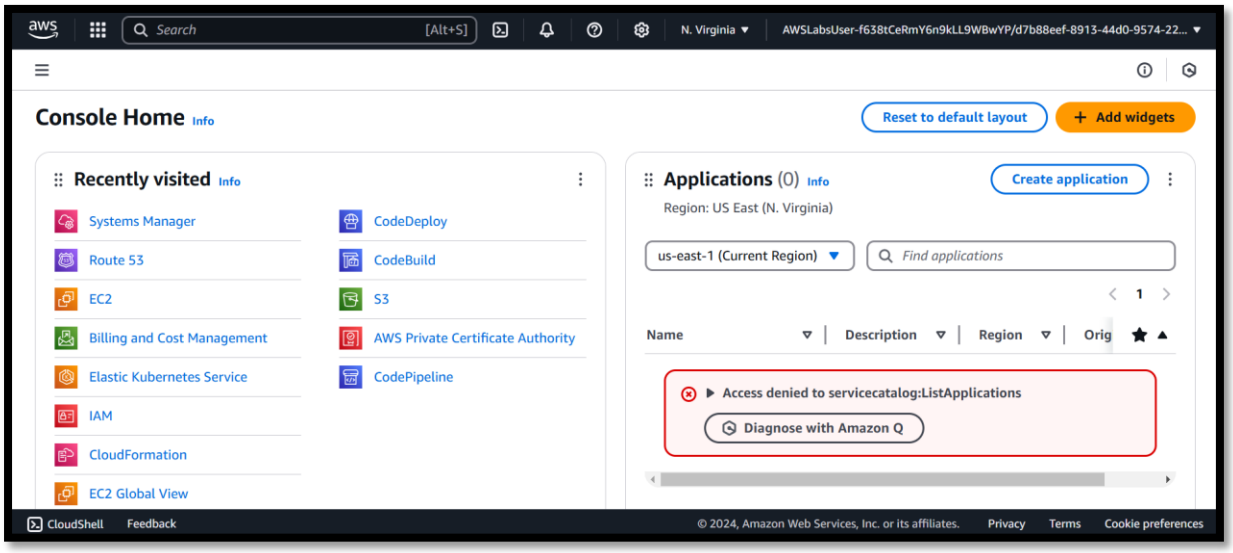
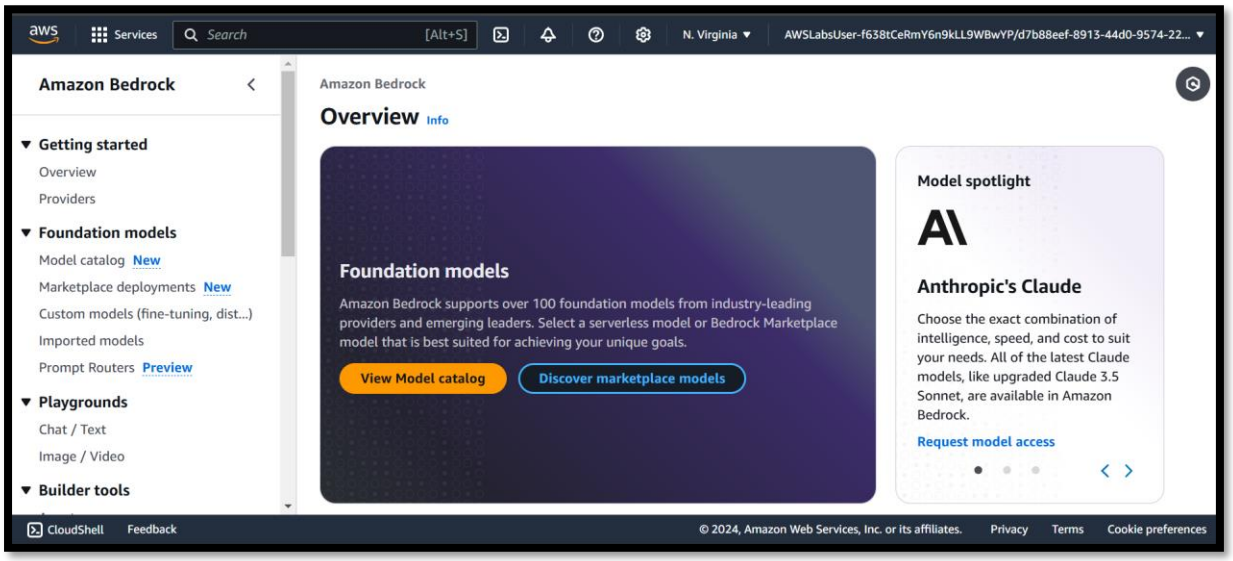
Objective: To use the Bedrock Titan model for question answering by providing factual responses to queries through context-included requests and receiving relevant responses.

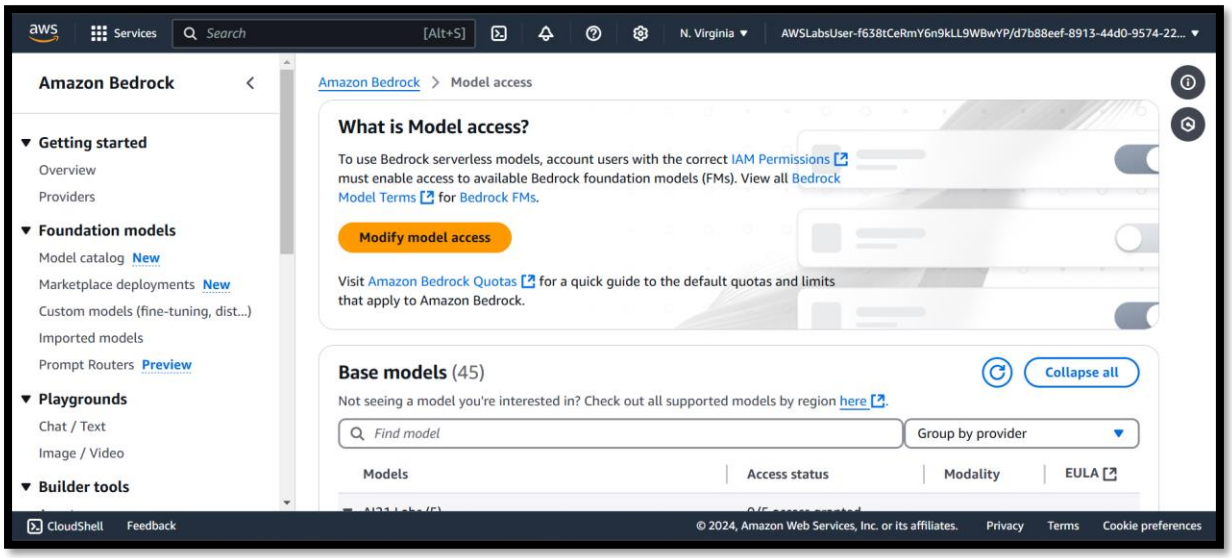
Task 0: Set up the environment

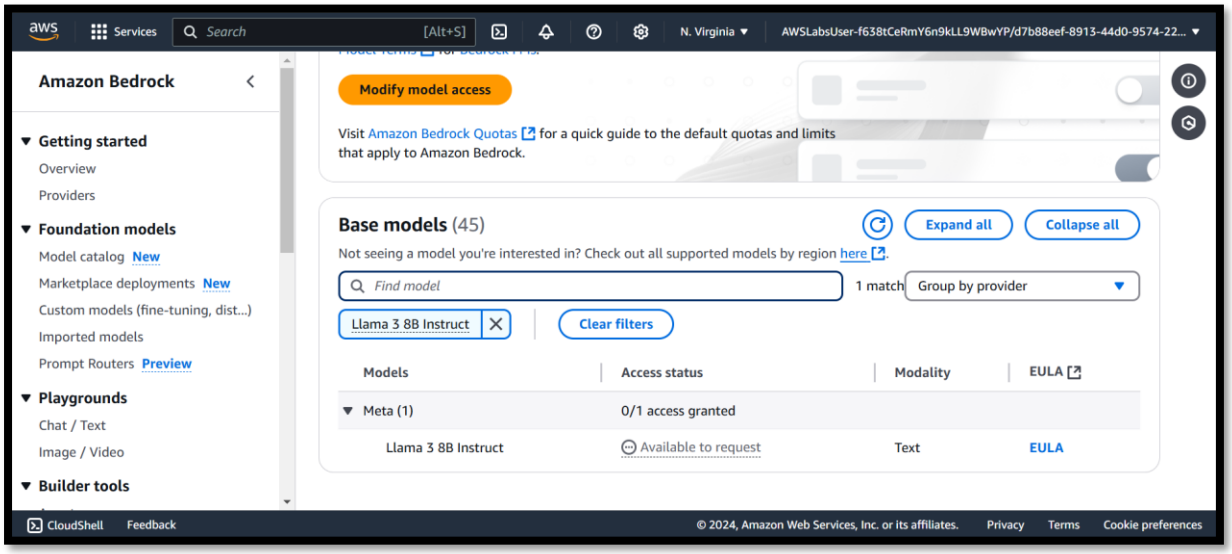
In this task, I registered the base models in the Amazon Bedrock console and launched an Amazon SageMaker Studio application to access my lab resources.



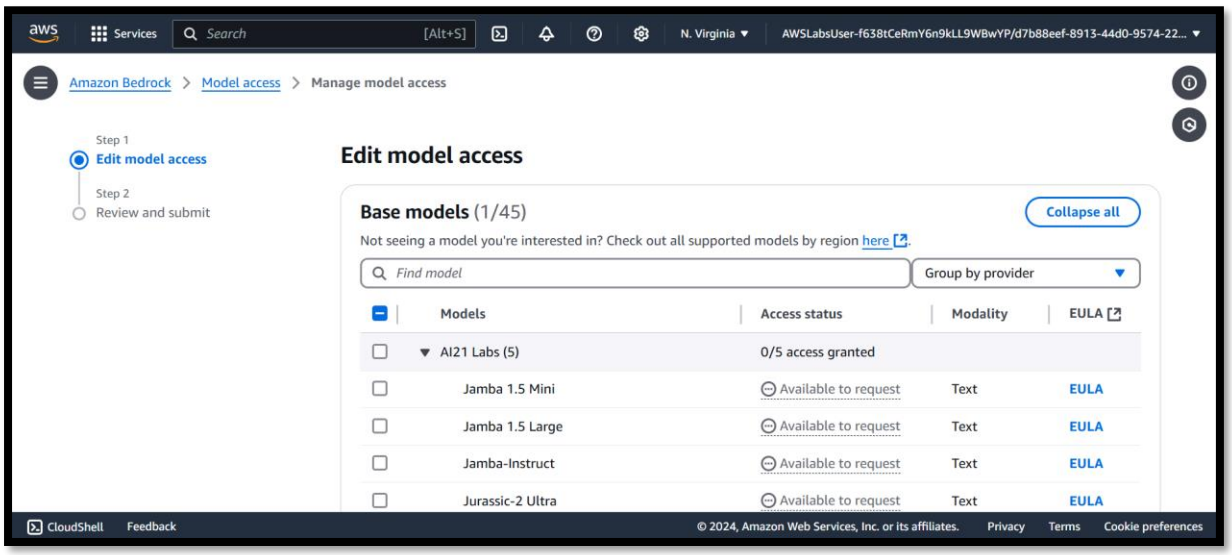


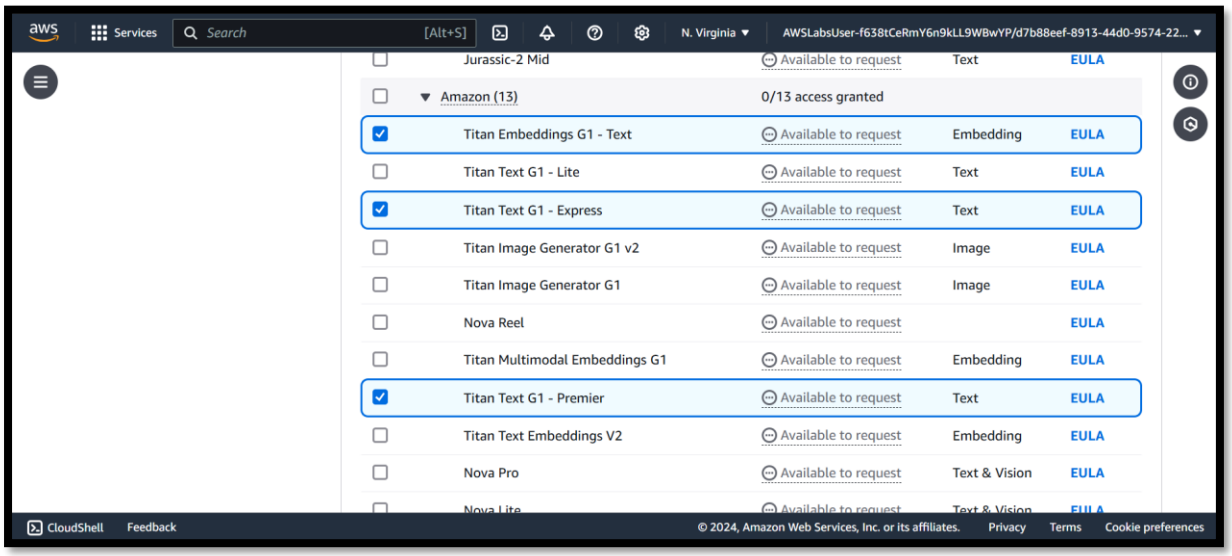
I reviewed the Access Status for each of the models. If the Access Status for one or more of the models was set to Available to request, I expanded this menu and followed the steps to enable access for them.

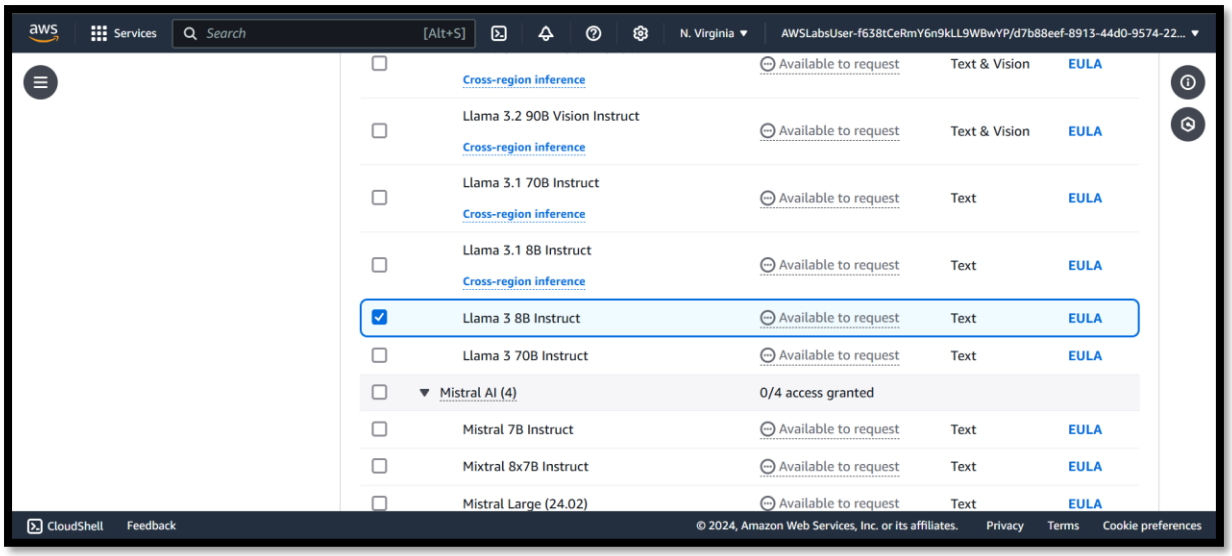


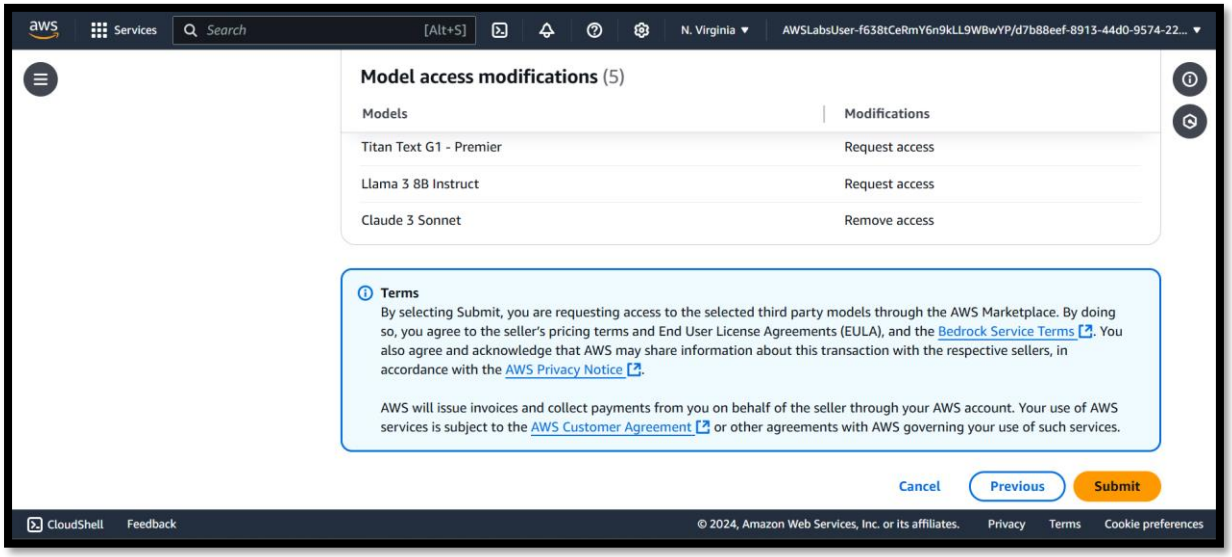


I chose Modify model access at the top of the screen.

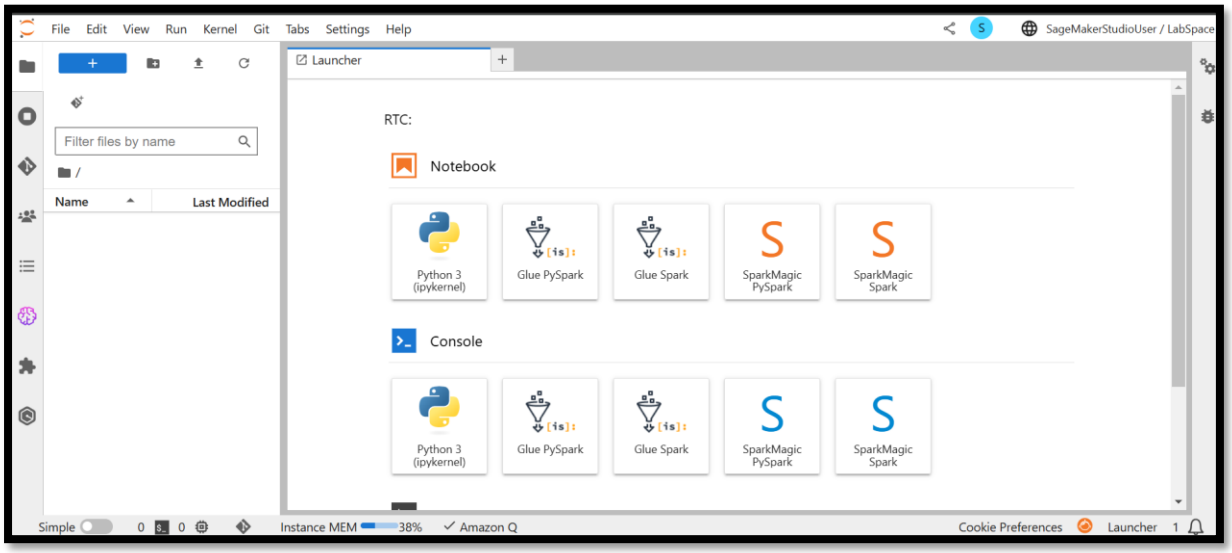





aws
Services
Search
[Alt+S]
N. Virginia
AWSLabsUser-f638tCeRmY6n9kLL9WBwYP/d7b88eef-8913-44d0-9574-22...
Cross-region inference
Available to request
Text & Vision
EULA
Llama 3.2 90B Vision Instruct
Cross-region inference
Available to request
Text & Vision
EULA
Llama 3.1 70B Instruct
Cross-region inference
Available to request
Text
EULA
Llama 3.1 8B Instruct
Cross-region inference
Available to request
Text
EULA
Llama 3 8B Instruct
Available to request
Text
EULA
Llama 3 70B Instruct
Available to request
Text
EULA
Mistral AI (4)
0/4 access granted
Mistral 7B Instruct
Available to request
Text
EULA
Mixtral 8x7B Instruct
Available to request
Text
EULA
Mistral Large (24.02)
Available to request
Text
EULA
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

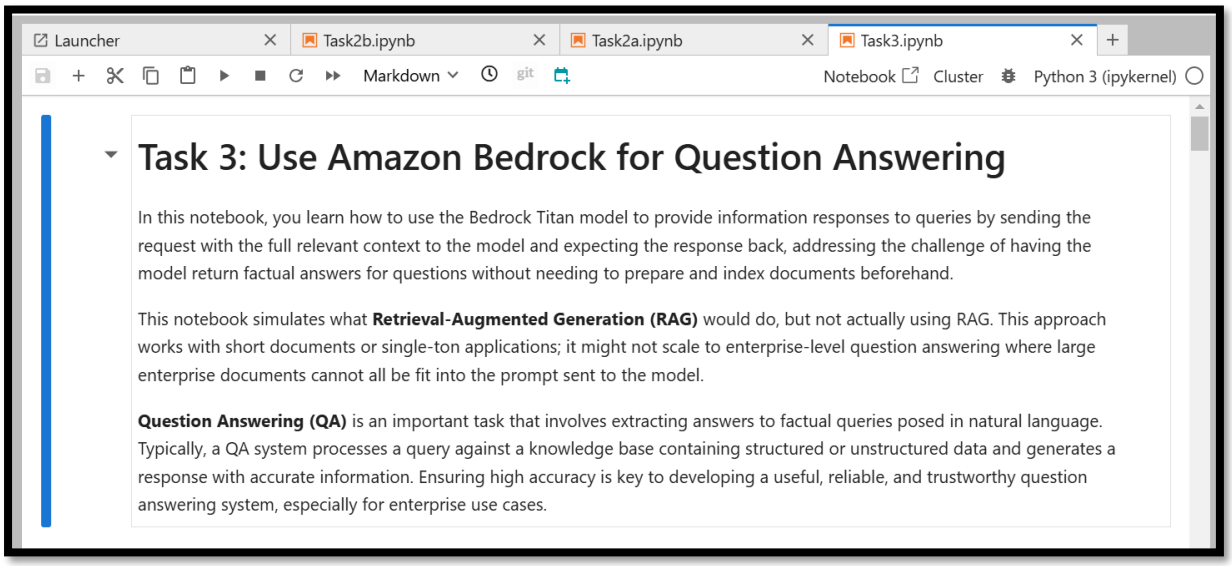

aws
Services
Search
[Alt+S]
N. Virginia
AWSLabsUser-f638tCeRmY6n9kLL9WBwYP/d7b88eef-8913-44d0-9574-22...
Model access modifications (5)
Models
Modifications
Titan Text G1 - Premier
Request access
Llama 3 8B Instruct
Request access
Claude 3 Sonnet
Remove access
Terms
By selecting Submit, you are requesting access to the selected third party models through the AWS Marketplace. By doing so, you agree to the seller's pricing terms and End User License Agreements (EULA), and the Bedrock Service Terms. You also agree and acknowledge that AWS may share information about this transaction with the respective sellers, in accordance with the AWS Privacy Notice.
AWS will issue invoices and collect payments from you on behalf of the seller through your AWS account. Your use of AWS services is subject to the AWS Customer Agreement or other agreements with AWS governing your use of such services.
Cancel
Previous
Submit
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Launch an Amazon SageMaker Studio application



Task 1: Use Amazon Bedrock for Question Answering

In this task, I utilized the Bedrock Titan model to provide factual responses to queries by sending context-included requests and receiving relevant responses.



Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | Notebook | Cluster | Python 3 (ipykernel)

Task 3.1: Environment setup

In this task, you set up your environment.

[1]:
```
#ignore warnings and create a service client by name using the default session.
import json
import os
import sys
import warnings

import boto3
import botocore

warnings.filterwarnings('ignore')
module_path = "../"
sys.path.append(os.path.abspath(module_path))
bedrock_client = boto3.client('bedrock-runtime',region_name=os.environ.get("AWS_DEFAULT_REGION", None))
```

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | Notebook | Cluster | Python 3 (ipykernel)

Task 3.2: Q&A with the knowledge of the model

In this section we try to use a model provided by Bedrock service to answer questions based on the knowledge it gained during the training phase.

In this task, you use the invoke_model() method of the Amazon Bedrock client. The mandatory parameters required to use this method are modelId, which represents the Amazon Bedrock model ARN, and body, which is the prompt for your task.

The body prompt changes depending on the foundation model provider you select. You walk through this in detail below.

```
{
   modelId= model_id,
   contentType= "application/json",
   accept= "application/json",
   body=body
}
```

You try to use models provided by the Bedrock service to answer questions based on the knowledge gained during the training phase.

[2]:
```
prompt_data = """You are an helpful assistant. Answer questions in a concise way. If you are unsure about the
answer say 'I am unsure'

Question: How can I fix a flat tire on my AnyCompany AC8?
Answer:"""
parameters = {
    "maxTokenCount":512,
    "stopSequences":[],
    "temperature":0,
    "topP":0.9
    }
```

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | Notebook | Cluster | Python 3 (ipykernel)

Task 3.3: Invoke the model by passing the JSON body to generate the response

[3]:
```
#model configuration
body = json.dumps({"inputText": prompt_data, "textGenerationConfig": parameters})
modelId = "amazon.titan-text-express-v1"  # change this to use a different version from the model provider
accept = "application/json"
contentType = "application/json"
try:

    response = bedrock_client.invoke_model(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )
    response_body = json.loads(response.get("body").read())
    answer = response_body.get("results")[0].get("outputText")
    print(answer.strip())

except botocore.exceptions.ClientError as error:
    if  error.response['Error']['Code'] == 'AccessDeniedException':
        print(f"\x1b[41m{error.response['Error']['Message']}\
        \nTo troubeshoot this issue please refer to the following resources.\
         \nhttps://docs.aws.amazon.com/IAM/latest/UserGuide/troubleshoot_access-denied.html\
```

```
         \nhttps://docs.aws.amazon.com/bedrock/latest/userguide/security-iam.html\x1b[0m\n")
        class StopExecution(ValueError):
            def _render_traceback_(self):
                pass
        raise StopExecution
    else:
        raise error
```

```
1. Find a safe place to park your car.
2. Turn on your hazard lights.
3. Remove the hubcap or wheel cover.
4. Loosen the lug nuts on the flat tire.
5. Use a jack to lift the car until the flat tire is off the ground.
6. Remove the lug nuts and the flat tire.
7. Install the spare tire.
8. Lower the car back to the ground.
9. Tighten the lug nuts on the spare tire.
10. Lower the car back to the ground.
11. Remove the jack and tools.
12. Put the hubcap or wheel cover back on.
13. Check the tire pressure.
14. Drive to a tire shop to have the flat tire repaired or replaced.
```

The model gives you an answer outlining the process of changing the car's flat tire, but the same explanation could be valid for any car. Unfortunately, this is not the right answer for an AnyCompany AC8, which does not have a spare tire. This occurs because the model has been trained on data containing instructions about changing tires on cars.

Another example of this issue can be seen by trying to ask the same question for a completely fake car brand and model, say an Amazon Tirana.

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | Notebook | Cluster | Python 3 (ipykernel)

[4]:
```python
prompt_data = "How can I fix a flat tire on my Amazon Tirana?"
body = json.dumps({"inputText": prompt_data,
                   "textGenerationConfig": parameters})
modelId = "amazon.titan-text-express-v1"  # change this to use a different version from the model provider
accept = "application/json"
contentType = "application/json"

response = bedrock_client.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
answer = response_body.get("results")[0].get("outputText")
print(answer.strip())
```

```
1. When you have a flat tire, you should move your car to a safe place.
2. You should put the parking brake on and turn on the hazard lights.
3. You should remove the hubcap or wheel cover.
4. You should loosen the lug nuts with a lug wrench.
5. You should lift the vehicle with a jack.
6. You should remove the lug nuts and the flat tire.
7. You should install the spare tire.
8. You should tighten the lug nuts with a lug wrench.
9. You should lower the vehicle with a jack.
10. You should tighten the lug nuts with a lug wrench.
11. You should remove the jack and the spare tire.
12. You should put the hubcap or wheel cover back on.
13. You should lower the vehicle to the ground.
14. You should tighten the lug nuts with a lug wrench.
15. You should check the tire pressure.
16. You should put the lug nuts back in the trunk.
17. You should drive to a tire shop to have the flat tire repaired or replaced.
```

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | Notebook | Cluster | Python 3 (ipykernel)

Given the prompt question, the model is unable to provide a realistic answer.

To fix this issue and have the model provide answers based on the specific instructions valid for your car model, you can augment the model's knowledge on-the-fly by providing an additional knowledge base as part of the prompt.

Let's see how you can use this to improve your application.

Assume the following is an excerpt from the manual of the AnyCompany AC8 (in reality, it is not the real manual, but treat it as such). This document is also conveniently short enough to fit entirely in the Titan Large context window.

```
Tires and Tire Pressure:

Tires are made of black rubber and are mounted on the wheels of your vehicle. They provide the
necessary grip for driving, cornering, and braking. Two important factors to consider are tire
pressure and tire wear, as they can affect the performance and handling of your car.

Where to Find Recommended Tire Pressure:

You can find the recommended tire pressure specifications on the inflation label located on the
driver's side B-pillar of your vehicle. Alternatively, you can refer to your vehicle's manual for
this information. The recommended tire pressure may vary depending on the speed and the number of
occupants or maximum load in the vehicle.
```

```
Reinflating the Tires:

When checking tire pressure, it is important to do so when the tires are cold. This means
allowing the vehicle to sit for at least three hours to ensure the tires are at the same
temperature as the ambient temperature.

To reinflate the tires:

    Check the recommended tire pressure for your vehicle.
    Follow the instructions provided on the air pump and inflate the tire(s) to the correct
pressure.
    In the center display of your vehicle, open the "Car status" app.
    Navigate to the "Tire pressure" tab.
    Press the "Calibrate pressure" option and confirm the action.
    Drive the car for a few minutes at a speed above 30 km/h to calibrate the tire pressure.

Note: In some cases, it may be necessary to drive for more than 15 minutes to clear any warning
symbols or messages related to tire pressure. If the warnings persist, allow the tires to cool
down and repeat the above steps.
```

```
Flat Tire:

If you encounter a flat tire while driving, you can temporarily seal the puncture and reinflate
the tire using a tire mobility kit. This kit is typically stored under the lining of the luggage
area in your vehicle.

Instructions for Using the Tire Mobility Kit:

    Open the tailgate or trunk of your vehicle.
    Lift up the lining of the luggage area to access the tire mobility kit.
    Follow the instructions provided with the tire mobility kit to seal the puncture in the tire.
    After using the kit, make sure to securely put it back in its original location.
    Contact Rivesla or an appropriate service for assistance with disposing of and replacing the
used sealant bottle.

Please note that the tire mobility kit is a temporary solution and is designed to allow you to
drive for a maximum of 10 minutes or 8 km (whichever comes first) at a maximum speed of 80 km/h.
It is advisable to replace the punctured tire or have it repaired by a professional as soon as
possible.
```

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | Notebook | Cluster | Python 3 (ipykernel)

[5]:
```
context = """Tires and tire pressure:

Tires are made of black rubber and are mounted on the wheels of your vehicle. They provide the necessary grip for dr

Where to find recommended tire pressure:

You can find the recommended tire pressure specifications on the inflation label located on the driver's side B-pil

Reinflating the tires:

When checking tire pressure, it is important to do so when the tires are cold. This means allowing the vehicle to s

To reinflate the tires:

    Check the recommended tire pressure for your vehicle.
    Follow the instructions provided on the air pump and inflate the tire(s) to the correct pressure.
    In the center display of your vehicle, open the "Car status" app.
    Navigate to the "Tire pressure" tab.
    Press the "Calibrate pressure" option and confirm the action.
    Drive the car for a few minutes at a speed above 30 km/h to calibrate the tire pressure.

Note: In some cases, it may be necessary to drive for more than 15 minutes to clear any warning symbols or messages
```

```
Flat Tire:

If you encounter a flat tire while driving, you can temporarily seal the puncture and reinflate the tire using a ti

Instructions for using the tire mobility kit:

    Open the tailgate or trunk of your vehicle.
    Lift up the lining of the luggage area to access the tire mobility kit.
    Follow the instructions provided with the tire mobility kit to seal the puncture in the tire.
    After using the kit, make sure to securely put it back in its original location.
    Contact AnyCompany or an appropriate service for assistance with disposing of and replacing the used sealant bo

Please note that the tire mobility kit is a temporary solution and is designed to allow you to drive for a maximum
```

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | git | Notebook | Cluster | Python 3 (ipykernel)

[6]:
```
question = "How can I fix a flat tire on my AnyCompany AC8?"
prompt_data = f"""Answer the question based only on the information provided between ## and give step by step guide
#
{context}
#

Question: {question}
Answer:"""
```

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | git | Notebook | Cluster | Python 3 (ipykernel)

Task 3.4: Invoke the model via boto3 to generate the response

[7]:
```
body = json.dumps({"inputText": prompt_data, "textGenerationConfig": parameters})
modelId = "amazon.titan-text-express-v1"  # change this to use a different version from the model provider
accept = "application/json"
contentType = "application/json"

response = bedrock_client.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
answer = response_body.get("results")[0].get("outputText")
print(answer.strip())
```

```
To fix a flat tire on your AnyCompany AC8, you can use the tire mobility kit provided in the luggage area. Follow t
hese steps:

1. Open the tailgate or trunk of your vehicle.
2. Lift up the lining of the luggage area to access the tire mobility kit.
3. Follow the instructions provided with the tire mobility kit to seal the puncture in the tire.
4. After using the kit, make sure to securely put it back in its original location.
5. Contact AnyCompany or an appropriate service for assistance with disposing of and replacing the used sealant bot
tle.
```

hon 3 (ipykerne... | Initialized (additional servers ne | Instance MEM 5 | Amazon | Cookie Preferen | Mode: Com... | Ln 1, ... | Task3.i... | 1

```
Please note that the tire mobility kit is a temporary solution and is designed to allow you to drive for a maximum
of 10 minutes or 8 km (whichever comes first) at a maximum speed of 80 km/h. It is advisable to replace the punctur
ed tire or have it repaired by a professional as soon as possible.
```

Since the model takes a while to understand the context and generate a relevant answer for you, this might lead to a poor user experience as they have to wait for a response for some seconds.

Bedrock also supports streaming capability where the service generates output as the model generates tokens. Here is an example of how you can implement that.

[8]:
```
from IPython.display import display_markdown,Markdown,clear_output
```

Launcher | Task2b.ipynb | Task2a.ipynb | Task3.ipynb

Markdown | Notebook | Cluster | Python 3 (ipykernel)

[9]:
```
# response with stream
response = bedrock_client.invoke_model_with_response_stream(body=body, modelId=modelId, accept=accept, contentType=c
stream = response.get('body')
output = []
i = 1
if stream:
    for event in stream:
        chunk = event.get('chunk')
        if chunk:
            chunk_obj = json.loads(chunk.get('bytes').decode())
            text = chunk_obj['outputText']
            clear_output(wait=True)
            output.append(text)
            display_markdown(Markdown(''.join(output)))
            i+=1
```

To fix a flat tire on your AnyCompany AC8, you can use the tire mobility kit provided in the luggage area. Here are the steps to follow:

1. Open the tailgate or trunk of your vehicle.
2. Lift up the lining of the luggage area to access the tire mobility kit.
3. Follow the instructions provided with the tire mobility kit to seal the puncture in the tire.
4. After using the kit, make sure to securely put it back in its original location.
5. Contact AnyCompany or an appropriate service for assistance with disposing of and replacing the used sealant bottle.

Please note that the tire mobility kit is a temporary solution and is designed to allow you to drive for a maximum of 10 minutes or 8 km (whichever comes first) at a maximum speed of 80 km/h. It is advisable to replace the punctured tire or have it repaired by a professional as soon as possible.

The response provides summarized, step-by-step instructions on how to change the tires.

You have now learned how you can leverage the Retrieval Augmented Generation (RAG) or the Augmentation process to generate a curated response tailored to the specific context and information provided.