again when you observe the prompt response.
Either you can use rejects directly. You can use return for. Don't verbose is something on runtime. OK, so don't verbose in the sense.
Let's suppose we are executing an agent. If you want to see that, that's the. You want to see what is happening in the runtime and don't work in GPT prompting is like. Adapt to my style.
But in technical details, the neural network is just to see what is happening in the back end when you are influencing. For example, I asked a question. So if you want to see that, you keep.
So there are two differences, though verbose is different, but in terms of technical details, verbose is different.
So now what we'll do is these are the parameters was the important and there is one article here. I'll just recommend you guys to read this article. It's a combination like how do these parameters work together and there are some combinations that it like.
So it's a very good article. Uh, probably everyone needs to read this and.
I'll share this for our notebook, but now we will write the code live. In data we will find this URL, but I'll paste this URL.
but in my case I personally believe you guys in temperature. Top key, top key, what the temperature is something everyone has to use right because sometimes you want to be top of the box, sometimes you just want
Don't overkill, right? Sometimes overkilling will overkill your entire performance. So temperature is something that we should always do, but we didn't catch top key and top gear because top key and top gear.
Default values are very good, so you don't want to play around with those values. Then you can bring those values right? The elements are non-discriminative models, right? You have to avoid sampling.
So that's the only purpose of this new parameter.

OK, so there are three ways we will use a refresh model. Uh, we will use Mistel. Open self model.
I just like 3 models which are very good enough in one is Mitchell. Why Mitchell? Because again it's hybrid compared to other models that we have and apart from Mitchell there is one. One is good, but sometimes I think it's not that consistent enough.
The good thing about Kuen is it's very good in generating the responses. Gokivo response generate the effects good, but sometimes it's not confusing. And second thing is gokik quenk, not just Kuen, most of the Chinese models.
They overfit on, they'll be benchmarking. You want to evaluate some benchmarking. They've used the data set in many new models. So never trust the benchmark scores of Chinese models, right?
One is good, but if you want to jump between 10 and Mistel, benchmarking, Mistel is not that on par of point. But when you see the difference, right, Mistel will be like, hey, we have all that. Mistel is good in translation. Mistel is good in structure responses.
And there is one test which is called leader in the case tag.
Let's suppose you have.
Context. I will extract the context. How do context I have K value is 5. Every K value has 2000, two thousand something. Let's just imagine I have 10,000 tokens. OK. Now what I will do is in my first 2000 total I will place a needle. That needle is a query. So the haste act is the context that you have. The needle is your query. 1st 10K. If I place my question, am I getting the right response or not?
If you're getting, you take that meter, you place that meter between 3K to 8K. Then you take that meter and you go to 8K to 10K. So what Michel does is Michel has 128K64K models.
Till last it will give you the response, but when it comes to point and Deepak, once you text 20K, 30K, the context starts degrading. So Michelle has very good following. And if you see, if you see the comparison of Vishal, they're doing it with GPD four and four. So it is that good in terms of the testing, but when you see the performance.
And once you start using it, you will notice the difference between digital and well. 10,000 will open single response, single response. So usually they have the back K value will be K equal to 7 occupy the baghbadvita baghbadvita has 800 pages.
It will fetch 7K value. Every K will be 2014, so 2017 14,000 AK value. These are my

input tokens. It will be basically.
14 data. So you have to test it just to test the context in the data. So if you know 30K is the data retained.
Fifth thing is of course latency is issue for that. Let's suppose your goal is to have better performance and good responsibility and not latency. So you can increase your context like we have core.
And you have a model which understands context. So either you can have memory. Memory is also context, right? Memory context is also concrete error. So you can include your.
Memory component.
OK.
So we believe Michelle, and next it is B.
Deep sink is good if you're directly using the API, right? Deep sink API, but it's also open. But if you're using API, it's faster. It's not box logo. First it will generate thinking programs.
So then your GPU.
Or the API man.
For Deepchik, though Deepchik itself is direct. The new company deepchik.com.
Mhm.
So we actually took the API case to test box lower then then we migrated to Azure Foundry other Foundry you can deploy it. There also it's very slow right? So then we have to switch to.
What works the AI compared to deepseek and the Azure it had partial responses for deepseek and the pricing are almost similar because.
It's deployed on the some server right?
So there are three days before now. The first way is we use inference. So.
Mr. I said, well, Mail text 22 BA there, but how do I find it? You know this. OK, I like that. So Michelle has chat.
Your phone.
Contacting shoulder here.
Legacy Model 8X 2 BR M.
Let me see what is.
For to sell if you get a working code, but I did.
OK, it is there in the levels model bottom. We'll roll down a little bit more under open models, I guess. This is a best model.

No, no, no. Alternative model we have. Mr. Small.
Currently phone to Mr. Maul 2506.
So this is only 32K, but 32K will not be that much. We just want people to that will be 128K. OK, they release one quarter in other this month on this month, OK, so.
We have a we start with your 3.1, though you see it in the.
So how how do we figure out like because we have a model set of how do we figure out? So usually how do we get a better benchmarking? Because I know Vishal will deliver. If you go with benchmarking, it's like you'll have to compare Vishal with Quinn or Quinn will have better benchmarking for sure.
Right, 2.5, but how do we get on usually? So usually though there are.
So what we really do is we have the waiting at port for that, so we just have to change the modeling and play, but you get a heat map.
So in this heatmap, if there is any more green data, if there is any. OK.
We can do that. We can do that. We can do that because there are so many.
So it's like, but in my case, then I'll go with one. That's like.
But for us, Vishal, it's like when this is not so.
In some, in some of these cases, we mean Saheli.
Now which is replaced as Mistral medium 3.1 or locally under the area what we usually see if we can.
So hold on my list. I was testing. It was 14 hours ago.
Kolama is good. So you can use neat chick and Poin.
So when it comes to product, product build and we are supposed to choose one open source model, then it's for sure because most of the time product comes to agencies because no one will wait.
If you're using for yourself, you'll wait for sometimes thinking tokens, then you're getting the final response. That way, deep chick is obviously better. Here also, if you think of a latency issue, then it's always good to pick deep chick.
OK, but up for latency and you'll have to shift it some model, no doubt.
They didn't need it to open it properly. So usually all these models are open bits, open so you can find the model properly. 2.5 you can find it. The thing was they didn't need the license properly. So many people are not sure whether to use that model on.
Puji is one of the reasons why good calling is not working properly.
Or maybe they need support. So these are the models. Mission, Deep Sequence, Agra for latency. It should make 100% weak. Agra for open source.

Use like if it's instructed in that type open, then you will have to go go with.
You can pick but now since a new model was released, this one, you can try the 3.1 parameters. So you copy this model and you just search it here.
And then I mean.
Oh, you could see this is the model, but the normal.
Locally in the room like CDU days.
And if you just want to use orama then.
Oorama Asbad Library.
Yes, Sir.
But most of the model links, exact model links are added in the.
And if you want to use on them, then like uh, hold on. So conference contacts or any otherwise they don't.
On them for all the lightweight. We can't offer cost. Then you'll be like.
So mostly people will correspond.
When it means billions of, let's suppose, because 70 in the India, means more data.
More data means text data.
And more detailed data they claimed it for longer time. So the size of the so the model size of 32 billion parameter it will be BCPA but whereas 7B it is at 4.7 BB.
So data size increase and then your model because it's for longer period of time in this complete this 4.73 or 4K.
Uh, we'll see that. Uh, it's my initial total add 8 to.
So first, um, the operation will use interest time. Second, we will run locally locally in the sense.
We will load model locally.
And third is your.
We will use a pin coding.
And decoding logic.
Third also we will run locally, but we will directly use pipeline.
Recommended.
So inference when API user yoga you will have to get HR token. So whatever model of this it will run on a new phase server local mini yoga and it will be but still inference line will not collect any data.
Data collector not collecting data in terms of. So if you have then you have the local one. How to run the entire model locally? So in our case we will use.
V3.

Vishal 7V Vishal 7V command support function. So can we open this URL Vishal 17? 03.
So you'll have to login into Aginjase or account create. Initially probably you will have to grant access. Nowadays every time like if you want to use any open source model. You want to agree to the license and then use a particular model.
This is a regulatory which people nowadays follow, but now all the models you have to get the access and then on the.
So what do you see here? So probably here you will not see you have been granted access to this model. There will be a entry box.
So website is anime.co. That's anime.co.
Where are steps you know how to get your HR token?
First step is create.
Your account.
Then click on your profile, go to the menu.
Here your profile. So if you see you have to click on this profile then you have settings. If not you can directly click on access tokens.
No, don't speak for time.
So first, let's let's see.
Is it for 403?
Because that's a common error that I get from everyone.
Fine. Just click on access token. Left hand side will have other options and here you'll see access token. OK, OK.
Is very good because we don't need any network.
Because we don't need any network to use models, right? So we can take it right there and it will want to time pass. This is lightweight and just use elements.
Yeah.
At least.
Here I'm not for a talk, so maybe I need a.
OK, so you have to click on create your token.
I will say while there is a light button.
Here you can give anywhere.
We can become secrets and then we have to give HF under score token.
Yep.
OK.
I hope for me it fix the session. So you can just have to create a new key. The key will

be HF under score token.
By HF under score token with the environment variable is saved under the name.
For 8 seconds notice on our screen everyone. Let's suppose here the HR program to accept data. Then here you have to give HR on the HR program to here just to match with what is there in their library. We don't confuse ourselves.
Is this best practice, right? But in secrets you can do any name, but just to keep it meaningful or I'll save it with the same name. So it's always good practice to have names.
Similar to what they have been there like to read, right? Again, Facebook, you need to accept.
No, not a new page. It's called a new page hub, then API key. But for a new page it is HF token. But rarely we will do a new page with uh plan change. Should we have better options like Sambanova, Gemini, Drop.
So these are the three ways. First of all we will start with influence line. Amarepa is already open and it will actually call grant access, grant access.
This is the.
In terms of technical job.
Is that the question? Then we have you for testing. I got you don't, and we will create evidence.
Uh, this word. Usually this you will read it from. Only in collab we will use this thing.
So when we we will be using VS code only, right? VS code will not use this like.
We said it would be from environment variable. Here we said right. So these two lengths of code, it will be on the.
OK, so now from a new place.
So if you're using influence, you can use it for free until certain amount of.
Model equals to, yeah.
Follow that model. I mean 7 weeks. I'll point these and if you look at the download, what this number of downloads.
And this is another validation point. You can look at any models, you'll never see. Just make sure you also see the graphs has this number.
For any testing purpose who have so.
OK so you see download from it. There is no graph in between someone has a spike.
If I use it for one testing.
Query here it will be none. That means I myself is 0 right? So here are some most of the models. We will see spike but that means only one spike or one more office that

means.
Internet team has used that model, but in terms of mutual, it's never that case.
You will always have and this number is.
It supports function calling, performing instance, audio visual, not audible.
So if you see it supports extended usually V2 of 7 million 8 million by sorry 8K contacts in. So they increase it to 32K.
OK, but now in the latest model I it's 128 bit, right? But this is still OK, which is still good number because.
Harikna ko sarvat ke bhi hidata.
So whenever you're using any model from a new phase, your model, your data set, copy the entire thing repo along with the model name.
And then that's equals to proper equals to. This is only for safety purposes.
Just make some modeling of the same.
Now what are the keyword audience I need to look for?
temperature. I don't want to be medium, so I'll just be 0.1, but if you're training from the education related, then you can give 0.8, 0.9 and then top peak.
I can keep it 0.95, which is fair enough and then I have Max over.
Nice opens.
Nice opens. I didn't need to open.
But I was going with default. So these two are defaults. These two are change. Max tokens, usually there's 256 or 512, they capital 2000, and for a 30,
So we completed this three lines of code.
Vivo for Van Sackin and this is online only.
So when I gave you the assignment right assignment of the portal to check add not the code base. So during that time I said there are some templates arguments for the writing side function same as that.
Hugging Pace, Fireworks, all are same similar to Openair, Jogi, Openair and after trying different create chat and after that Jogi response format attack Hugging Pace and Openair is same.
So if you want to try to copy what openly added so that it is compatible to all the frameworks, right? So these are the same here. So now what you can do is you can reset any system from.
And then you can define the user account.
Saturday if you want to reuse the same thing or you can just copy it.
Yes.

That's right.
OK, now that we are using open source.
It's better to use the prompt template than the model provides, right? So every single local what you can do is you can use the prompt template of it. So prompt the question.
I was just copied it from which is correct. So if you see S S is nothing but startup token. INSC is nothing but it's for instruction based.
So what model are we using? Mr. Laya, Mr. 17, instead now GPT. What is GPT?
Chat equity, forget about chat equity.
Mhm.
OK.
So what is Sat GPT? Sat GPT is a instance based model. OK, GPT is a base model, right? How did you? How did you define?
Question.
What was the question?
So if you see it.
You also have Google generated AI. Google generated AI is generated base model and you also have chat models. Chat models must have instruct base.
Whenever that conversational skills comes to that model, it is insert base. So there are always two words, right? First you release a base model and then you release the chat model because many people prefer conversational base model. So you use conversational base.
Every single API that we use are conversation. So you see model of but fine tuning is different right? So you have two different findings one is.
Just fine tune the base model and one more is fine tune instant model which is instant based and when you do instant based fine tuting data data it is conversational based and what kind of data can you do?
What prom technique? Huh This is a role.
You have the things set by set and COP related data set. You have instructions, you have input, you have output instructions and then.
Which one's COP related data set or it's for of the.
We can change this model would be. Insert this point. Alpha is nothing but COD.
Right. Since our board is to build a conversational board, we will use instruct based based model. For now directly Mr. AI 7B is. So Mr. AI 7B is a base model. Instruct is a instruct based model.

So failure it makes sense to see about instruct something like that.
So now what we will do is every single time, if you want to design a form, a four template is coming start of the end of the tokens.
SSR and then you have again SV.
First thing is after some wrong there, you'll be funny to come from and you close it.
We don't know my output.
So output in the sense of a BPD level and if you ask a lot it will be some generate thing. So if it is in such a model same way but it will be like how can I help you today like for another question as if you want to continue the conversation.
In Simmba Vojya EPA logo, all the EPAS are.
Well, that's in back to that.
M.
OK.
So he starts with S, then INSV, then he gives in circum close S.
OK, OK. First we close and I think I need to write.
OK.
You started model answer. Model answer is role. System role, user role answer, model response. How are you model response case or we just start with instruction?
We give the
And then you can define.
We will try both the message, take message of both. We will do system from in one way we just pass the base from template of what I'm sorry.
So the of the other one, we define the line, we just have to run it. How did we run it? We have a system form, we have a user form or a role defined. Role system equals to that particular name, then role user equals to that particular name. We will follow that also in the locally example.
But I'm also telling you the second way on how to do it. Most of the. So this is a prompt template, right for open.
And then I'll show you the example. If you have your own system from the user from, how can you convert it without prompt templates? OK, so one example is with prompt template, one is without prompt template. So when you show without prompt template, then tell it right.
Roll the system and then roll it to the for here since this is the template with the different paper. Whereas we will do one of the reason why public.
That is very no concatenate meaning.

Not that system input my system problem. No more. That's what it because I was. Either when it because I wanted to write in multiple. So single quotes, double quotes, triple quotes all are free until you define it in a variable. This is a comment now.
OK, but if you save it in a variable, it's a variable.
So what is keyword arguments? Keyword arguments takes a example of that by passing double as.
OK, that will be.
Stop in a year.
Talk.
So now aware of the format. This is what you call here. So what kind of models open AI like models open AI documentation. Open AI Python.
Now what you are doing is using the API.
So what they are finding is that open search will have more flexibility when you run a type in local. Now what is this?
Open AI. So my AI will sketch that completion. Now once you run the chat completion, it will be variable. You have it that is an object. So if you see the output, I mean in the sense you have the results.
Now what is this thing? Stand completion output. It is a class because the class are the same representation. So now what we need to do is we need to take the final response.
Aggie Festa formatting is similar to Open. If I copy this line, I could get the results.
So inside response.
I have choices.
Inside response I have choice now choice format list.
Message. Yeah, you have no message.
No, did you get direct response? I got the response. One of which didn't direct results. Otherwise I thought about change your prompt. It's like I met the Prime Minister of Narendra Modi and then you got some more results.
So this is something that you have to tweak from. As I said, right, Vishnu is very direct. You know, long only answer to the point. It will never be so.
Increase sample extra value. OK, keep it 0.9 Agarapore created in a SA.
Now that you see the response, it's like actual pilot's angle error you have done. I've been not sure, but I do know the prime minister is a high ranch and profession much like.
IM under score M.

M.
It's for I like. I can understand you. Can you approach it?
OK.
Yeah, OK, this is a total token. This is for a iron.
Yeah, yeah, yeah. So everybody from.
OK, now you're obviously we can.
OK, it is.
OK, OK, OK.
Are you?
So if I'm getting.
Just to see if you're getting the so that I can.
Yeah, yeah.
Oh, something else. So usually you'll have something that goes.
0.1 at this point, 0.98 million.
So now what you have to do is XML and then it was misunderstood the XML format, that's all.
Format. My mark on format works good with open source. So so as I said right open source cellular you have to do experimentation.
Sometimes our poly girls have some leg up, but the only thing is Mistral and Deepshika are good. So now create a new somebody. Somebody create a Nigamas just like Lukkal.
Oh yeah, this is.
Now just write local.
So far we were running it on CPU. CPU, it's not possible, right? Because we want to load the very based model, so we'll click on runtime.
Then change runtime right and then select I will select the model.
It V2 dash TP watcher better than it GPU. It's one OK because every time we change runtime.
Runtime disconnect over. OK, so you have to for now it's fine, but the next time when you do change runtime.
Now when you do restart, not restart, restart, but if you change runtime, all the things will be gone. It's like you're running Colab for first time once you restart the runtime.
The model we will use the same but quantize currently right? For quantize we will use bits and bytes.

Yeah.
In that sense, you have collab one and you have collab two. That's not possible. We can run it only from one collab at one time.
Separate, separate. So both are bits and bytes. So what bits and bytes will do is.
So.
Floating point. Let's suppose I have thirty-two bit floating point, so total I'll have.
So just like that we have 32 bits. So can anyone tell me a bit here? Fine, thank you.
But it's called exponent. It's my data.
So usually... So all the models do we have this fellow by recorded in 91, 32-bit floating point.
A lot of framework. You can bring that down to 16 grid.
So this is the normal format format. If you want to run any model minimum.
And now if you confine, just for inference, with the model of load, loading of model only once. If you can ask as much of questions to read.
If I click on this, I'm going to GPU RAM for GPU, right? So.
Yeah.
13 GB is only taking to load the model of the passenger 15 GB.
It's not possible, right? Even though you can run one query, there might be chances it will be limited, right? So now what we will do is go be a 16 bit floating point model.
We will run it in four.
4 bit in 8 sorry 4 bit NF4 OK formatted NF4 is nothing but.
Normal floor 4 bed.
That of four bit and a format and that of 8 bit in 8 format. So what is 8 bit now? 8 bit is the floating representation and it is in integer 8 bit.
So once you start working with fine tuning, after model of and once you load your model, there are two ways you can use fine tuning.
But I'll just give one example. These are people use.
They fine tune it and they make it real estate and then more over the people can use it. Now what they did was they made it real estate. Let's suppose you build text to image.
And what one guy did was it built a LoRa adapters where you can convert that text to image only for anime characters. OK, if anime character, you can't do it. So if you want to use a base modeler, then someone has to train that model on anime data set.
And then that Lora is distributed across. You can reuse that so that Lora adapters is

only one of.
One quantification or you can just say single quantification.
Oh, Laura, this is Laura, this is.
More than.
Now, there is one more concept called Pulora. Pulora is nothing but quantized pluora.
So usually quantization, there is a parameter called quantization, quantization factor.
So let's suppose you have.
Kalpan Li is a Bhargavi Bharam.
OK.
Yeah, so.
Hello Andreka basically 2627 eight. So do you know this range? So for 16 bit your range is from -127.
To 127. So now what we are trying to do is when you are doing quantization, you need the range, you need the quantization factor and then up you do the values and down my exact formula.
But that's not.
But won't take it. It won't give you workload performance figure. So I'm going to you're running it on 4 bit and in recent days there is a library part on slot.
And there is a research for 1.58 bit quantization. You're running it in just one bit. The performance of opposite finds at 6 fours giga for 32 bit to 16, one bit.
Most of the people are.
There is no performance drop. Hardly scores which is benchmarking. Perplexity is very good evaluation of this, not the startup.
So that is a huge thing for a very long time perplexity.
One bit, but which is negligible. Perplexity is a metrics, perplexity laws.
And you'll get your quantization factor formula for signal and you'll get the slides.
Chesa Yogana Di Vaibhavi.
Mhm.
It's there where Sriland is. So because obviously talk about, I mean.
So this is how your vector look like or is the maximum cap.
5.4 this is your 5.4 or range here. So you just divide your range by Max you get a quantization factor. Now once you get this quantization factor up multiply. So it is in 16 bit.
Once you multiply, this is your you do it again. No, this is the quantities. So this is

nothing and the formula is same range like a.
And then you just multiply it. So no one can not do that with the. So this is 127 of 8 bit, 8 bit per XKR, 127 to 127, minus 63 to 16 bit.
So you would change over. And now what you are doing is 63 / 5.4. OK, whatever value you get was multiplied. OK, for whole multiplied value. So now these values will increase the above.
64 So when you're using 4 bit floating point data values, it will be one, but it is compressed.
So this is mainly used for two Laura. Laura, one quantization of Laura, two Laura have a double quantization. There will be 16 bit there. First you bring that to 8 bit and eight bit like you are going to 4 bit which is called double quantization.
And this is NF 4 for.
So now what we will do now is we directly have a model which is 600 bits. Using bits and bytes, we are directly moving that to 4 bit close part. And then we will influence the model.
M.
And if you are working with error.
Don't run, Thomas.
Import and I told you this. If you're using initial model, what I will do is they have auto classes, right? So to three models, I'll pick the tokenization for that.
So you pick the model and use auto crashes. Auto crashes can give any model and you can give any and ugly phase will do that everything on their end.
So auto model for CASLM is where we will define our mixture. So there are only two types in again phase models auto model for masking LLM. Masking LLM doesn't for BERT Roberta P5.
You might have seen one special argument, pad, UNK, CLS, SEP and there was one more called mask. Mask is mainly for BERT and everything. So here we are using casual element.
LM is auto regressive models. If you want to use that, you can use Kaiser LM. LM is nothing but language model.
And then you have auto tokenizer, auto tokenizer, whichever model you pick for model, same tokenization will be used for now.
Yeah, we use it in a lot of form for training process.
So we have to rethink if you want to create. No, they're just quantitative. So there are no different, there are a lot of different. We just have to define the quantitative. OK,

no, I mean like team has generated.
Can I on is it or do I have to?
You can. So once you leave your adapters, it's not a model adapter. Let me combine it with your base model to make it fine-tuned. OK, so when you combine it, you can add.
Yeah, I thought that didn't do that. Once adapter is built, it will be.
OK, so the GPM is here. So now what you can do is define device as.
who write a programming language which is GP programming language.
OK, so this is basically a GPU programming language that they have GPUs use CUDA and the developers of CUDA is Nvidia.
Yeah.
So nowadays people are also using Krytan. I'm not gone in that, but Huda is 1 or fix ratings are there is Krytan.
So now now they're getting cheaper and the and now they support program.
GPU the best. If we get the supporting GPU, then probably the other things. Now if they're migrating, there might be items that shift to.
But Cretan is very complex. Still, the developers of Cretan are open AI and and media.
There's still be any Nvidia product at the end of the year. Nvidia.
TMC teens a bout you.
OK, uh, either that something. Now what we're doing is we want to load the model in four bit. So load in four bit is true. And what kind of four bit is it? What kind of quantization?
16 bit man, you're bringing it to 8 bit with a single quant. Double quant is 4 bit. So B&B is nothing but bits and bytes. 4 bit use double quant is true. So these are attributes of the particular class, right?
And then BNB 4 bit quant type. What is the type?
And a fault. So a quota of float 16 which is 6 bit bit floating point, 32 bit token. Then you have int 8. Int 8 is nothing but integer unsigned 8. And there is also OK, there are two one is.
One is int8 and then there is UN8 which is unsigned integer 8 and then you have NL4 which is normal for one bit of because is 1 bit.
But this is the recent 1.58 quantity.
And pi 8 is there. It's not just one bit, 1.58 works on 0 and then you have BNB 4 bit quantum. This is compute. Compute data type is stored system.

OK, I didn't. Sorry.
I will.
What did you run? Oh, once we load the model, everyone are done. Yeah, yes. OK.
So now what we'll do is we'll define the model.
So you have auto model for caselim dot from retrain. So there are two functions here. Both the functions are important. Let's suppose if you're running it on VM right redefined your only from.
One time. OK, yeah, from three different saver, you can define a path where you can save all the model threads. OK, next time when you are loading the model again, you can use from save.
You can use it from concrete. So now what are we doing from pre-defined? Pre-defined, same modeling, Device Mapsair, Device Mapsair now running it on CUDA. Either you can do PUDA, then you can keep auto. So based on your system requirement, it will pick if PUDA is available, it is PUDA. If PUDA is not available, it is CPA.
So how do we get that? It's just two lines of code. Import Torch. Torch is the number one library. It is Torch Pytorch every single LLM.
Is built using this library. If this library doesn't exist, everyone has to write Max and build neural networks, build CNN. They build neural networks for.
Pulsar.
So will be Neural Networks, CNN. The one who proposed CNN was so this team is there and Facebook was the one who built the the best open source framework is it.
So we can get, but what in fact, that is probably very.
So we'll use this for. Thank you.
I don't know how much Kalpeshar.
92.72.
Very close to on that place.
He is available.
True. So if it is true, it will be CUDA. So it will directly write CUDA. If it is false, it will take C3. So since you already defined CUDA by this.
Add device and now Amit, we want to load this in fourth bit, so we're using quantization concrete, BNB concrete.
So do we load that? OK, but when you do it, it's not converting. You cannot convert and save.
No, no. Even the loading that. So whatever that calculation is on, we'll load it and tell

you. OK, you are getting error because you need license.
All these models are now licensed. Only when you grant access, you can use it. So you just run this again and just make sure the line opens the door because model act is there.
Typically you don't need an API key, you need only an API key to bypass this. It is just checking whether the accounts are no API, whether that account has the access or not. Can you see this slide?
Does anyone know this? Any of this?
I think granted access to this model. You have to click on agree and access the positive. Now we have access.
We can come back here and run this cell, and now rerun this.
That is only permission set. It's not.
Bhagavpuri Sawan, Vijay Kumar.
No, I think about it.
So now these are safe tensors. There are two types. Either of safe tensors you'll see model weights, right? So if you see safe tensors, that means all these files are weights.
That means you build a model and that weight is available to public. So everything neural network is like you have weight and other weights are available and you can reuse it.
So this is what you call this open weight. That means you can reuse that model. DPT gemini are not open, weights are not available. Other weights are available there.
Then you can fine tune it for three. Three doesn't give you a yoga, but once you have GPU, you can fine tune it.
18 download on Akbar, your model is loaded.
So this will take time.
Because I can write the.
After model, we have to defend to organizer.
Until that is running, we can just define this.
For infant it will again take time 3 seconds, 3 seconds.
Tokenizer again same syntax from predefined or the name it should be same. Now can anyone tell me before running this?
What is it?
Tokenizer, tokenizer, tokenizer config dot Jasonicoga and then there is one model.
Yeah, yeah. And then you have special tokens. When we trained, this one, this one

and this one.
Is this here? Hmm. We have to use specific to that model.
And what we did was we just built BP. We didn't use the exact. So that's the reason why we didn't have model.
We only had the vocab complete, then tokenizer and what is the mapping of the fresher document?
So that after the port, this should take one second. We won't need. Thank you. I will take one second.
OK, now this is for who is using keyboard.
Most of the time extra extra RAM you so we might have already seen this regular one.
So once you have any remove any digit variable, you just report DC and do DC dot correct.
So this is 1 and then the only import talks here. There are some empty space in Ura, so you have to empty the space.
Once the model is built up, it will be. It will not be much, but something it will be.
Now what you can do is you can write it in the format. Different role for system, different role for user.
Yeah, we will not use the the compensate. There is a function to bypass.
This is what we already saw here.
But the only thing is it's in dictionary rather than a tupple.
And there's some documentation, because every frameworks have their own way to define it. In a line chain, open AI, if you're using direct elements, what is Line Chain and Laman? They have frameworks around open AI.
It will be.
Input variables, because input variables is a template, though.
So what what is 1919 wrapper. Data you are loading, you have data loaders then prompt.
So it's like templator. We just want to use the template and you have to run it. It's a wrapper thing.
So now if you see here, you have a list dictionary. This is a common template of every element, but you have change over if there is any framework.
We still have four documentations, but this is standard practice.
But you'll see this one later.
This probably no one has any doubt, right? Yeah.

Whatever variable you're looking at, you can do that. So let's suppose you have user from.
We like, we like, I'd say, and here you like. So this issue is already defined here.
Yeah, I'll either you or not. Yeah, I'll so this will.
Yeah, no checkpoints.
But this is only one time. But this is for VM at this point.
So I'm just copying my system and then running the message. Yeah, one is collab and one is scattered.
So model whatever model same and these files you already know what this file contains. I just want to see some space.
And the format instead of tuple. This is a standard time.
We have everyone are done.
Not running, but so once it is executed.
So there are two formats we will use. One is pipeline, but you just use the pipeline signal also parameters.
The second hardware is no. There I want to show you generated IDs. So tokenizer is only IDs like the reason why we use IDs is because we have to pass it in the morning. You can write this. You can also avoid it because they use. This is just for understanding.
So when I every time when we define any open source, we have to use a prompt template. So what has done is in tokenizer they created a function called apply chat template.
So whatever prompt template any programming languages has, they will map it in their background. So your goal is only to be and they will take care of the chat template and.
And you can add your message.
So what are we doing? We define our prompt and we want that prompt to be in a prompt template that a open source model can understand, right? Now The thing is still here you understood and then you have return tensor is equal to CP.
Tensor is nothing but metrics, vectors. Yogi a frameworks and as well as Tensorflow vectors for they call it as tensors. And what is this PT? PT is nothing but pytots. Other of Tensorflow, instead of PT it is PA.
Is it clear?
So PD, is this clear? OK.
And now model inputs.

It is in CPU once it converts, so we just want to make it in CPU.
And also.
When could we be listening to the list of all the IDs?
OK.
Yeah.
OK, this is written pencils.
And now your arguments with this popular click we get list of on the but now it is on GPU compatible. OK so N codes is same N codes and model input but the only thing is everything.
Pytorch format sensors and applicable compatible for GP.
And now let's define our keyword ordinance. And every time we should also tell the padding token. The pad token. So pad token ID will be my oh that was the tokenizer EOS.
This is constant, OK.
So for fine tuning also, every time you define the total angle, do you want to start it from the deep?
So we just have to define that.
Temperature. I'm keeping 0.9.
So we use encoding and now we have our QR arguments. The next step is we need to decode.
Yeah, yeah.
And now in code on I'll just write generated IDs equal to model.
Instead of invoke, either they have a method called generate.
So here model generate is taking a decode. Yeah, model model inputs. It's not a normal prompt because of course we can see how the data is passed in the model when you're getting the response.
This is how typically the response is going to model. If you ask who is that is only for user, but for a model the input is.
It's only IDs. It's only then it is based on the vocab correct. But now when you use pipeline and if you use Modelingo, so you will never know IDs using.
So now you have tokenizers. Do you know what is embedding? Embedding for input IDs. You also know of this going to model though encoder. Once you have the model, the decoder, the decoder will give you the output.
So number set. You never have any text in deep learning. It's always numbers. So other than pipeline, this could have missed, right? Because you never know output

there. So I will do model or generate model IDs.
One keyword all the attention masking the fact open ID were not said.
OK, now that we are using the model, model for maximum token 3, it's maximum token. Before we were using OpenAI compatible. Now we are using model for parameters. So you can also have top gear.
So now if I print the generated IDs, I'll.
And now if I print generated ID, it's still numbers.
OK, OK. When you call the model, because it is just to show since you are there, may I repeat Gautam?
When we discuss tokenizers, every tokenizer digit, right? Your digit is what is passing in your model.
Or model B takes care of, which is wrong. Basically your input is converted into IDs.
Deploy is same. So can you take this example for me?
Raw data or raw data. You have token ID. Token ID is now going to embedding so that you can find semantic means. Model is only getting your input IDs. If you see you have input IDs, output is also IDs.
So when you use any function pipeline, you will bind site, you'll get text as output, but you will actually not know what is happening behind these things. So behind the first encoding over with our.
Then you use model generate encode, decode, decode. You give your input which is messages and then you get output also in IDs.
And now we just have to record. So I just do import them because I don't need.
I don't need the input ID to be so input ID.
So shape one. So if you see you have 2D array. You have one method which is shape of one. Now if I do input length.
I'm getting 31, whatever I'm asking is 30 minutes. I don't want this 31 program to be repeated.
So now if I do new.
Uh, because you're generating Uh.
Generated IDs. What is generated IDs? Then this is your output. Then you just have to get the 0 index.
So you just have to decode. Even if you don't use new tokens, you can directly pass generated ID 0. So let me show that one first.
Decoded.
We aligned once before. Don't admit. I'll show you the actual result first. You already

got the response. What is the decoded ID? Either you just have to use tokenizer dot. Here it's decode.
Because autocomplete will always show. So here I'll just show generate ID.
0 and skip pressure token.
Huh. If you use that report.
But batch record also, so I'm just avoiding that too.
And I want to tokens and now if I print decode I'll have my active output.
So what will I do? Are you able to understand what I'm saying here? So generated IDs, give me the output program, but it has system problem.
I want to avoid model IDs. So Vijay Karuna, model ID is already here.
For now, my new for consists model IDs, generated 0. This is my output.
OK, so I just do index.
New tokens.
Now you have your actual response. So here is he done. May I see if I.
Now what you need to do is go up the model encoding and decoding. This is only for you for understanding purpose. You can directly use something called as pipeline.
Go transformers from transformers to import pipeline.
And pipeline where you have to do a generation model equals to model tokenizer equal to tokenizer device. So when you run this, why? Because when.
But if you don't use CUDA, like suddenly if you don't use auto and if you directly use CUDA, it is also safe.
So is this clear? We are not encoding, we are not decoding. We are just getting under that. So we are directly defining Python.
Yeah, we integrate. So now response is going to under pipe is here pipeline. What we can do is we can pass our message.
And.
Mhm.
OK, so this is the response. When I pipeline define pipe by variable.
Generated 0 is a dictionary where generated text you already have a output.
So now I just tell the -1 -, 1 it will take the last the assistant test. We require system user assistant. These are the 0 tested on Saturdays and we just have to print content.
OK, not.
I hope. Now X input and X output. There is no ID, there is no output.
It'll go second time. Sorry, not a second. It does come.
Bharat Bhar.

Oh.
So the one not a one is how to use inference, right? It was IDs and then pipeline which is recommended. This is recommended because you know you know the input is not a string, right? It's ID.
No, that was just for understanding, encoding, decoding. Because I just wanted to show complete. We came to this flow embedding over here. Since we are not building models for tracks, I wanted to show how this actually works.
Now you have your model.
Hello.

◉ **Mitesh Rathod** stopped transcription