AWS Bedrock Workshop

Michael Lin

Sr. Solutions Architect
Amazon Web Services

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

- Choice of leading FMs through a single API
- Model customization
- Retrieval Augmented Generation (RAG)
- Agents that execute multistep tasks
- Security, privacy, and safety

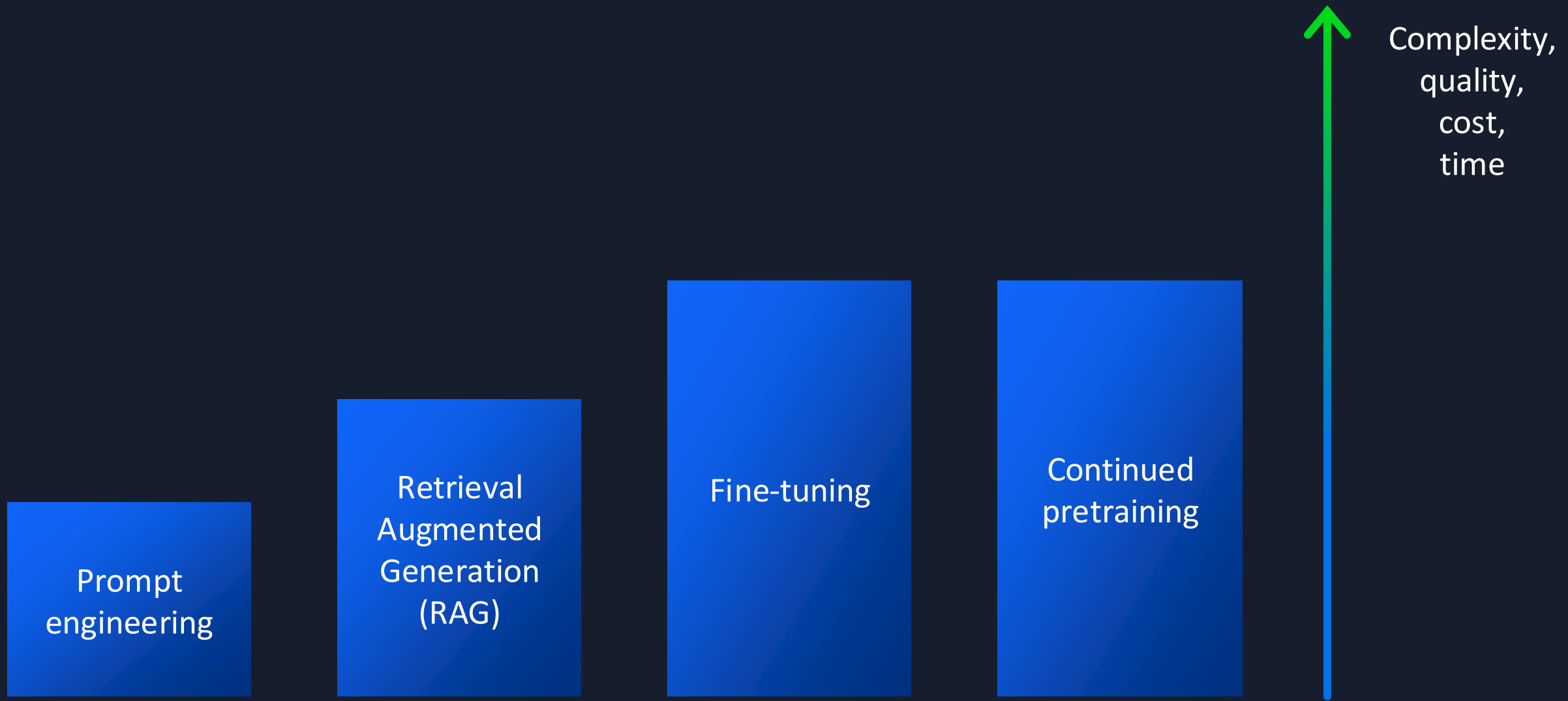
Amazon Bedrock

BROAD CHOICE OF MODELS

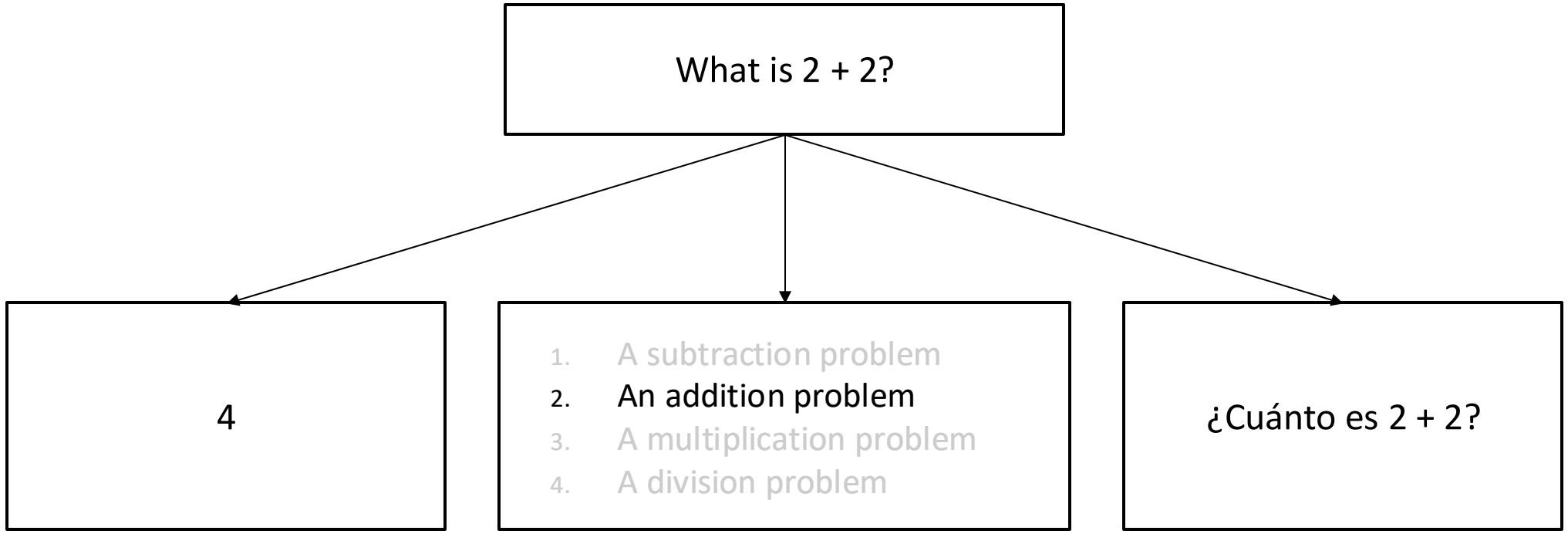
AI21labs	amazon	ANTHROPIC	cohere	Meta	MISTRAL AI_	stability.ai
Contextual answers, summarization, paraphrasing	**Text summarization, generation, Q&A, search, image generation**	**Summarization, complex reasoning, writing, coding**	**Text generation, search, classification**	**Q&A and reading comprehension**	**Text summarization, text classification, text completion, code generation, Q&A**	**High-quality images and art**
Jamba-Instruct	Amazon Titan Text Premier	Claude 3.5 Sonnet	Command	Llama 3 8B	Mistral Small	Stable Diffusion XL1.0
Jurassic-2 Ultra	Amazon Titan Text Lite	Claude 3 Opus	Command Light	Llama 3 70B	Mistral Large	Stable Diffusion XL 0.8
Jurassic-2 Mid	Amazon Titan Text Express	Claude 3 Sonnet	Embed English	Llama 2 13B	Mistral 7B	
	Amazon Titan Text Embeddings	Claude 3 Haiku	Embed Multilingual	Llama 2 70B	Mixtral 8x7B	
	Amazon Titan Text Embeddings V2	Claude 2.1	Command R+			
	Amazon Titan Multimodal Embeddings	Claude 2	Command R			
	Amazon Titan Image Generator	Claude Instant				

aws

Common approaches for customizing FMs



What is prompt engineering?



Prompt engineering is the process of **controlling model behavior** by **optimizing your prompt to elicit high performing LLM responses** (as assessed by rigorous evaluations tailored to your use case).

Parts of a prompt

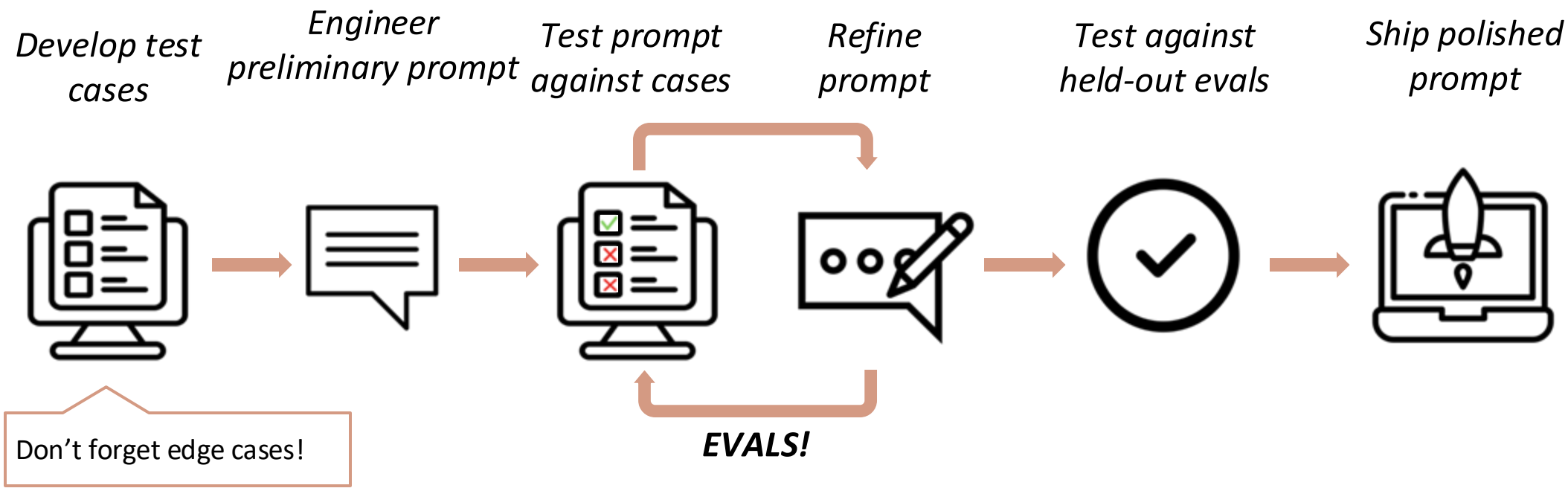
1. Task context
2. Tone context
3. Background data, documents, and images
4. Detailed task description & rules
5. Examples
6. Conversation history
7. Immediate task description or request
8. Thinking step by step / take a deep breath
9. Output formatting
10. Prefilled response (if any)

Example:

User	You will be acting as an AI career coach named Joe created by the company AdAstra Careers. Your goal is to give career advice to users. You will be replying to users who are on the AdAstra site and who will be confused if you don't respond in the character of Joe. You should maintain a friendly customer service tone. Here is the career guidance document you should reference when answering the user: <guide>{{DOCUMENT}}</guide> Here are some important rules for the interaction: - Always stay in character, as Joe, an AI from AdAstra careers - If you are unsure how to respond, say "Sorry, I didn't understand that. Could you repeat the question?" - If someone asks something irrelevant, say, "Sorry, I am Joe and I give career advice. Do you have a career question today I can help you with?" Here is an example of how to respond in a standard interaction: <example> User: Hi, how were you created and what do you do? Joe: Hello! My name is Joe, and I was created by AdAstra Careers to give career advice. What can I help you with today? </example> Here is the conversation history (between the user and you) prior to the question. It could be empty if there is no history: <history> {{HISTORY}} </history> Here is the user's question: <question> {{QUESTION}} </question> How do you respond to the user's question? Think about your answer first before you respond. Put your response in <response></response> tags.
Assistant (prefill)	<response>

How to engineer a good prompt

Empirical science: always test your prompts & iterate often!



Knowledge Bases for Amazon Bedrock

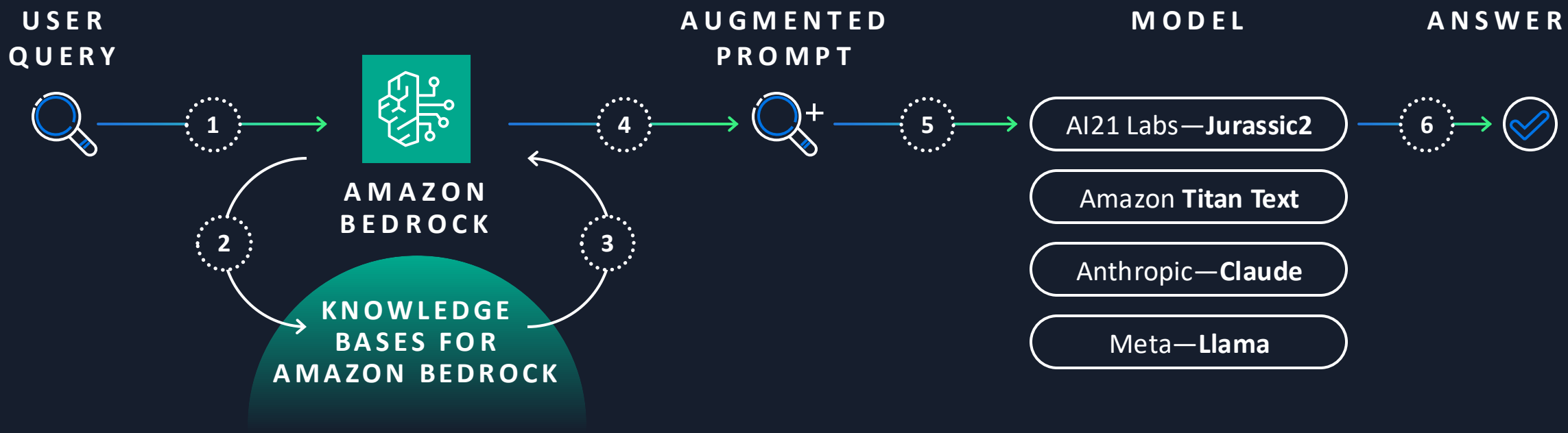
NATIVE SUPPORT FOR RAG

Securely connect FMs to data sources for RAG to deliver more relevant responses

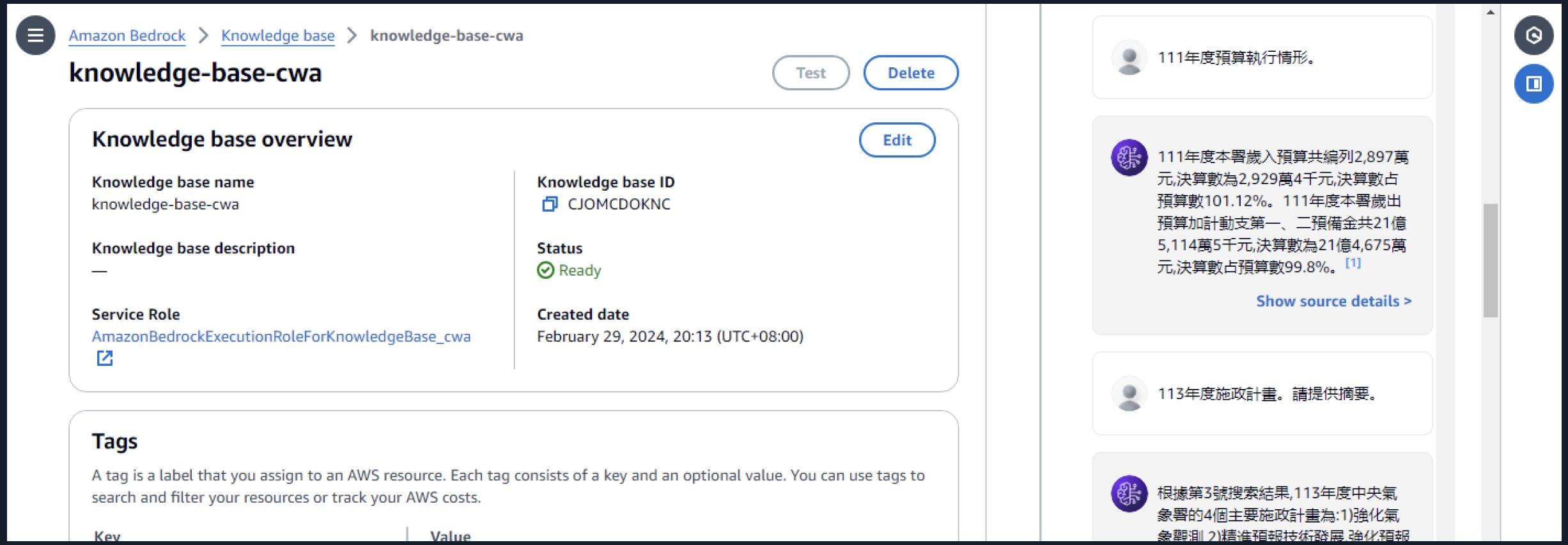
Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multiturn conversations

Automatic citations with retrievals to improve transparency



Bedrock Knowledge Base in Action



中華民國 112 年 10 月
立法院第 10 屆第 8 會期

交通部中央氣象署 113 年度

施政計畫及預算編列口頭報告

一、111 年度預算執行情形

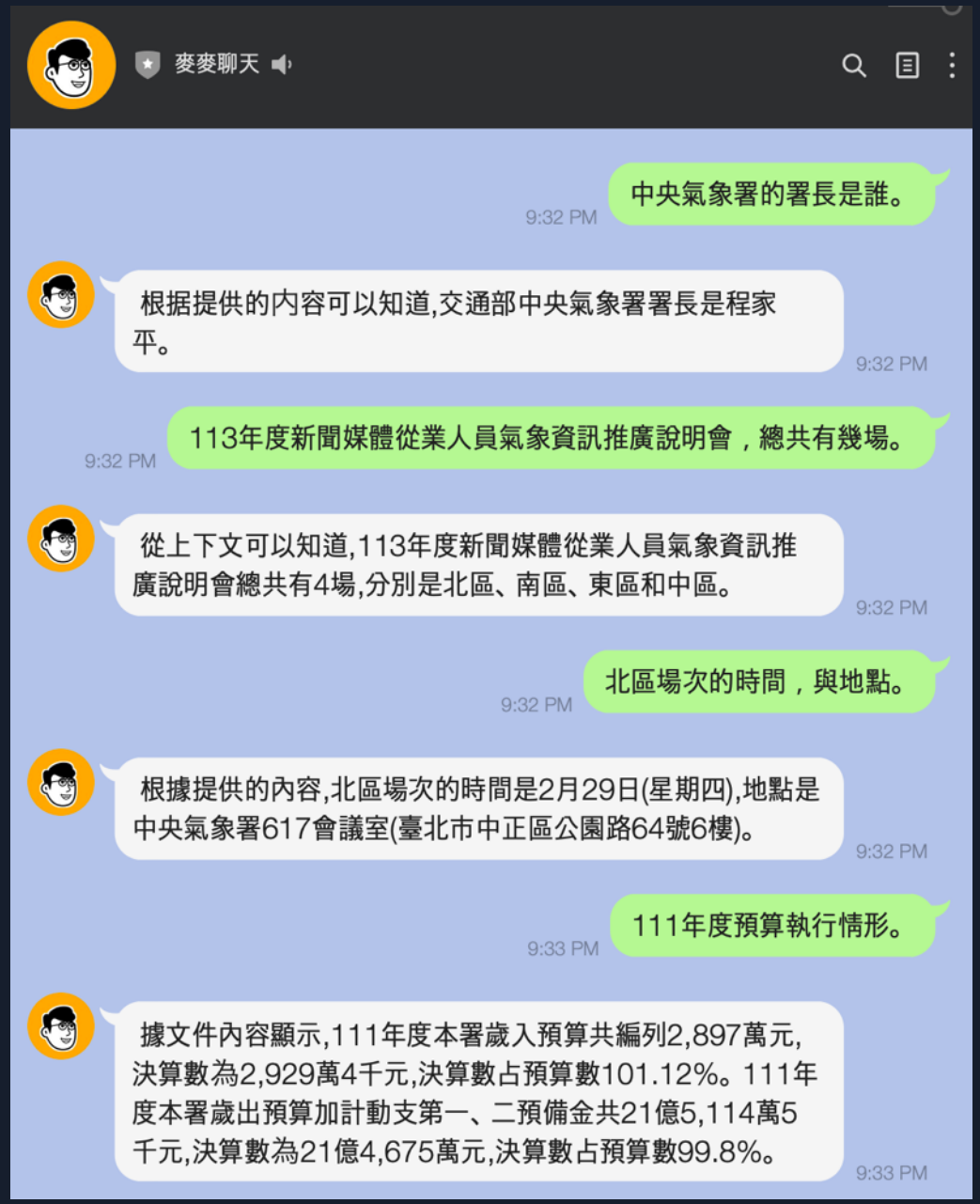
(一)111 年度本署歲入預算共編列 2,897 萬元，決算數為 2,929 萬 4 千元，決算數占預算數 101.12%。

(二)111 年度本署歲出預算加計動支第一、二預備金共 21 億 5,114 萬 5 千元，決算數為 21 億 4,675 萬元，決算數占預算數 99.8%。

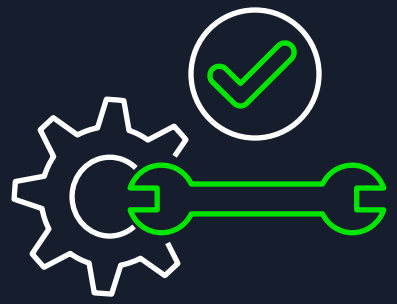
二、112 年度截至 9 月止預算執行情形

(一)112 年度本署歲入預算共編列 2,997 萬元，截至 9 月止預算分配數 1,782 萬 4 千元，實收數 2,548 萬 2 千元，占預算分配數 142.96%。

(二)112 年度本署歲出預算(含預備金動支數)共編列 19 億 692 萬 8 千元，截至 9 月止預算分配數 12 億 2,413 萬 6 千元，執行數 11 億 7,312 萬 7 千元，執行數占預算分配數 95.83%。



Customizing model responses for your business



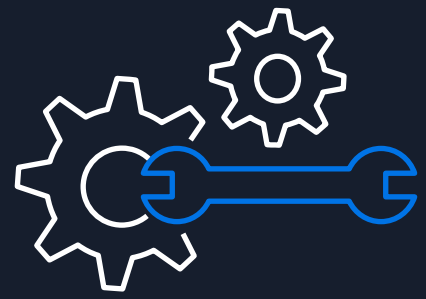
Fine-tuning

PURPOSE

Maximizing accuracy for **specific tasks**

DATA NEED

Small number of labeled examples



Continued pretraining

PURPOSE

Maintaining model accuracy for **your domain**

DATA NEED

Large number of unlabeled datasets

Fine-Tuning in Action: Text Summarization

```
prompt = """
Summarize the simplest and most interesting part of the following conversation.

#Person1#: Hello. My name is John Sandals, and I've got a reservation.
#Person2#: May I see some identification, sir, please?
#Person1#: Sure. Here you are.
#Person2#: Thank you so much. Have you got a credit card, Mr. Sandals?
#Person1#: I sure do. How about American Express?
#Person2#: Unfortunately, at the present time we take only MasterCard or VISA.
#Person1#: No American Express? Okay, here's my VISA.
#Person2#: Thank you, sir. You'll be in room 507, nonsmoking, with a queen-size bed. Do you approve, sir?
#Person1#: Yeah, that'll be fine.
#Person2#: That's great. This is your key, sir. If you need anything at all, anytime, just dial zero.

Summary:
"""

body = {
    "prompt": prompt,
    "temperature": 0.5,
    "top_p": 0.9,
    "max_gen_len": 512,
}
```

Fine-Tuning in Action: Baseline Completion

```
response = bedrock_runtime.invoke_model(
    modelId="meta.llama2-13b-chat-v1", # compare to chat model
    body=json.dumps(body)
)

response_body = response["body"].read().decode('utf8')
print(json.loads(response_body)["generation"])
```

```
A man named John Sandals checks into a hotel and provides identification and a credit card. The hotel only takes M
asterCard or VISA, so he uses his VISA card. He is given room 507, a nonsmoking room with a queen-size bed.
```

Fine-Tuning in Action: Improved Completion

```
response = bedrock_runtime.invoke_model(
    modelId=provisioned_model_arn, # custom fine-tuned model
    body=json.dumps(body)
)

response_body = response["body"].read().decode('utf8')
print(json.loads(response_body)["generation"])
```

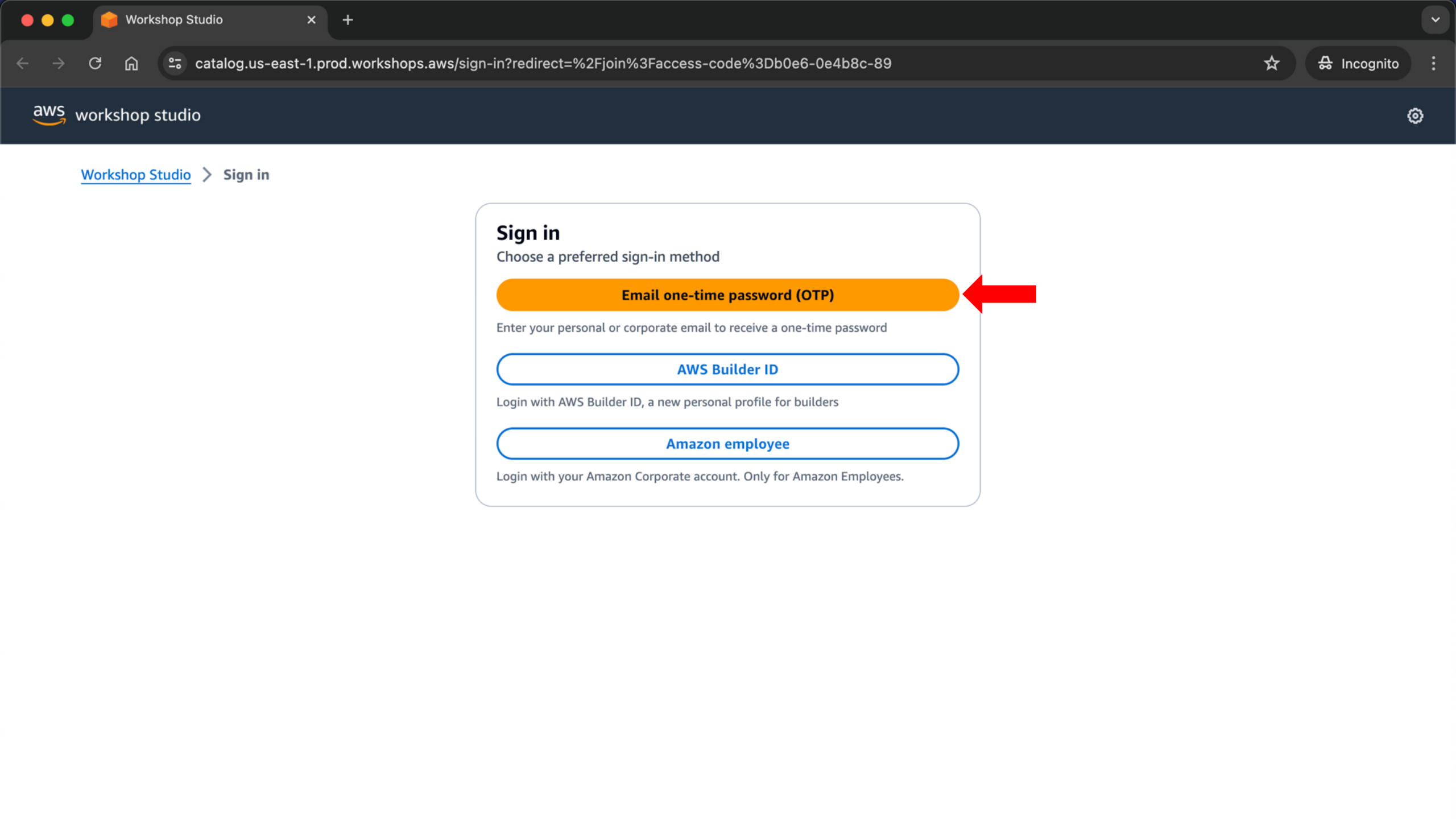
```
John Sandals checks in the hotel with VISA and is assigned room 507, nonsmoking, with a queen-size bed.
```

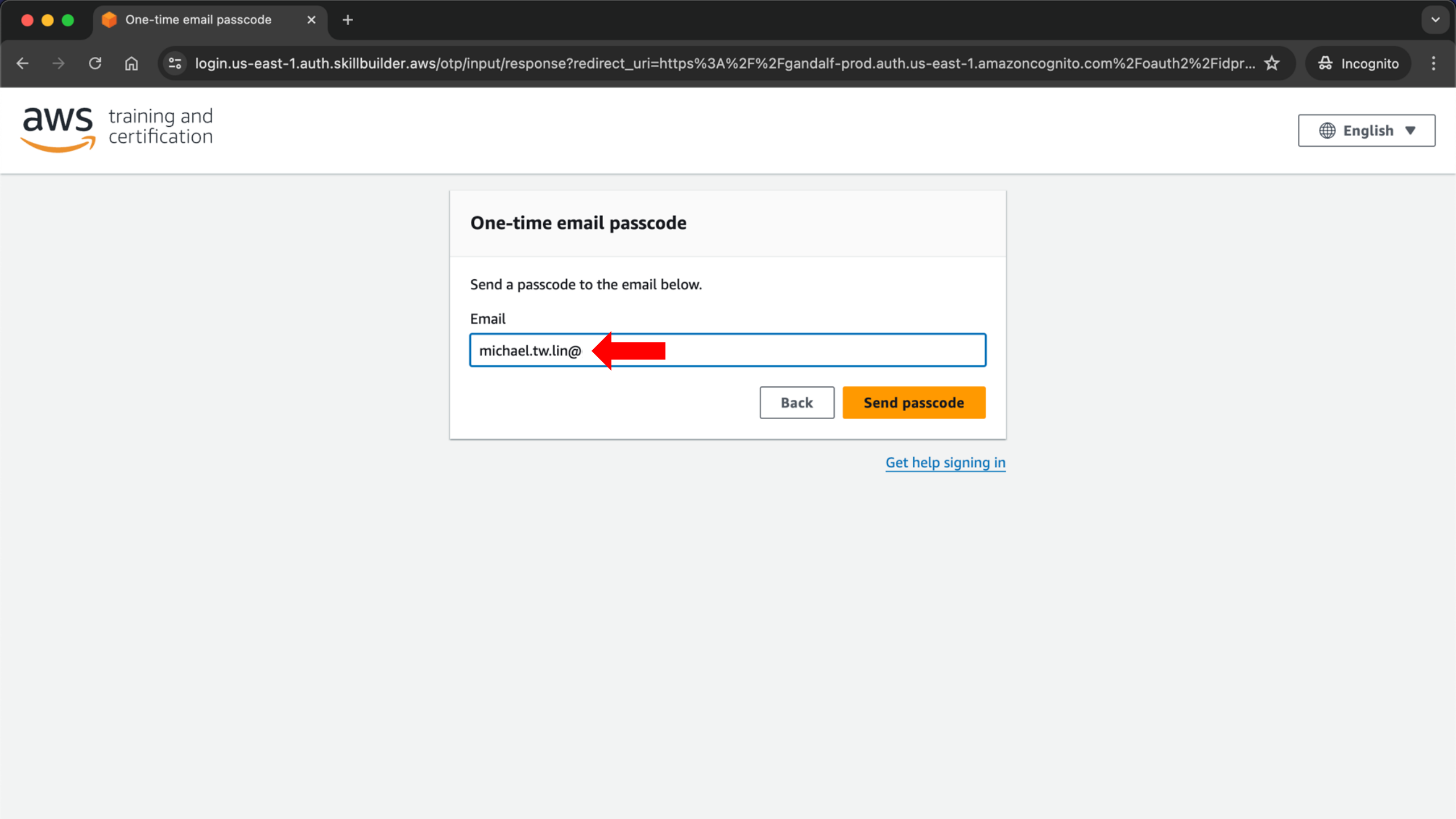
https://reurl.cc/2jmo66

https://reurl.cc/NlRZbk

Workshop Setup

- Workshop Access
- Bedrock Model Access
- SageMaker Studio Access


Workshop Studio
catalog.us-east-1.prod.workshops.aws/sign-in?redirect=%2Fjoin%3Faccess-code%3Db0e6-0e4b8c-89
Incognito
aws workshop studio
Workshop Studio
Sign in
Sign in
Choose a preferred sign-in method
Email one-time password (OTP)
Enter your personal or corporate email to receive a one-time password
AWS Builder ID
Login with AWS Builder ID, a new personal profile for builders
Amazon employee
Login with your Amazon Corporate account. Only for Amazon Employees.


One-time email passcode
login.us-east-1.auth.skillbuilder.aws/otp/input/response?redirect_uri=https%3A%2F%2Fgandalf-prod.auth.us-east-1.amazoncognito.com%2Foauth2%2Fidpr...
Incognito
aws training and certification
English
One-time email passcode
Send a passcode to the email below.
Email
michael.tw.lin@
Back
Send passcode
Get help signing in

Verify one-time email passco

login.us-east-1.auth.skillbuilder.aws/otp/challenge?redirect_uri=https%3A%2F%2Fgandalf-prod.auth.us-east-1.amazoncognito.com%2Foauth2%2Fidprespo...

Incognito

aws training and certification

English

One-time email passcode

We sent a passcode to michael.tw.lin@gmail.com. You should receive it within 5 minutes.

Passcode (9-digit) | Resend passcode

529717102

Back | Sign in

Get help signing in

Step 1
Enter event access code

Step 2
Review and join

Review and join

Event details

Name	Start time	Duration	Level
TCC-Bedrock-Dryrun	7/27/2024 03:02 PM	72 hours	300

Description
TCC-Bedrock-Dryrun

Terms and Conditions

Read and accept before joining the event

Read and accept before joining the event:

1. By using AWS Workshop Studio for the relevant event, You agree to the AWS Event Terms and Conditions, the AWS Responsible AI Policy, and the AWS Acceptable Use Policy.
2. If You are under 18 years old, you may participate in the relevant event using AWS Workshop Studio: (a) if You are at least the minimum age below based on the country or region in which You reside, and (b) with the involvement of a parent, guardian, or educator.

Country or region	Minimum age
All countries or regions not listed below (including the United States, Brazil, the United Kingdom, and India)	13
Canada, China, Republic of Korea (South Korea)	14
Australia	15

3. You acknowledge and agree that You are using an AWS-owned account that You will only be able to access during the relevant event. You have no ownership rights over this AWS-owned account.
4. During the relevant event, while using this AWS-owned account, You will not use, import, input, or introduce any data, dataset, or other material that contains personal data, financial information, or any other data or materials that may be subject to laws and regulations (such as the General Data Protection Regulation or The Health Insurance Portability and Accountability Act of 1996).
5. If You find residual resources or materials in this AWS-owned account, You will notify your Event Operator immediately.
6. AWS, its affiliates, and any entities or persons acting on AWS's behalf reserves the right to terminate this AWS-owned account and to delete its contents at any time, without any notice to You.
7. During the relevant event, while using this AWS-owned account, You will not process or run any operation on any data other than test datasets or lab materials that have been approved by AWS.
8. You will not copy, import, export or otherwise create derivative works of materials provided by AWS for use outside of the relevant event.
9. AWS, its affiliates, and any entities or persons acting on AWS's behalf have no obligation to enable the transmission of Your materials through AWS Workshop Studio, and may, in their discretion, edit, block, refuse to post, or remove Your materials at any time, without notice to You.
10. If You access and use a service and/or third-party models that have their own terms during the relevant event, while in the AWS-owned account, You agree to review those terms and comply with them during the event.
11. If You are an AWS Partner using AWS Workshop Studio as part of Your participation in the AWS Partner Network Program, Your use of AWS Workshop Studio is governed by these terms, the AWS Partner Network Terms and Conditions, and the AWS Customer Agreement or other agreement with us governing your use of AWS Services.
12. Your use of AWS Workshop Studio will comply with these terms and all applicable laws. If You fail to comply with any of these terms, Your access to AWS Workshop Studio may be immediately terminated, without notice to You.

☐ I agree with the Terms and Conditions

Cancel Previous Join event

3. You acknowledge and agree that You are using an AWS-owned account that You will only be able to access during the relevant event. You have no ownership rights over this AWS-owned account.
4. During the relevant event, while using this AWS-owned account, You will not use, import, input, or introduce any data, dataset, or other material that contains personal data, financial information, or any other data or materials that may be subject to laws and regulations (such as the General Data Protection Regulation or The Health Insurance Portability and Accountability Act of 1996).
5. If You find residual resources or materials in this AWS-owned account, You will notify your Event Operator immediately.
6. AWS, its affiliates, and any entities or persons acting on AWS's behalf reserves the right to terminate this AWS-owned account and to delete its contents at any time, without any notice to You.
7. During the relevant event, while using this AWS-owned account, You will not process or run any operation on any data other than test datasets or lab materials that have been approved by AWS.
8. You will not copy, import, export or otherwise create derivative works of materials provided by AWS for use outside of the relevant event.
9. AWS, its affiliates, and any entities or persons acting on AWS's behalf have no obligation to enable the transmission of Your materials through AWS Workshop Studio, and may, in their discretion, edit, block, refuse to post, or remove Your materials at any time, without notice to You.
10. If You access and use a service and/or third-party models that have their own terms during the relevant event, while in the AWS-owned account, You agree to review those terms and comply with them during the event.
11. If You are an AWS Partner using AWS Workshop Studio as part of Your participation in the AWS Partner Network Program, Your use of AWS Workshop Studio is governed by these terms, the AWS Partner Network Terms and Conditions, and the AWS Customer Agreement or other agreement with us governing your use of AWS Services.
12. Your use of AWS Workshop Studio will comply with these terms and all applicable laws. If You fail to comply with any of these terms, Your access to AWS Workshop Studio may be immediately terminated, without notice to You.

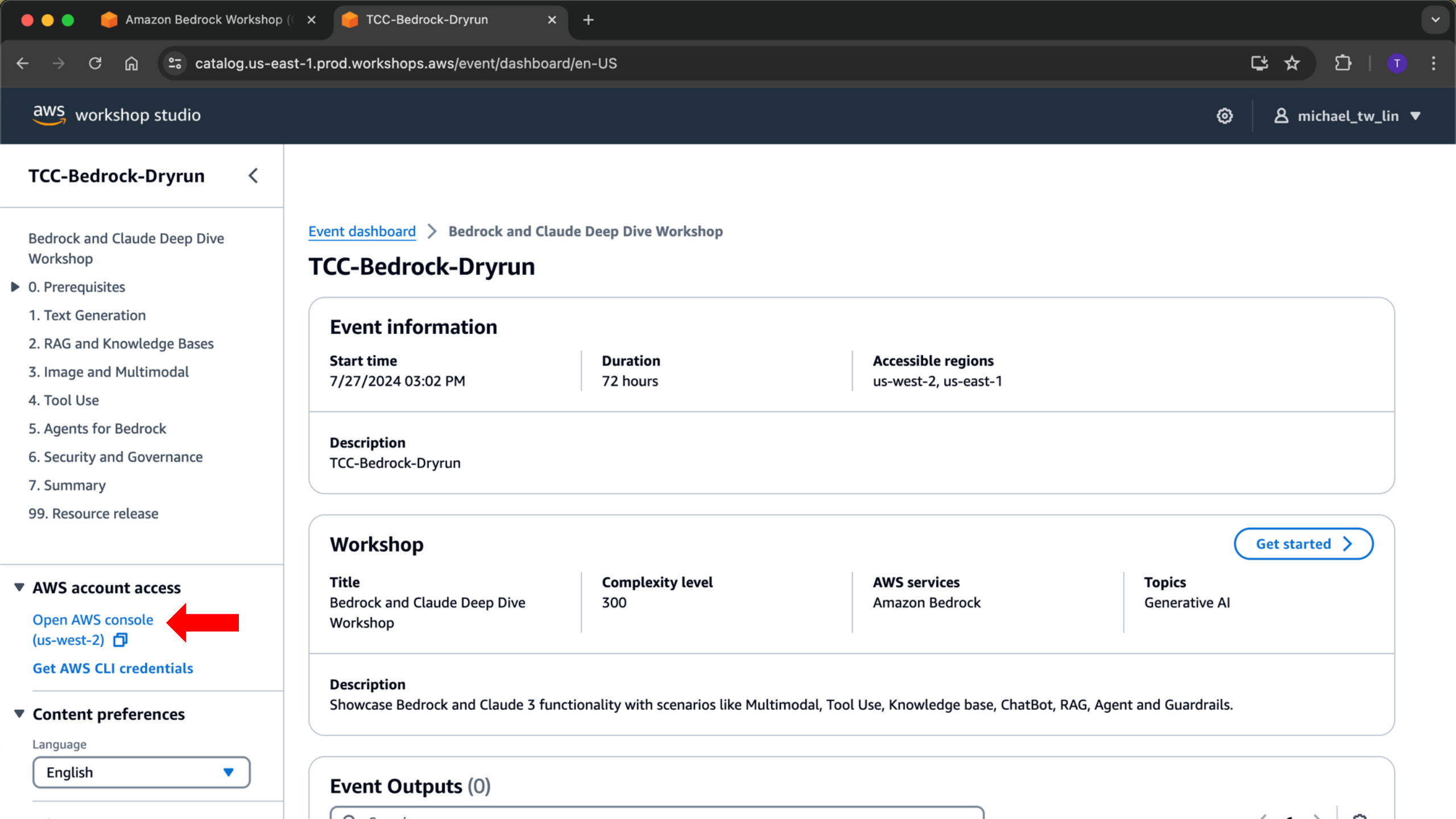
☑ I agree with the Terms and Conditions

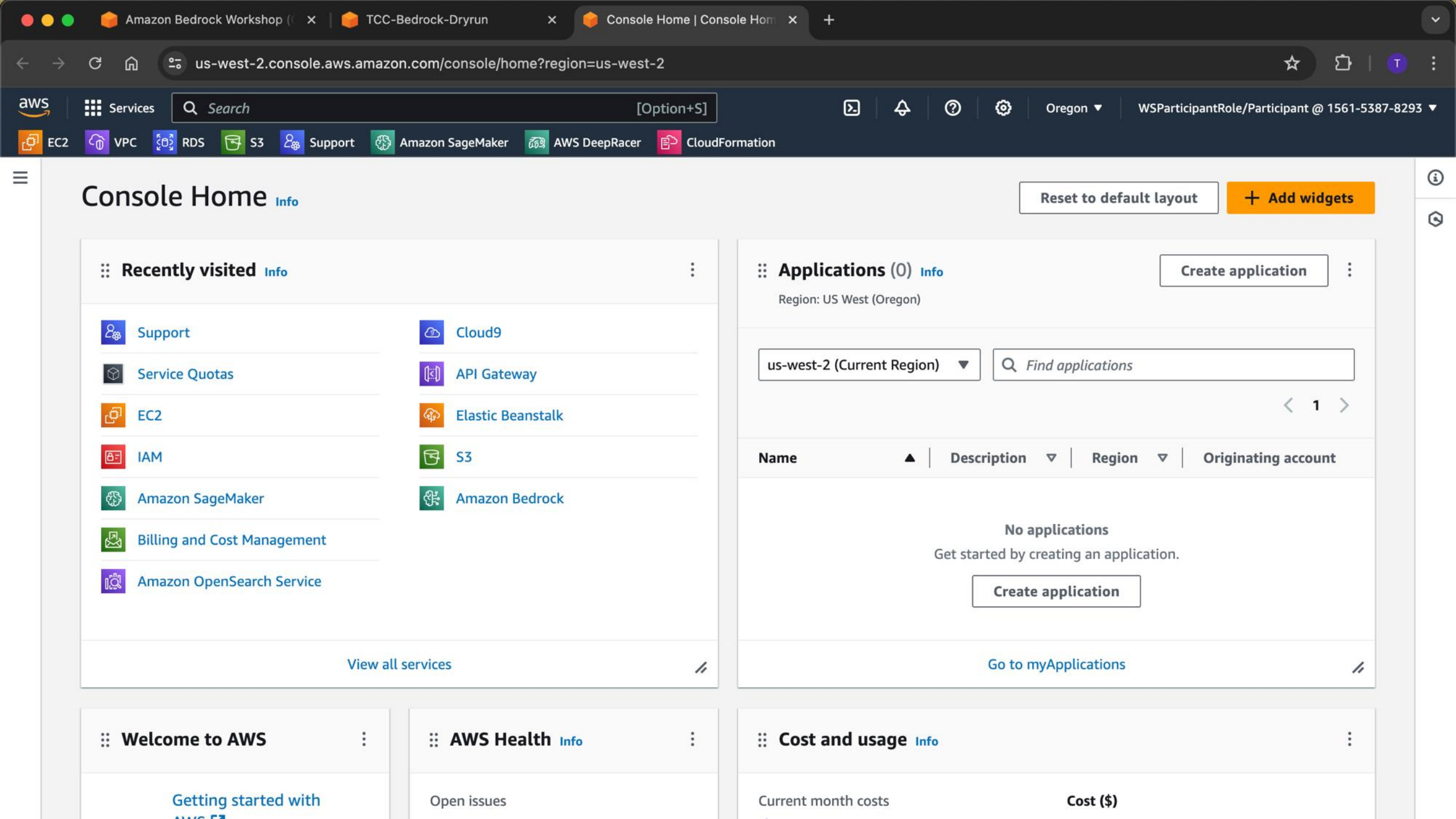
Cancel | Previous | Join event

3. You acknowledge and agree that You are using an AWS-owned account that You will only be able to access during the relevant event. You have no ownership rights over this AWS-owned account.
4. During the relevant event, while using this AWS-owned account, You will not use, import, input, or introduce any data, dataset, or other material that contains personal data, financial information, or any other data or materials that may be subject to laws and regulations (such as the General Data Protection Regulation or The Health Insurance Portability and Accountability Act of 1996).
5. If You find residual resources or materials in this AWS-owned account, You will notify your Event Operator immediately.
6. AWS, its affiliates, and any entities or persons acting on AWS's behalf reserves the right to terminate this AWS-owned account and to delete its contents at any time, without any notice to You.
7. During the relevant event, while using this AWS-owned account, You will not process or run any operation on any data other than test datasets or lab materials that have been approved by AWS.
8. You will not copy, import, export or otherwise create derivative works of materials provided by AWS for use outside of the relevant event.
9. AWS, its affiliates, and any entities or persons acting on AWS's behalf have no obligation to enable the transmission of Your materials through AWS Workshop Studio, and may, in their discretion, edit, block, refuse to post, or remove Your materials at any time, without notice to You.
10. If You access and use a service and/or third-party models that have their own terms during the relevant event, while in the AWS-owned account, You agree to review those terms and comply with them during the event.
11. If You are an AWS Partner using AWS Workshop Studio as part of Your participation in the AWS Partner Network Program, Your use of AWS Workshop Studio is governed by these terms, the AWS Partner Network Terms and Conditions, and the AWS Customer Agreement or other agreement with us governing your use of AWS Services.
12. Your use of AWS Workshop Studio will comply with these terms and all applicable laws. If You fail to comply with any of these terms, Your access to AWS Workshop Studio may be immediately terminated, without notice to You.

☑ I agree with the Terms and Conditions

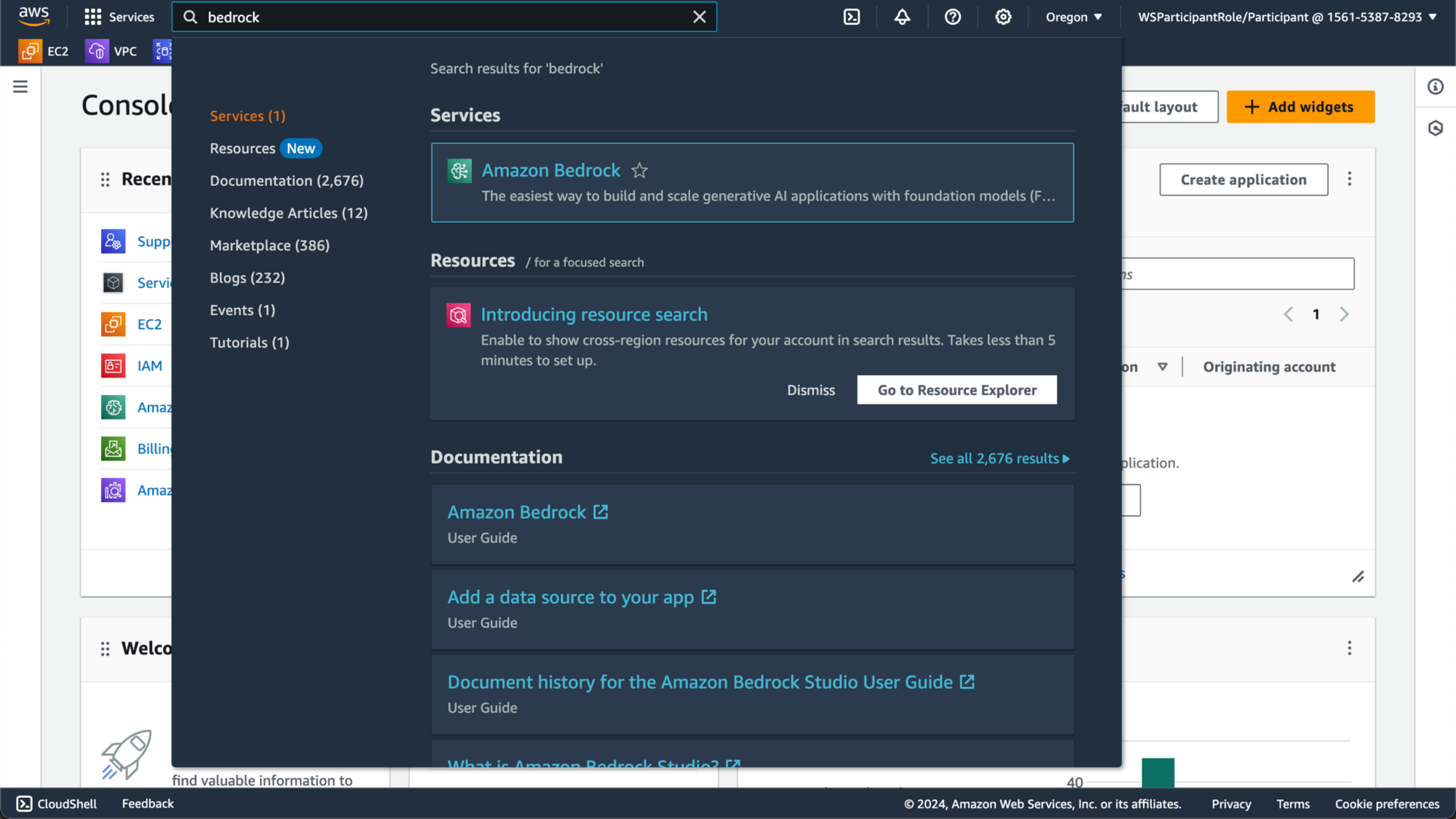
Cancel | Previous | Join event

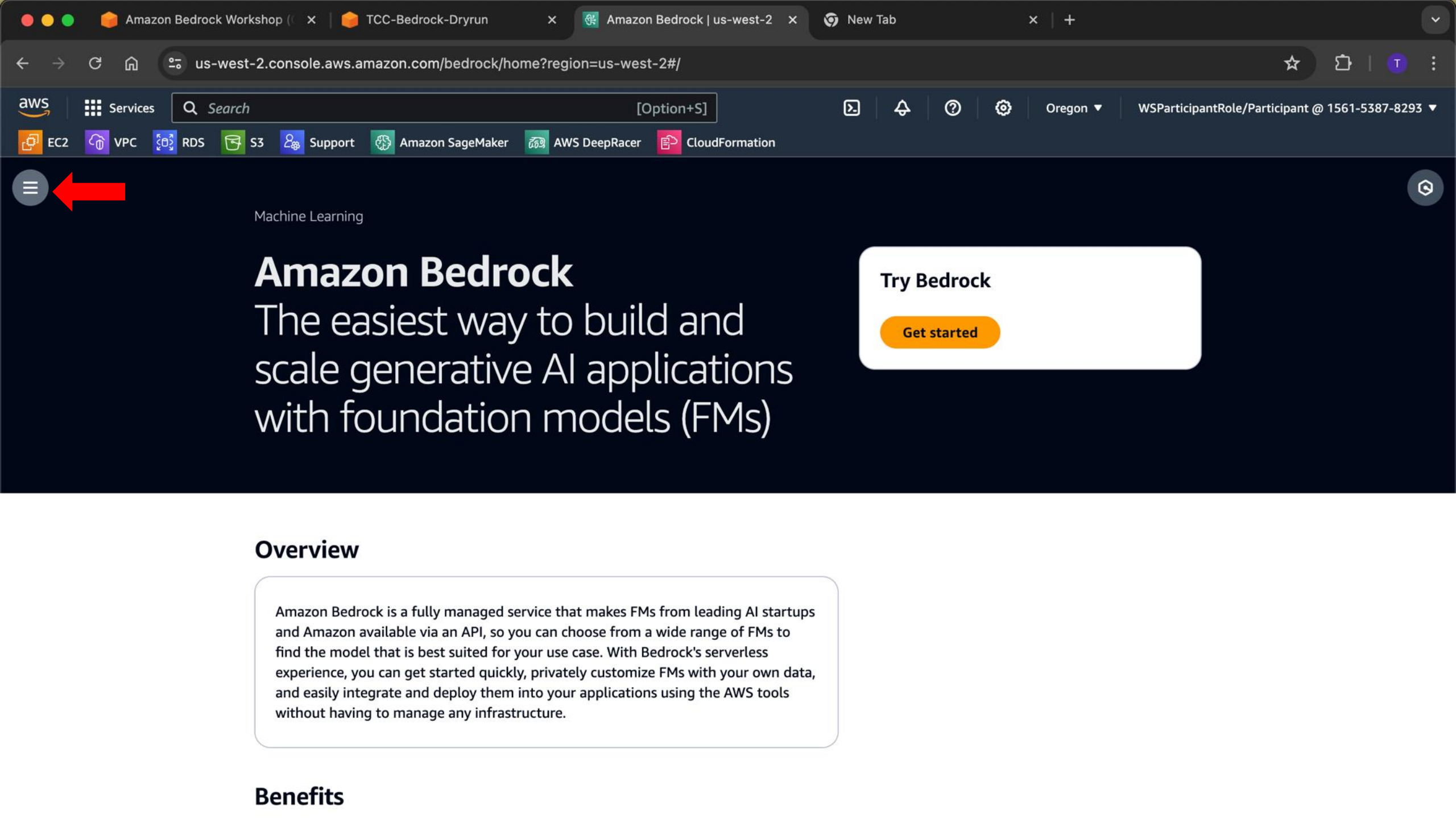

Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
catalog.us-east-1.prod.workshops.aws/event/dashboard/en-US
aws workshop studio
michael_tw_lin
TCC-Bedrock-Dryrun
Bedrock and Claude Deep Dive Workshop
0. Prerequisites
1. Text Generation
2. RAG and Knowledge Bases
3. Image and Multimodal
4. Tool Use
5. Agents for Bedrock
6. Security and Governance
7. Summary
99. Resource release
AWS account access
Open AWS console (us-west-2)
Get AWS CLI credentials
Content preferences
Language
English
Event dashboard
Bedrock and Claude Deep Dive Workshop
TCC-Bedrock-Dryrun
Event information
Start time
7/27/2024 03:02 PM
Duration
72 hours
Accessible regions
us-west-2, us-east-1
Description
TCC-Bedrock-Dryrun
Workshop
Get started
Title
Bedrock and Claude Deep Dive Workshop
Complexity level
300
AWS services
Amazon Bedrock
Topics
Generative AI
Description
Showcase Bedrock and Claude 3 functionality with scenarios like Multimodal, Tool Use, Knowledge base, ChatBot, RAG, Agent and Guardrails.
Event Outputs (0)

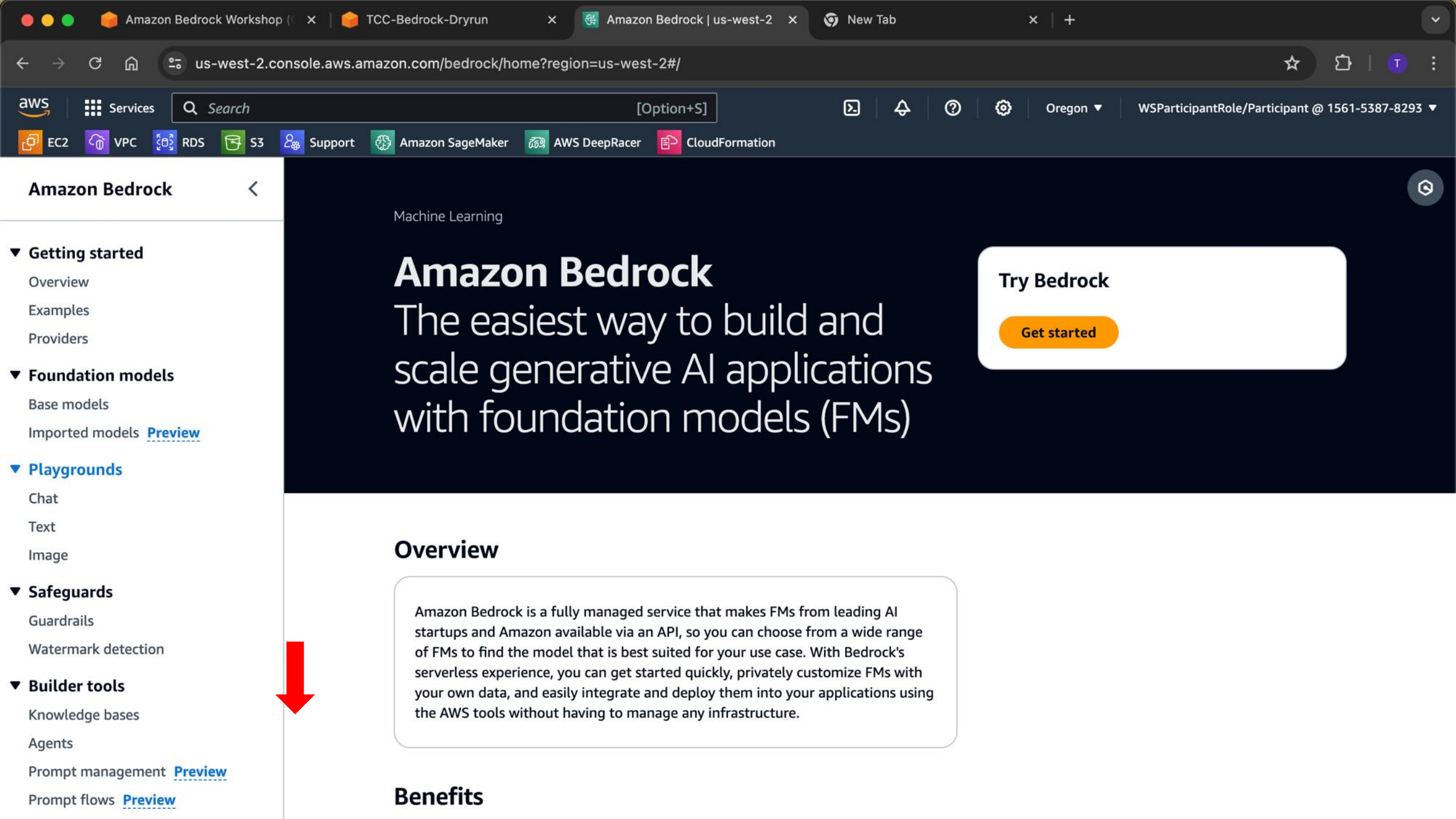

Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Console Home | Console Hom
us-west-2.console.aws.amazon.com/console/home?region=us-west-2
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Console Home Info
Reset to default layout
+ Add widgets
Recently visited Info
Support
Service Quotas
EC2
IAM
Amazon SageMaker
Billing and Cost Management
Amazon OpenSearch Service
Cloud9
API Gateway
Elastic Beanstalk
S3
Amazon Bedrock
View all services
Applications (0) Info
Region: US West (Oregon)
Create application
us-west-2 (Current Region)
Find applications
1
Name
Description
Region
Originating account
No applications
Get started by creating an application.
Create application
Go to myApplications
Welcome to AWS
Getting started with
AWS Health Info
Open issues
Cost and usage Info
Current month costs
Cost ($)

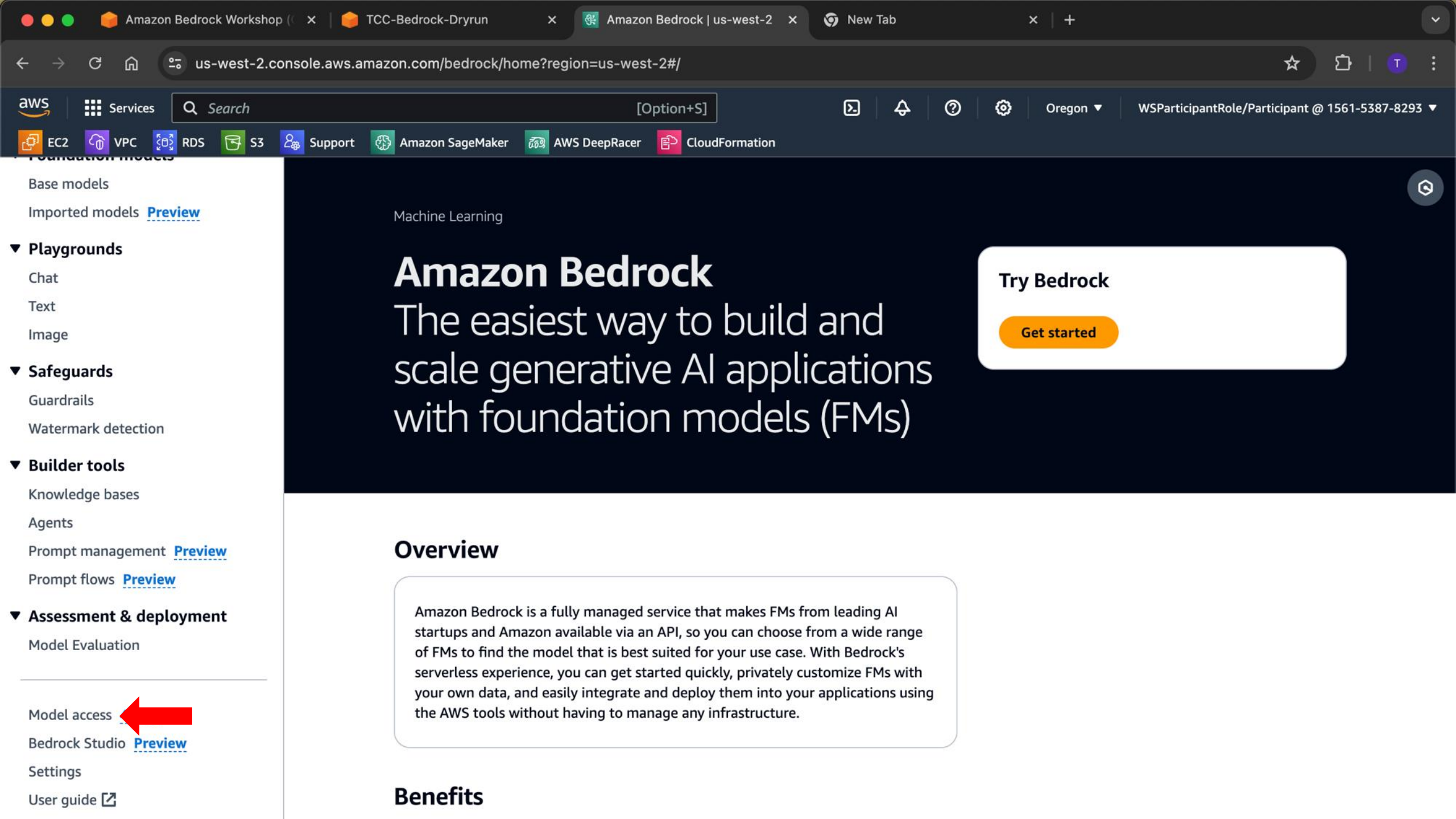
Workshop Setup

- Workshop Access
- Bedrock Model Access
- SageMaker Studio Access


aws
Services
bedrock
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
Search results for 'bedrock'
Services (1)
Resources New
Documentation (2,676)
Knowledge Articles (12)
Marketplace (386)
Blogs (232)
Events (1)
Tutorials (1)
Services
Amazon Bedrock
The easiest way to build and scale generative AI applications with foundation models (F...
Resources / for a focused search
Introducing resource search
Enable to show cross-region resources for your account in search results. Takes less than 5 minutes to set up.
Dismiss
Go to Resource Explorer
Documentation
See all 2,676 results
Amazon Bedrock
User Guide
Add a data source to your app
User Guide
Document history for the Amazon Bedrock Studio User Guide
User Guide
Add widgets
Create application
Originating account
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Machine Learning
Amazon Bedrock
The easiest way to build and scale generative AI applications with foundation models (FMs)
Try Bedrock
Get started
Overview
Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.
Benefits


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Machine Learning
Amazon Bedrock
The easiest way to build and scale generative AI applications with foundation models (FMs)
Try Bedrock
Get started
Overview
Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.
Benefits


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Base models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Assessment & deployment
Model Evaluation
Model access
Bedrock Studio Preview
Settings
User guide
Machine Learning
Amazon Bedrock
The easiest way to build and scale generative AI applications with foundation models (FMs)
Try Bedrock
Get started
Overview
Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.
Benefits

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2 | New Tab

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/modelaccess

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

What is Model access?

To use Bedrock, account users with the correct IAM Permissions must enable access to available Bedrock foundation models (FMs). View all Bedrock Model Terms for Bedrock FMs.

Enable all models | Enable specific models

Visit Amazon Bedrock Quotas for a quick guide to the default quotas and limits that apply to Amazon Bedrock.

Base models (33)

Collapse all

Not seeing a model you're interested in? Check out all supported models by region here.

Find model

Group by provider

Models	Access status	Modality	EULA
AI21 Labs (2)	0/2 access granted		
Jurassic-2 Ultra	Available to request	Text	EULA
Jurassic-2 Mid	Available to request	Text	EULA
Amazon (6)	0/6 access granted		
Titan Embeddings G1 - Text	Available to request	Embedding	EULA

Step 1
Edit model access
Step 2
Review and submit

Edit model access

Base models (33/33)

Collapse all

Not seeing a model you're interested in? Check out all supported models by region here.

Find model

Group by provider

Models	Access status	Modality	EULA
AI21 Labs (2)	0/2 access granted		
Jurassic-2 Ultra	Available to request	Text	EULA
Jurassic-2 Mid	Available to request	Text	EULA
Amazon (6)	0/6 access granted		
Titan Embeddings G1 - Text	Available to request	Embedding	EULA
Titan Text G1 - Lite	Available to request	Text	EULA
Titan Text G1 - Express	Available to request	Text	EULA
Titan Image Generator G1	Available to request	Image	EULA
Titan Multimodal Embeddings G1	Available to request	Embedding	EULA
Titan Text Embeddings V2	Available to request	Embedding	EULA
Anthropic (5)	0/5 access granted		

☑	Llama 3.1 70B Instruct	Available to request	Text	EULA
☑	Llama 3.1 8B Instruct	Available to request	Text	EULA
☑	Llama 3 8B Instruct	Available to request	Text	EULA
☑	Llama 3 70B Instruct	Available to request	Text	EULA
☑	Llama 2 Chat 13B	Unavailable	Text	EULA
☑	Llama 2 Chat 70B	Unavailable	Text	EULA
☑	Llama 2 13B	Unavailable	Text	EULA
☑	Llama 2 70B	Unavailable	Text	EULA
☑	▼ Mistral AI (4)	0/4 access granted		
☑	Mistral Large (2407)	Available to request	Text	EULA
☑	Mistral 7B Instruct	Available to request	Text	EULA
☑	Mixtral 8x7B Instruct	Available to request	Text	EULA
☑	Mistral Large (2402)	Available to request	Text	EULA
☑	▼ Stability AI (1)	0/1 access granted		
☑	SDXL 1.0	Available to request	Image	EULA

Cancel Next

Step 1
Edit model access

Step 2
Review and submit

Review and submit

Step 1: Edit model access

Edit

Model access modifications (33)

Models	Modifications
Mistral Large (2407)	Request access
Jurassic-2 Ultra	Request access
Jurassic-2 Mid	Request access
Claude 3 Opus	Request access
Claude 3 Sonnet	Request access
Claude 3 Haiku	Request access
Claude	Request access
Claude Instant	Request access
SDXL 1.0	Request access
Command R+	Request access

Model access modifications (33)

Models	Modifications
Llama 3 8B Instruct	Request access
Llama 3 70B Instruct	Request access
Llama 2 Chat 13B	Request access
Llama 2 Chat 70B	Request access
Llama 2 13B	Request access
Llama 2 70B	Request access
Mistral 7B Instruct	Request access
Mixtral 8x7B Instruct	Request access
Mistral Large (2402)	Request 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

Model access modifications (33)

Models	Modifications
Llama 3 8B Instruct	Request access
Llama 3 70B Instruct	Request access
Llama 2 Chat 13B	Request access
Llama 2 Chat 70B	Request access
Llama 2 13B	Request access
Llama 2 70B	Request access
Mistral 7B Instruct	Request access
Mixtral 8x7B Instruct	Request access
Mistral Large (2402)	Request 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

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2 | New Tab

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/modelaccess

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
  - Prompt management Preview

Access request for 7 models failed

- Claude 3 Opus - Unauthorized to perform action due to private marketplace eligibility
- Command R+ - Unauthorized to perform action due to private marketplace eligibility
- Command R - Unauthorized to perform action due to private marketplace eligibility
- Llama 2 13B - Could not create agreement - Agreement already exists
- Llama 2 70B - Could not create agreement - Agreement already exists
- Llama 2 Chat 13B - Failed to create regional entitlement. Model not available at the moment. Try again later.
- Llama 2 Chat 70B - Failed to create regional entitlement. Model not available at the moment. Try again later.

Notifications 1 0 0 1 0

Amazon Bedrock > Model access

What is Model access?

To use Bedrock, account users with the correct IAM Permissions must enable access to available Bedrock foundation models (FMs). View all Bedrock Model Terms for Bedrock FMs.

Modify model access

Visit Amazon Bedrock Quotas for a quick guide to the default quotas and limits that apply to Amazon Bedrock.

Base models (33)

Collapse all

Not seeing a model you're interested in? Check out all supported models by region here.

Find model

Group by provider

Models	Access status	Modality	EULA

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2 | New Tab

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/modelaccess

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Model access updates submitted
It may take several minutes to receive or remove access to models. Meanwhile, you can use other Bedrock console pages. Once your access is updated, you can use the models in Bedrock. Refresh the base models table to view the updated statuses.

What is Model access?

To use Bedrock, account users with the correct IAM Permissions must enable access to available Bedrock foundation models (FMs). View all Bedrock Model Terms for Bedrock FMs.

Modify model access

Visit Amazon Bedrock Quotas for a quick guide to the default quotas and limits that apply to Amazon Bedrock.

Base models (33)

Collapse all

Not seeing a model you're interested in? Check out all supported models by region here.

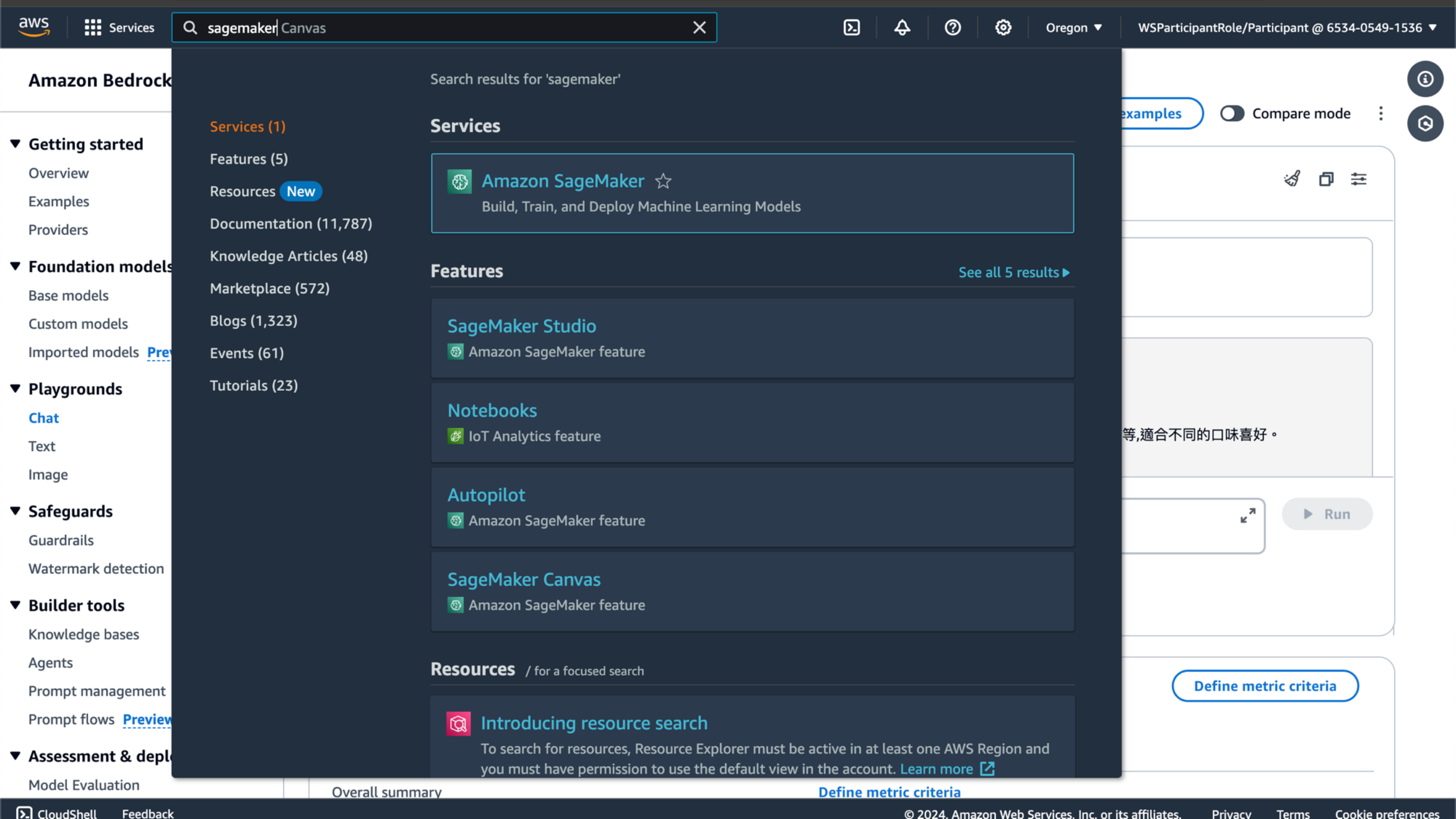
Find model

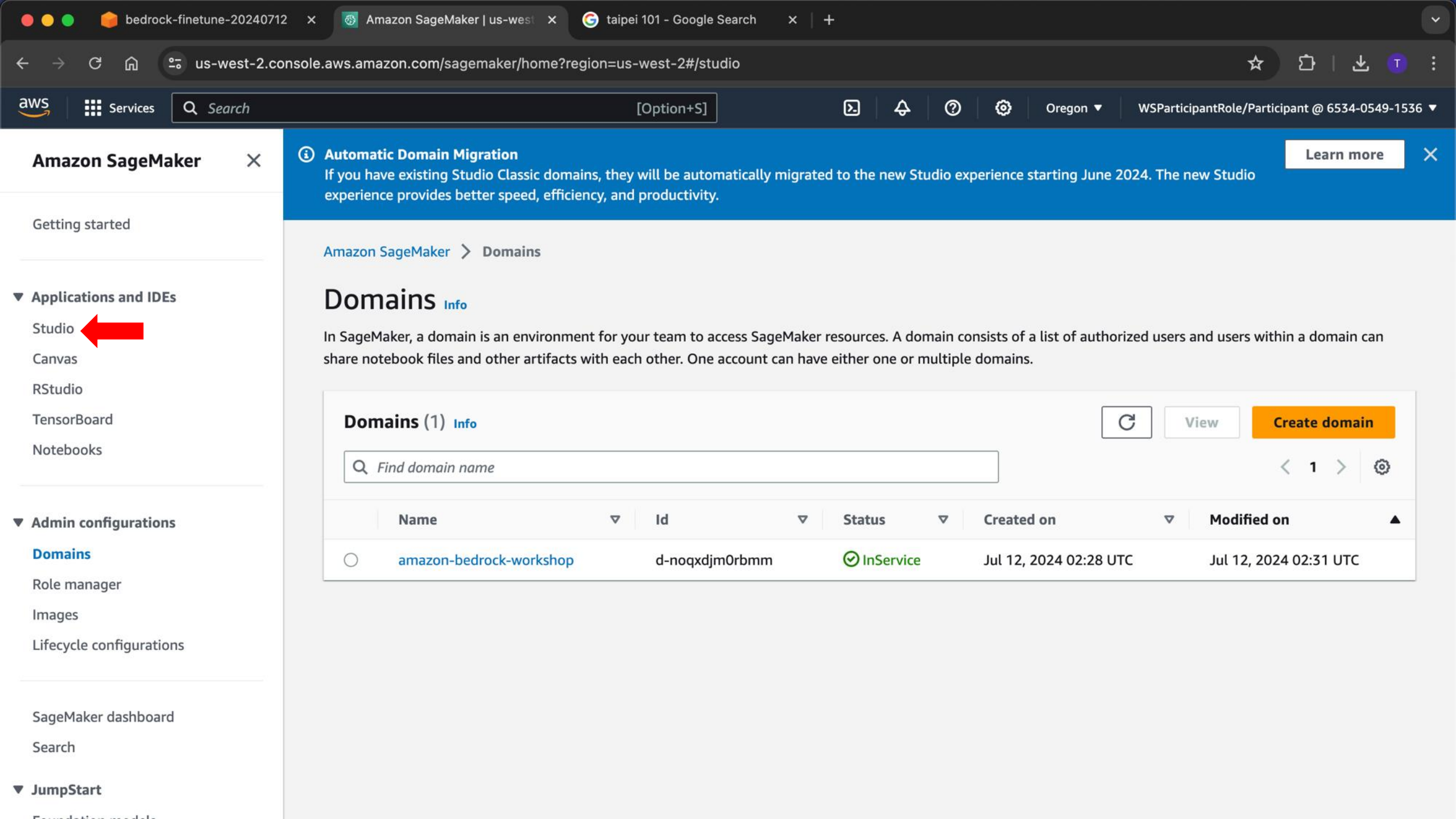
Group by provider

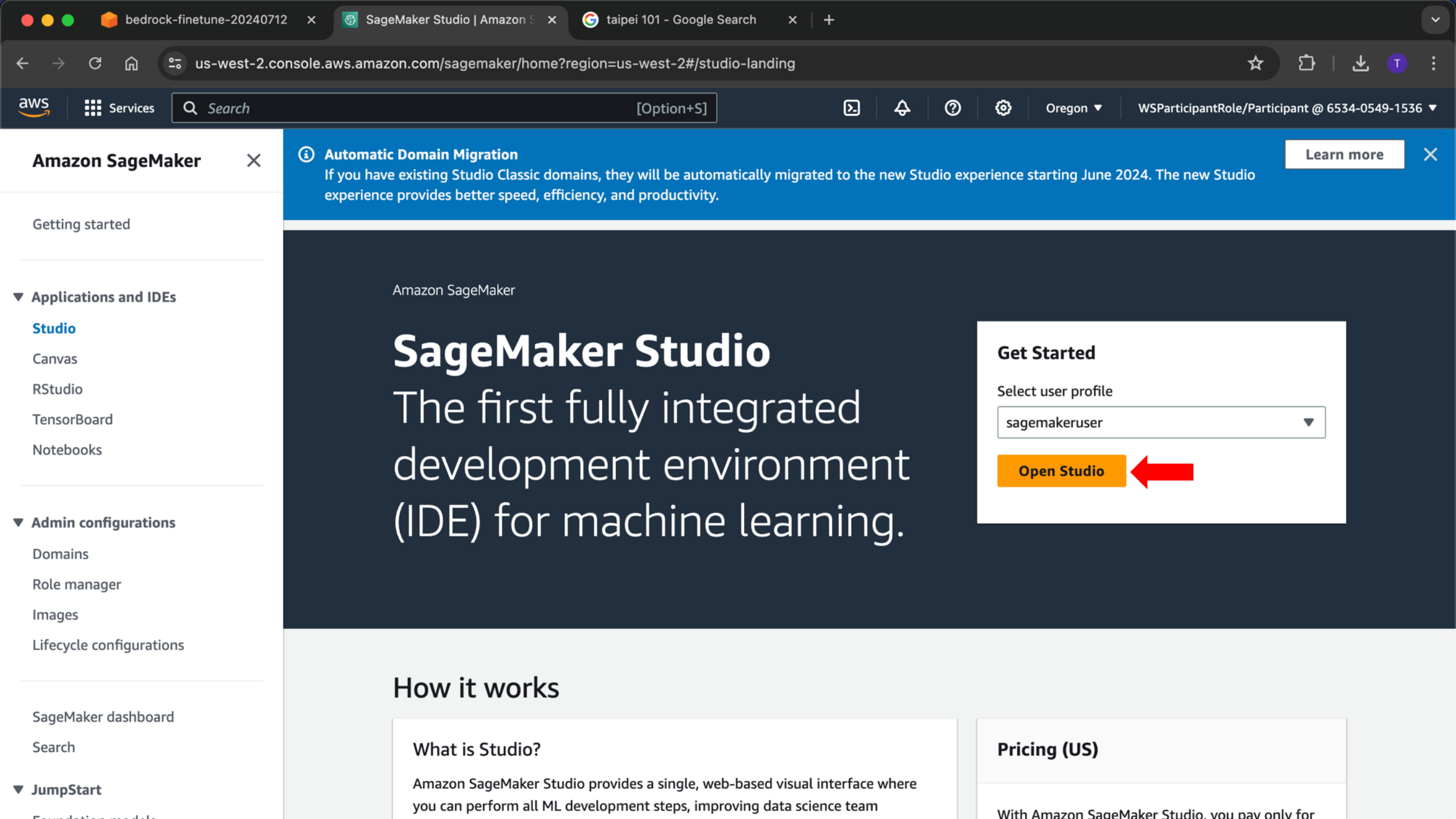
Models	Access status	Modality	EULA
AI21 Labs (2)	0/2 access granted		
Jurassic-2 Ultra	In Progress	Text	EULA
Jurassic-2 Mid	In Progress	Text	EULA

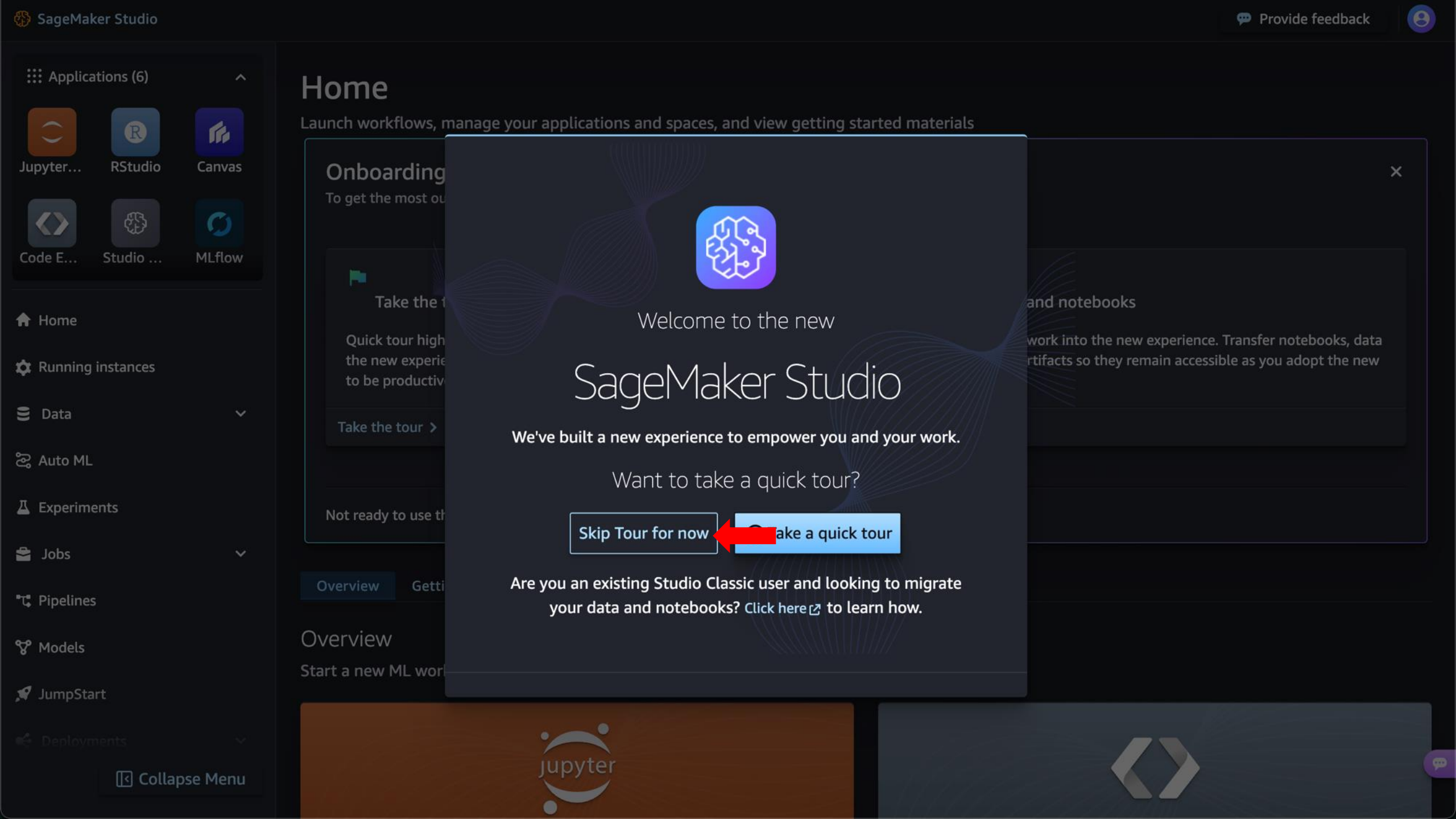
Workshop Setup

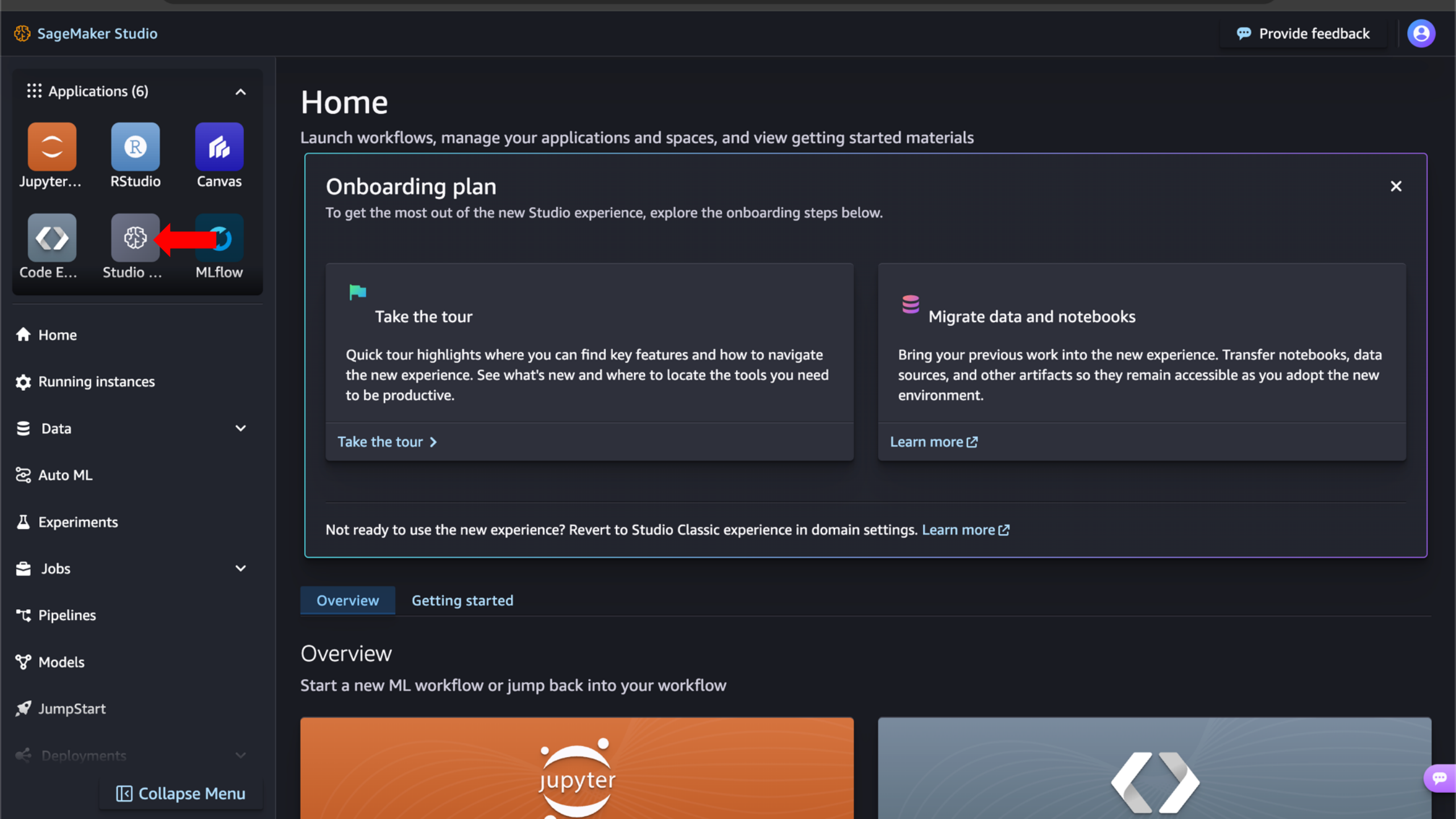
- Workshop Access
- Bedrock Model Access
- SageMaker Studio Access


aws
Services
sagemaker Canvas
Oregon
WSParticipantRole/Participant @ 6534-0549-1536
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Prompt flows
Assessment & depl
Model Evaluation
Search results for 'sagemaker'
Services (1)
Features (5)
Resources New
Documentation (11,787)
Knowledge Articles (48)
Marketplace (572)
Blogs (1,323)
Events (61)
Tutorials (23)
Services
Amazon SageMaker
Build, Train, and Deploy Machine Learning Models
Features
See all 5 results
SageMaker Studio
Amazon SageMaker feature
Notebooks
IoT Analytics feature
Autopilot
Amazon SageMaker feature
SageMaker Canvas
Amazon SageMaker feature
Resources / for a focused search
Introducing resource search
To search for resources, Resource Explorer must be active in at least one AWS Region and you must have permission to use the default view in the account. Learn more
examples
Compare mode
等,適合不同的口味喜好。
Run
Define metric criteria
Overall summary
Define metric criteria
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


bedrock-finetune-20240712
Amazon SageMaker | us-west
taipei 101 - Google Search
us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 6534-0549-1536
Amazon SageMaker
Getting started
Applications and IDEs
Studio
Canvas
RStudio
TensorBoard
Notebooks
Admin configurations
Domains
Role manager
Images
Lifecycle configurations
SageMaker dashboard
Search
JumpStart
Automatic Domain Migration
If you have existing Studio Classic domains, they will be automatically migrated to the new Studio experience starting June 2024. The new Studio experience provides better speed, efficiency, and productivity.
Learn more
Amazon SageMaker > Domains
Domains Info
In SageMaker, a domain is an environment for your team to access SageMaker resources. A domain consists of a list of authorized users and users within a domain can share notebook files and other artifacts with each other. One account can have either one or multiple domains.
Domains (1) Info
View
Create domain
Find domain name
1
Name
Id
Status
Created on
Modified on
amazon-bedrock-workshop
d-noqxdjm0rbmm
InService
Jul 12, 2024 02:28 UTC
Jul 12, 2024 02:31 UTC


bedrock-finetune-20240712
SageMaker Studio | Amazon
taipei 101 - Google Search
us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio-landing
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 6534-0549-1536
Amazon SageMaker
Getting started
Applications and IDEs
Studio
Canvas
RStudio
TensorBoard
Notebooks
Admin configurations
Domains
Role manager
Images
Lifecycle configurations
SageMaker dashboard
Search
JumpStart
Automatic Domain Migration
If you have existing Studio Classic domains, they will be automatically migrated to the new Studio experience starting June 2024. The new Studio experience provides better speed, efficiency, and productivity.
Learn more
Amazon SageMaker
SageMaker Studio
The first fully integrated development environment (IDE) for machine learning.
Get Started
Select user profile
sagemakeruser
Open Studio
How it works
What is Studio?
Amazon SageMaker Studio provides a single, web-based visual interface where you can perform all ML development steps, improving data science team
Pricing (US)
With Amazon SageMaker Studio, you pay only for


SageMaker Studio
Provide feedback
Applications (6)
Jupyter...
RStudio
Canvas
Code E...
Studio ...
MLflow
Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models
JumpStart
Deployments
Collapse Menu
Home
Launch workflows, manage your applications and spaces, and view getting started materials
Onboarding
To get the most ou
Take the t
Quick tour high
the new experie
to be productiv
Take the tour
Not ready to use th
and notebooks
work into the new experience. Transfer notebooks, data
rtifacts so they remain accessible as you adopt the new
Overview
Getti
Overview
Start a new ML wor
Welcome to the new
SageMaker Studio
We've built a new experience to empower you and your work.
Want to take a quick tour?
Skip Tour for now
ake a quick tour
Are you an existing Studio Classic user and looking to migrate your data and notebooks? Click here to learn how.
jupyter



Home

Launch workflows, manage your applications and spaces, and view getting started materials

Onboarding plan

To get the most out of the new Studio experience, explore the onboarding steps below.

Take the tour

Quick tour highlights where you can find key features and how to navigate the new experience. See what's new and where to locate the tools you need to be productive.

Take the tour >

Migrate data and notebooks

Bring your previous work into the new experience. Transfer notebooks, data sources, and other artifacts so they remain accessible as you adopt the new environment.

Learn more

Not ready to use the new experience? Revert to Studio Classic experience in domain settings. Learn more

Overview | Getting started

Overview

Start a new ML workflow or jump back into your workflow

Applications (6)
JupyterLab
RStudio
Canvas
Code Editor
Studio Cl...
MLflow

Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models
JumpStart
Deployments
Collapse Menu

SageMaker Studio Classic

+ Create Studio Classic space

About

Studio Classic is a legacy IDE that allows you to access the previous iteration of SageMaker Studio from within the new Studio experience.

See features | Learn more about Studio Classic

Search...

Filter spaces: Running

Name	Application	Status	Type	Last modified	Action
sagemakeruser	Studio Classic	Stopped	Private	0 seconds ago	Run

1 results Results are cached Refresh Go to page 1 Page 1 of 1

Introducing spaces New

JupyterLab and Code Editor now come with durable instances that allow for faster startup, privacy options, and configurable storage.

Learn more

Applications (6)
Jupyter...
RStudio
Canvas
Code E...
Studio ...
MLflow
Studio Classic
Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models
JumpStart
Deployments
Collapse Menu

SageMaker Studio Classic

+ Create Studio Classic space

About

Studio Classic is a legacy IDE that allows you to access the previous iteration of SageMaker Studio from within the new Studio experience.

See features | Learn more about Studio Classic

Search...

Filter spaces: Running

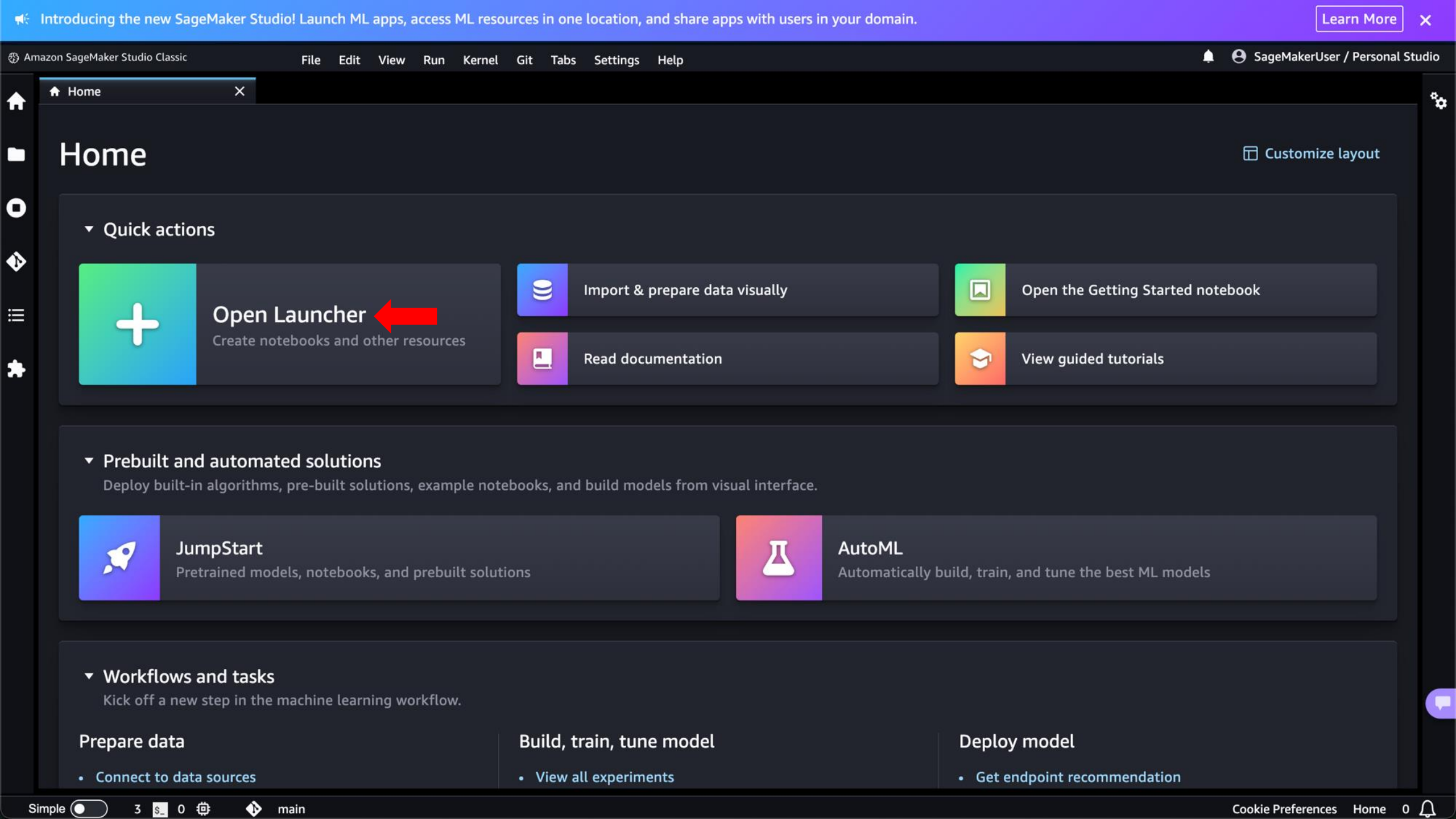
Name	Application	Status	Type	Last modified	Action
sagemakeruser	Studio Classic	Running	Private	15 seconds ago	Stop Open

1 results Results are cached Refresh Go to page 1 Page 1 of 1

Introducing spaces New

JupyterLab and Code Editor now come with durable instances that allow for faster startup, privacy options, and configurable storage.

Learn more


Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain.
Learn More
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
SageMakerUser / Personal Studio
Home
Home
Customize layout
Quick actions
Open Launcher
Create notebooks and other resources
Import & prepare data visually
Open the Getting Started notebook
Read documentation
View guided tutorials
Prebuilt and automated solutions
Deploy built-in algorithms, pre-built solutions, example notebooks, and build models from visual interface.
JumpStart
Pretrained models, notebooks, and prebuilt solutions
AutoML
Automatically build, train, and tune the best ML models
Workflows and tasks
Kick off a new step in the machine learning workflow.
Prepare data
Connect to data sources
Build, train, tune model
View all experiments
Deploy model
Get endpoint recommendation
Simple
3
0
main
Cookie Preferences
Home
0

Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain. Learn More

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help SageMakerUser / Personal Studio

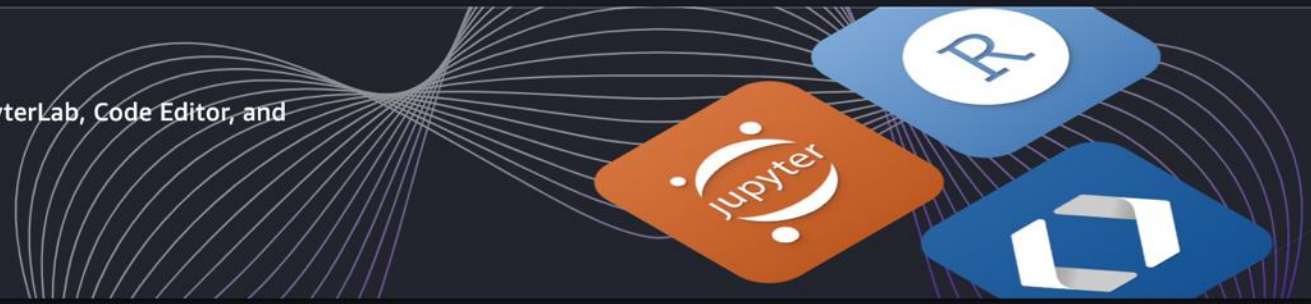
Home Launcher

Launcher

Interested in launching different applications?

Introducing a centralized hub for launching all your favorite apps including JupyterLab, Code Editor, and RStudio.

Learn more



Notebooks and compute resources

/

Create notebooks, code console, image terminal with custom environment in the active folder.

Image	Kernel	Instance	Start-up script	
Data Science 3.0	Python 3	ml.t3.medium	No script	Change environment

Create notebook

Open code console

Open image terminal

Learn more about SageMaker images and how to customize compute environment

Utilities and files

Simple 1 0 Cookie Preferences Launcher 0

Notebooks and compute resources

Create notebooks, code console, image terminal with custom environment in the active folder.

/

Image	Kernel	Instance	Start-up script	
Data Science 3.0	Python 3	ml.t3.medium	No script	Change environment

Create notebook

Open code console

Open image terminal

Learn more about SageMaker images and how to customize compute environment

Utilities and files

System terminal

Text file

Markdown file

Python file

Notebook jobs

Contextual help

git clone https://github.com/aws-samples/amazon-bedrock-workshop

d-ebucdpoajiiy.studio.us-west-2.sagemaker.aws/jupyter/default/lab

Paused

Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain. Learn More

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

SageMakerUser / Personal Studio

Filter files by name

/

Name

Home | Launcher | Terminal 3

```
!!!!!!!! Welcome to SageMaker Studio System Terminal !!!!!!!!

Below are some useful tips:

* Activate studio conda environment using "conda activate studio" to install extensions, list extension etc. For more information, see https://docs.aws.amazon.com/sagemaker/latest/dg/studio-jl.html#studio-jl-install

* Post JupyterServer extension installation, if needed, restart just the server(not app) using "restart-jupyter-server"

sagemaker-user@studio$ git clone https://github.com/aws-samples/amazon-bedrock-workshop
```

d-ebucdpoajiiy.studio.us-west-2.sagemaker.aws/jupyter/default/lab

Paused

Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain. Learn More

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

SageMakerUser / Personal Studio

Filter files by name

/

Name

amazon-bedrock-wor...

Home | Launcher | Terminal 3

```
!!!!!!!! Welcome to SageMaker Studio System Terminal !!!!!!!!

Below are some useful tips:

* Activate studio conda environment using "conda activate studio" to install extensions, list extension etc. For more information, see https://docs.aws.amazon.com/sagemaker/latest/dg/studio-jl.html#studio-jl-install

* Post JupyterServer extension installation, if needed, restart just the server(not app) using "restart-jupyter-server"


sagemaker-user@studio$ git clone https://github.com/aws-samples/amazon-bedrock-workshop
Cloning into 'amazon-bedrock-workshop'...
remote: Enumerating objects: 2761, done.
remote: Counting objects: 100% (1835/1835), done.
remote: Compressing objects: 100% (510/510), done.
remote: Total 2761 (delta 1492), reused 1496 (delta 1321), pack-reused 926
Receiving objects: 100% (2761/2761), 28.97 MiB | 23.18 MiB/s, done.
Resolving deltas: 100% (1850/1850), done.
Updating files: 100% (161/161), done.
sagemaker-user@studio$
```

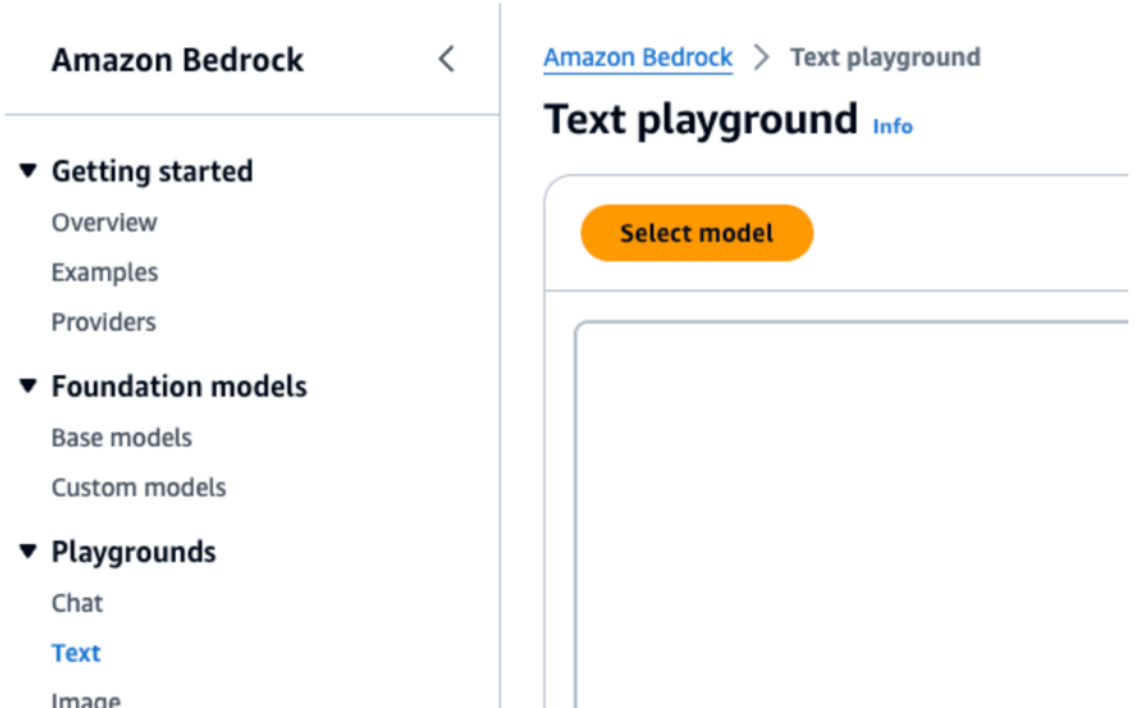
https://reurl.cc/8Xpvq4

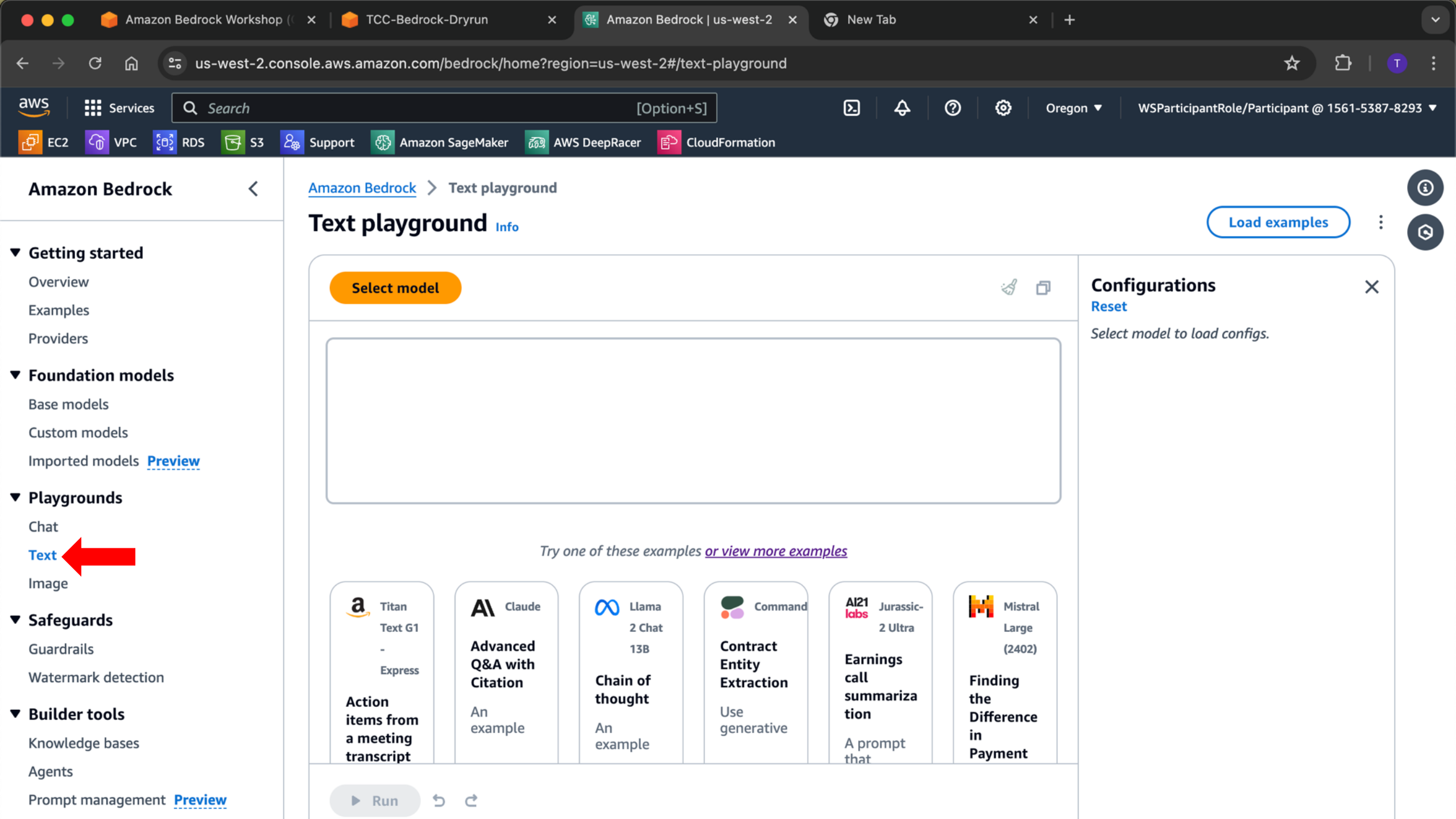
翻譯

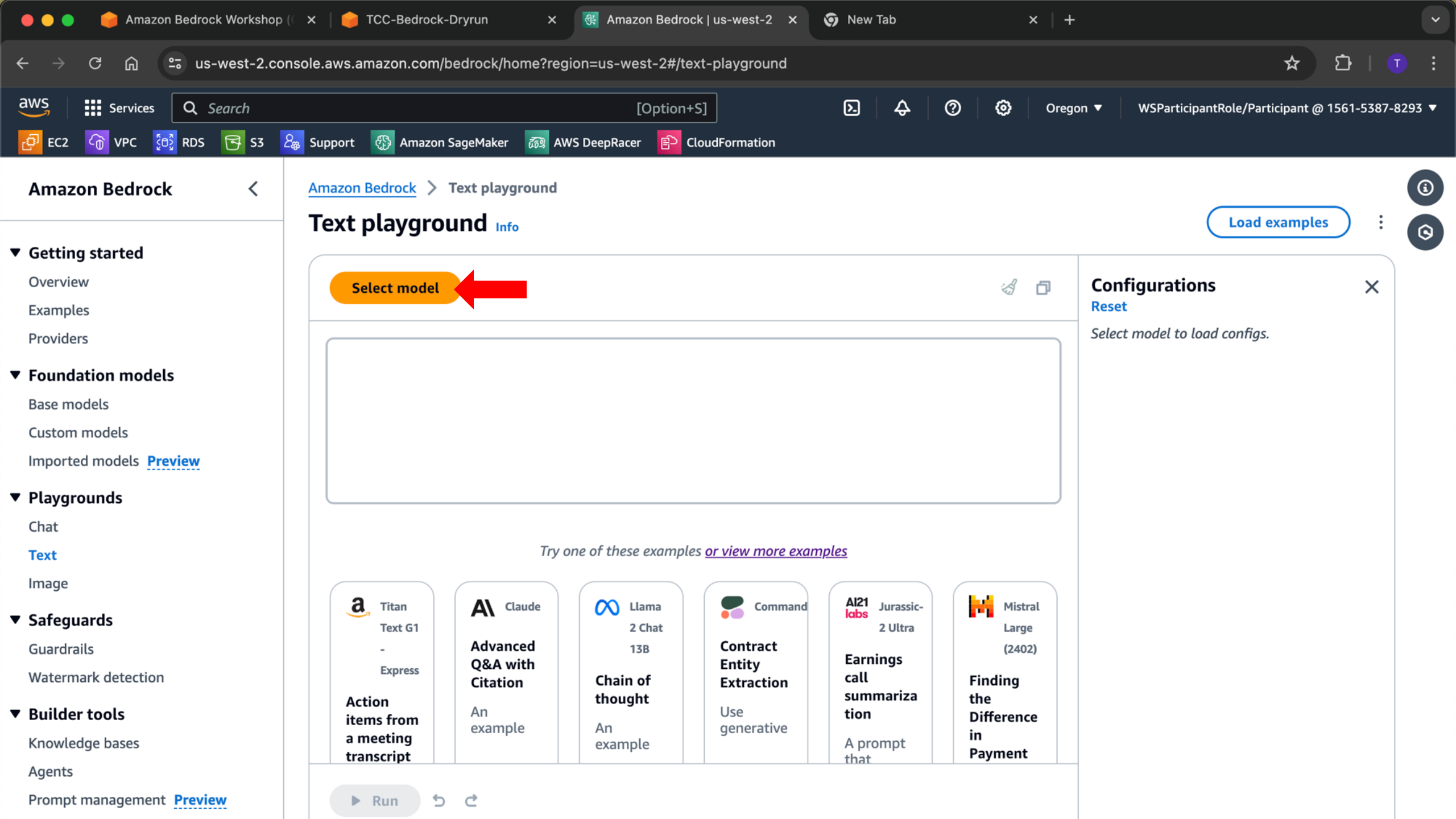
下列實驗演示在 Amazon Bedrock playground 上使用 Claude 3 進行文字翻譯：

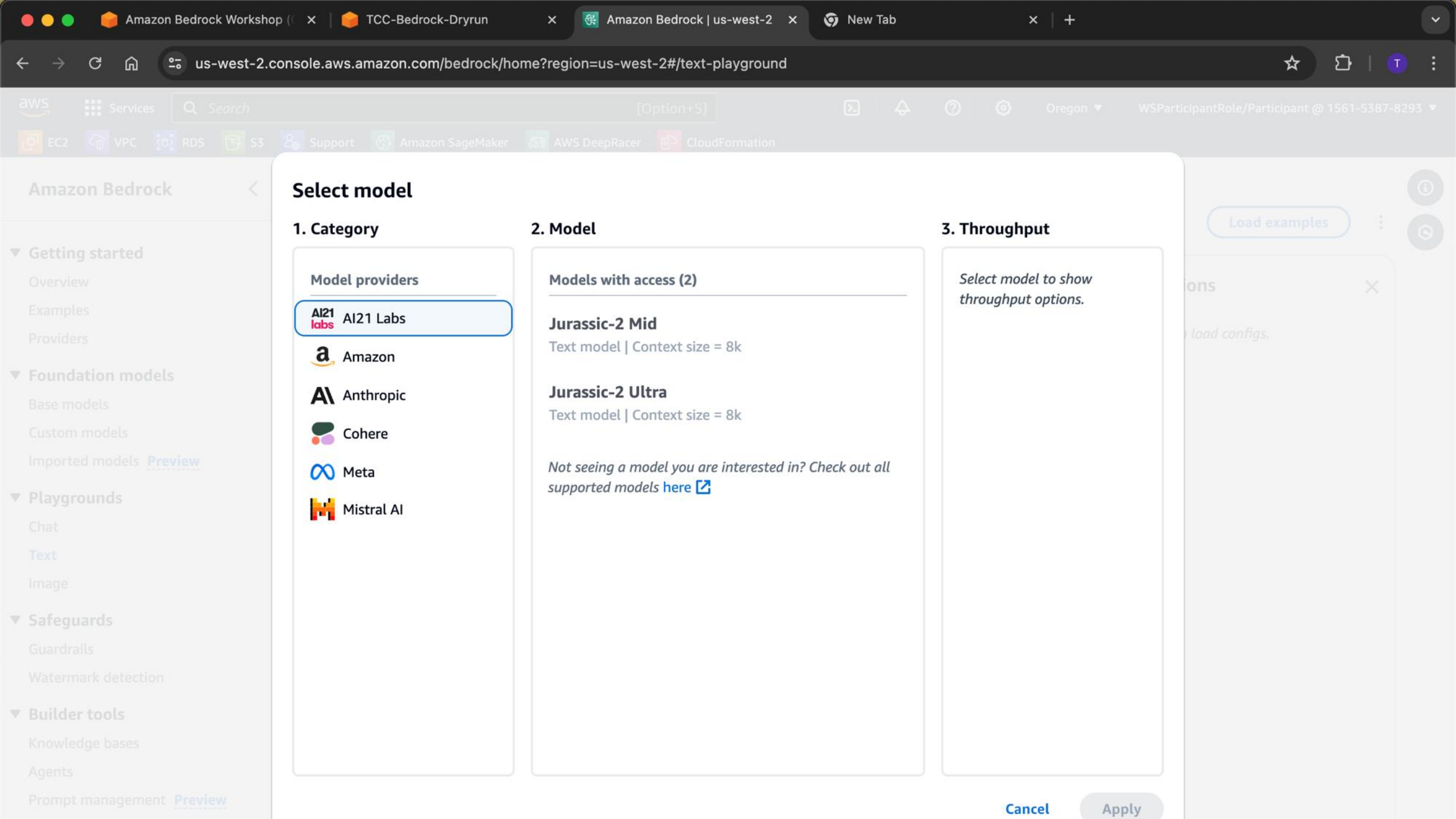
控制台及模型選擇

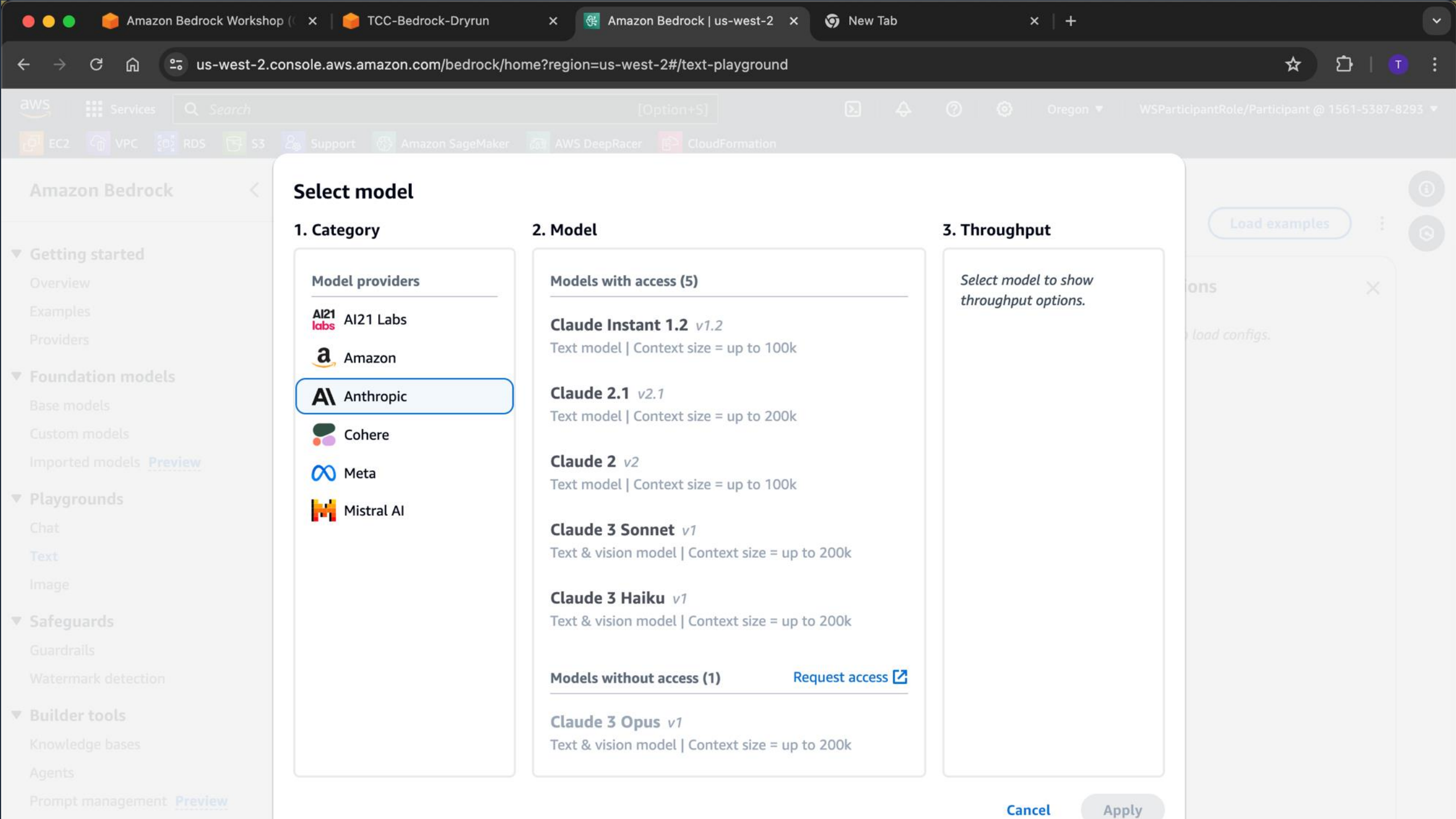
- 打開 Amazon Bedrock 控制台，左側選單選擇 **Playgrounds -> Text**

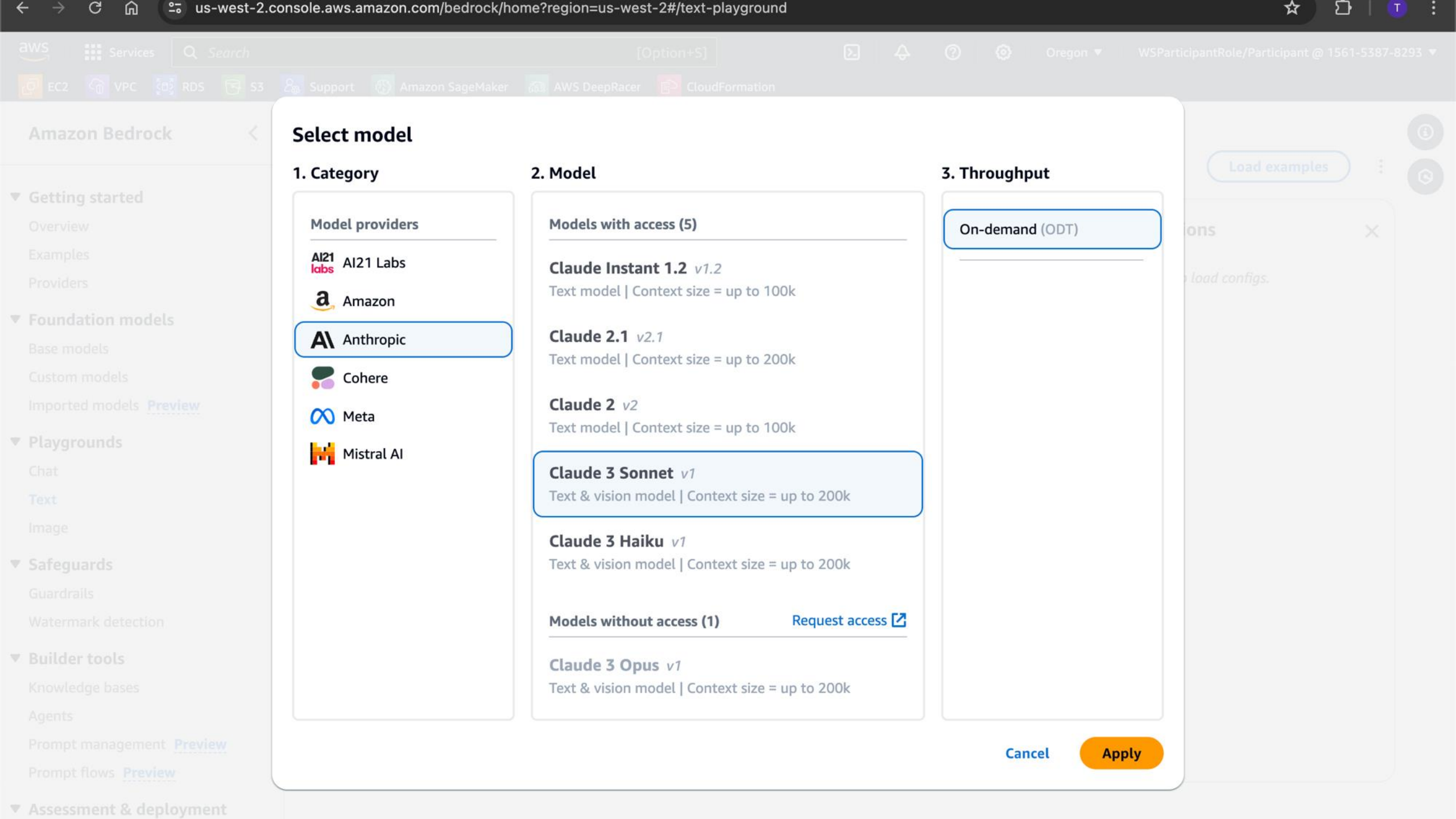


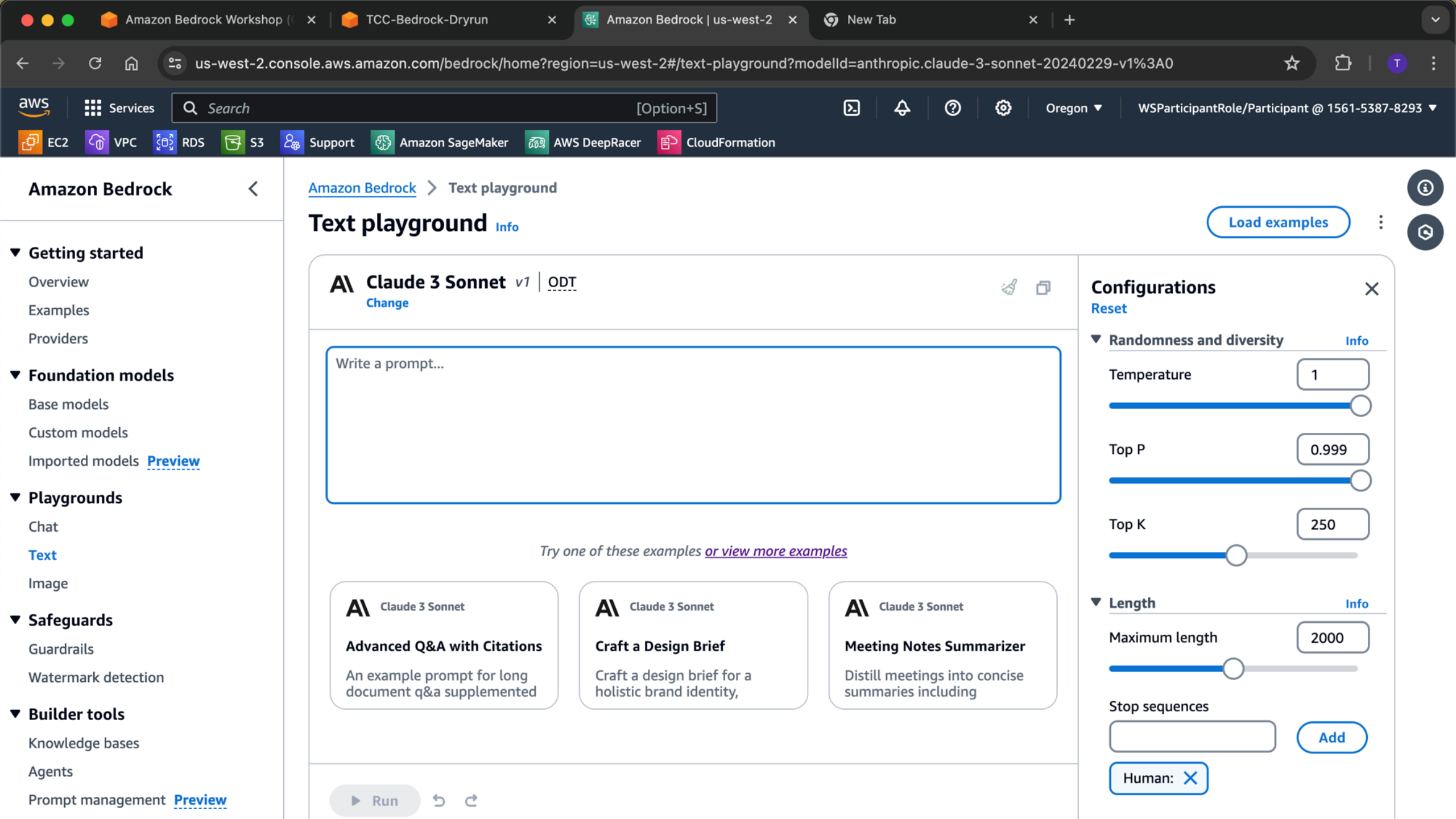

Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-playground
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Amazon Bedrock
Text playground
Text playground
Info
Load examples
Select model
Configurations
Reset
Select model to load configs.
Try one of these examples or view more examples
Titan Text G1 - Express
Action items from a meeting transcript
Claude
Advanced Q&A with Citation
An example
Llama 2 Chat 13B
Chain of thought
An example
Command
Contract Entity Extraction
Use generative
AI21 labs
Jurassic-2 Ultra
Earnings call summarization
A prompt that
Mistral Large (2402)
Finding the Difference in Payment
Run


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-playground
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Amazon Bedrock
Text playground
Text playground
Info
Load examples
Select model
Configurations
Reset
Select model to load configs.
Try one of these examples or view more examples
Titan Text G1 - Express
Action items from a meeting transcript
Claude
Advanced Q&A with Citation
An example
Llama 2 Chat 13B
Chain of thought
An example
Command
Contract Entity Extraction
Use generative
AI21 labs
Jurassic-2 Ultra
Earnings call summarization
A prompt that
Mistral Large (2402)
Finding the Difference in Payment
Run


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-playground
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Load examples
Select model
1. Category
Model providers
AI21 Labs
Amazon
Anthropic
Cohere
Meta
Mistral AI
2. Model
Models with access (2)
Jurassic-2 Mid
Text model | Context size = 8k
Jurassic-2 Ultra
Text model | Context size = 8k
Not seeing a model you are interested in? Check out all supported models here
3. Throughput
Select model to show throughput options.
Cancel
Apply


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-playground
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Load examples
Select model
1. Category
Model providers
AI21 Labs
Amazon
Anthropic
Cohere
Meta
Mistral AI
2. Model
Models with access (5)
Claude Instant 1.2 v1.2
Text model | Context size = up to 100k
Claude 2.1 v2.1
Text model | Context size = up to 200k
Claude 2 v2
Text model | Context size = up to 100k
Claude 3 Sonnet v1
Text & vision model | Context size = up to 200k
Claude 3 Haiku v1
Text & vision model | Context size = up to 200k
Models without access (1)
Request access
Claude 3 Opus v1
Text & vision model | Context size = up to 200k
3. Throughput
Select model to show throughput options.
Cancel
Apply


us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-playground
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Prompt flows
Preview
Assessment & deployment
Load examples
Select model
1. Category
Model providers
AI21 Labs
Amazon
Anthropic
Cohere
Meta
Mistral AI
2. Model
Models with access (5)
Claude Instant 1.2 v1.2
Text model | Context size = up to 100k
Claude 2.1 v2.1
Text model | Context size = up to 200k
Claude 2 v2
Text model | Context size = up to 100k
Claude 3 Sonnet v1
Text & vision model | Context size = up to 200k
Claude 3 Haiku v1
Text & vision model | Context size = up to 200k
Models without access (1)
Request access
Claude 3 Opus v1
Text & vision model | Context size = up to 200k
3. Throughput
On-demand (ODT)
Cancel
Apply


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-playground?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Amazon Bedrock
Text playground
Text playground
Info
Load examples
Claude 3 Sonnet
v1
ODT
Change
Write a prompt...
Try one of these examples
or view more examples
Claude 3 Sonnet
Advanced Q&A with Citations
An example prompt for long document q&a supplemented
Claude 3 Sonnet
Craft a Design Brief
Craft a design brief for a holistic brand identity,
Claude 3 Sonnet
Meeting Notes Summarizer
Distill meetings into concise summaries including
Run
Configurations
Reset
Randomness and diversity
Info
Temperature
1
Top P
0.999
Top K
250
Length
Info
Maximum length
2000
Stop sequences
Add
Human:

Content preferences

Language

中文(繁體)

- 我們以一個字幕翻譯情境為例，在 playground 中輸入以下提示詞 prompt:

Copied!

```
你是一個幫助翻譯劇本的助理。
你的任務是將<text>中的英文原文翻譯成繁體中文。翻譯時，請遵守以下規則:
0.不要改變原意。
1.翻譯前先瞭解上下文，保持語義連貫、閱讀流暢，但不要故意誇張。
2.原文大多是對話式的，因此翻譯仍應符合短影音/影音部落格/Youtube影片的上下文環境。注意避免使用平常不會在日常聊天中出現的字詞。
3.適當的時候保留一些專有名詞或專業術語未翻譯，注意前後一致性。
4.標點符號必須使用全形。例如，不可以使用小寫的 ","，必須使用 "，"。
5.在<result></result> 中回覆翻譯。不要包含任何額外的內容。
<example>
H: Welcome to the Amazon Bedrock Workshop, this workshop will help you quickly get started on your journey to generative AI application
A: 歡迎來到 Amazon Bedrock Workshop，透過此 Workshop 將協助您快速開始生成式 AI 應用程式的旅程。
</example>
<text>
"We'll cover all of those things in a moment, but before we get started, this video doesn't have a sponsor, but it is supported by the
</text>
```

你是一個幫助翻譯劇本的助理。
你的任務是將<text>中的英文原文翻譯成繁體中文。翻譯時，請遵守以下規則:
0.不要改變原意。
1.翻譯前先瞭解上下文，保持語義連貫、閱讀流暢，但不要故意誇張。
2.原文大多是對話式的，因此翻譯仍應符合短影音/影音部落格/Youtube影片的上下文環境。注意避免使用平常不會在日常聊天中出現的字詞。
3.適當的時候保留一些專有名詞或專業術語未翻譯，注意前後一致性。
4.標點符號必須使用全形。例如，不可以使用小寫的 ","，必須使用 "，"。
5.在<result></result> 中回覆翻譯。不要包含任何額外的內容。
<example>
H: Welcome to the Amazon Bedrock Workshop, this workshop will help you quickly get started on your journey to generative AI applications.
A: 歡迎來到 Amazon Bedrock Workshop，透過此 Workshop 將協助您快速開始生成式 AI 應用程式的旅程。
</example>
<text>
"We'll cover all of those things in a moment, but before we get started, this video doesn't have a sponsor, but it is supported by the thousands of you wonderful people who get value out of all of my courses, prints, presets and ebooks over at patk.com."
</text>

▼ Randomness and diversity Info

Temperature 1

Top P 0.999

Top K 250

▼ Length Info

Maximum length 2000

Stop sequences

Amazon Bedrock

Getting started
Overview
Examples
Providers

Foundation models
Base models
Custom models
Imported models Preview

Playgrounds
Chat
Text
Image

Safeguards
Guardrails
Watermark detection

Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview

Text playground Info

Load examples

Claude 3 Sonnet v1 | ODT

Change

你是一個幫助翻譯劇本的助理。
你的任務是將<text>中的英文原文翻譯成繁體中文。翻譯時，請遵守以下規則:
0.不要改變原意。
1.翻譯前先瞭解上下文，保持語義連貫、閱讀流暢，但不要故意誇張。
2.原文大多是對話式的，因此翻譯仍應符合短影音/影音部落格/Youtube影片的上下文環境。注意避免使用平常不會在日常聊天中出現的字詞。
3.適當的時候保留一些專有名詞或專業術語未翻譯，注意前後一致性。
4.標點符號必須使用全形。例如，不可以使用小寫的 "," ，必須使用 " ，" 。
5.在<result></result> 中回覆翻譯。不要包含任何額外的內容。
<example>
H: Welcome to the Amazon Bedrock Workshop, this workshop will help you quickly get started on your journey to generative AI applications.
A: 歡迎來到 Amazon Bedrock Workshop，透過此 Workshop 將協助您快速開始生成式 AI 應用程式的旅程。
</example>
<text>
"We'll cover all of those things in a moment, but before we get started, this video doesn't have a sponsor, but it is supported by the thousands of you wonderful people who get value out of all of my courses, prints, presets and ebooks over at patk.com."
</text>

Try one of these examples or view more examples

Configurations

Reset

Randomness and diversity Info

Temperature 1

Top P 0.999

Top K 250

Length Info

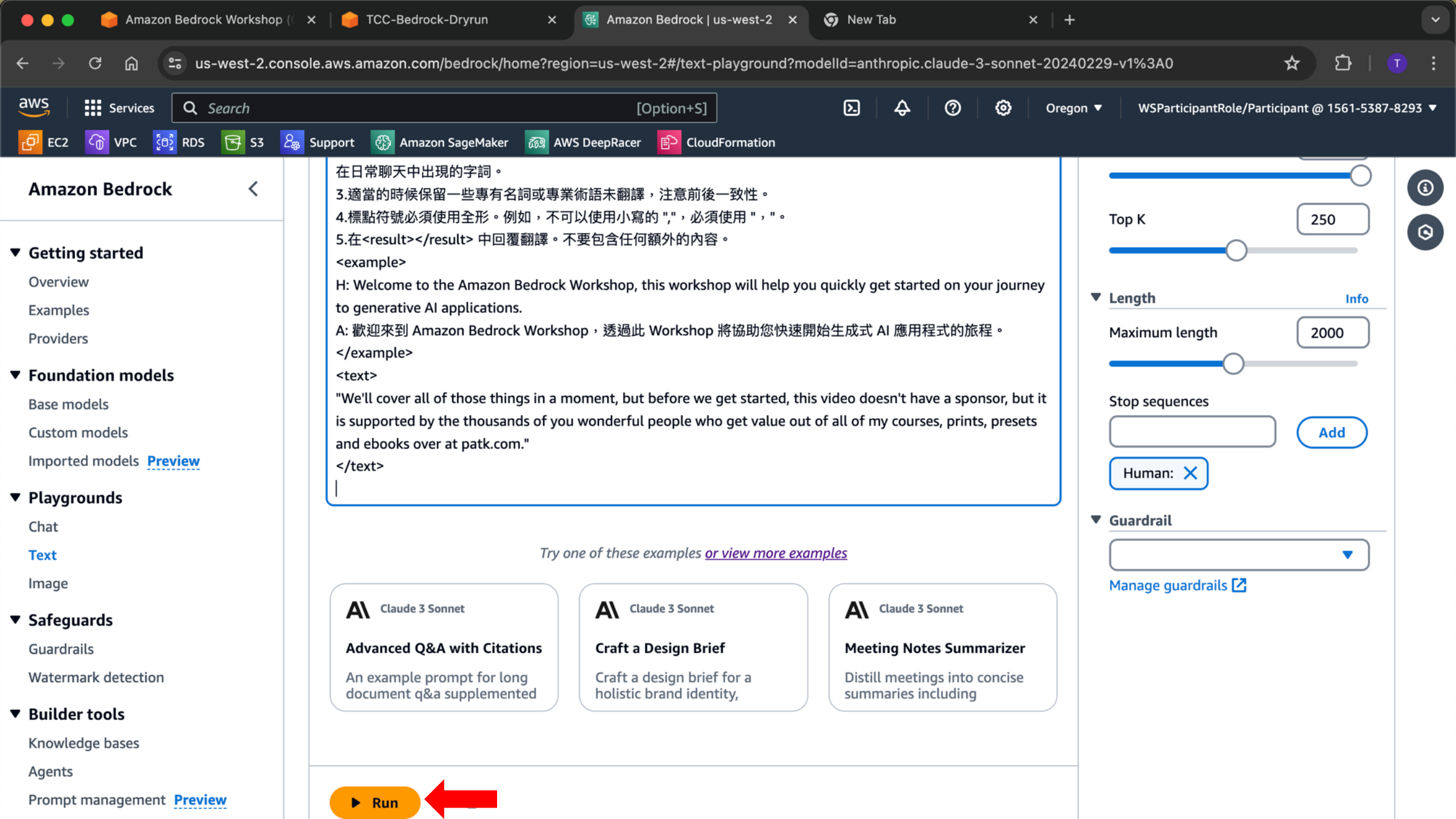
Maximum length 2000

Stop sequences

Add

Human:

Guardrail


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
New Tab
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-playground?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
在日常聊天中出現的字詞。
3.適當的時候保留一些專有名詞或專業術語未翻譯，注意前後一致性。
4.標點符號必須使用全形。例如，不可以使用小寫的 ","，必須使用 "，"。
5.在<result></result> 中回覆翻譯。不要包含任何額外的內容。
<example>
H: Welcome to the Amazon Bedrock Workshop, this workshop will help you quickly get started on your journey to generative AI applications.
A: 歡迎來到 Amazon Bedrock Workshop，透過此 Workshop 將協助您快速開始生成式 AI 應用程式的旅程。
</example>
<text>
"We'll cover all of those things in a moment, but before we get started, this video doesn't have a sponsor, but it is supported by the thousands of you wonderful people who get value out of all of my courses, prints, presets and ebooks over at patk.com."
</text>
Try one of these examples or view more examples
Claude 3 Sonnet
Advanced Q&A with Citations
An example prompt for long document q&a supplemented
Claude 3 Sonnet
Craft a Design Brief
Craft a design brief for a holistic brand identity,
Claude 3 Sonnet
Meeting Notes Summarizer
Distill meetings into concise summaries including
Run
Top K
250
Length
Info
Maximum length
2000
Stop sequences
Add
Human:
Guardrail
Manage guardrails

Amazon Bedrock

Getting started
Overview
Examples
Providers

Foundation models
Base models
Custom models
Imported models Preview

Playgrounds
Chat
Text
Image

Safeguards
Guardrails
Watermark detection

Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview

Assessment & deployment

Text playground Info

Load examples

Claude 3 Sonnet *v1* | ODT
Change

在日常聊天中出現的字詞。
3.適當的時候保留一些專有名詞或專業術語未翻譯，注意前後一致性。
4.標點符號必須使用全形。例如，不可以使用小寫的 ","，必須使用 "，"。
5.在<result></result> 中回覆翻譯。不要包含任何額外的內容。
<example>
H: Welcome to the Amazon Bedrock Workshop, this workshop will help you quickly get started on your journey to generative AI applications.
A: 歡迎來到 Amazon Bedrock Workshop，透過此 Workshop 將協助您快速開始生成式 AI 應用程式的旅程。
</example>
<text>
"We'll cover all of those things in a moment, but before we get started, this video doesn't have a sponsor, but it is supported by the thousands of you wonderful people who get value out of all of my courses, prints, presets and ebooks over at patk.com."
</text>

<result>「我們稍後會涵蓋所有這些內容，但在開始之前，這段影片並沒有贊助商，不過它獲得來自成千上萬位從我在 patk.com 上的各種課程、美工圖案、預設和電子書中獲益良多的出色觀眾支持。」</result>

Run

Configurations
Reset

Randomness and diversity Info

Temperature 1

Top P 0.999

Top K 250

Length Info

Maximum length 2000

Stop sequences

Add

Human:

Guardrail

▼ Content preferences

Language

中文(繁體)

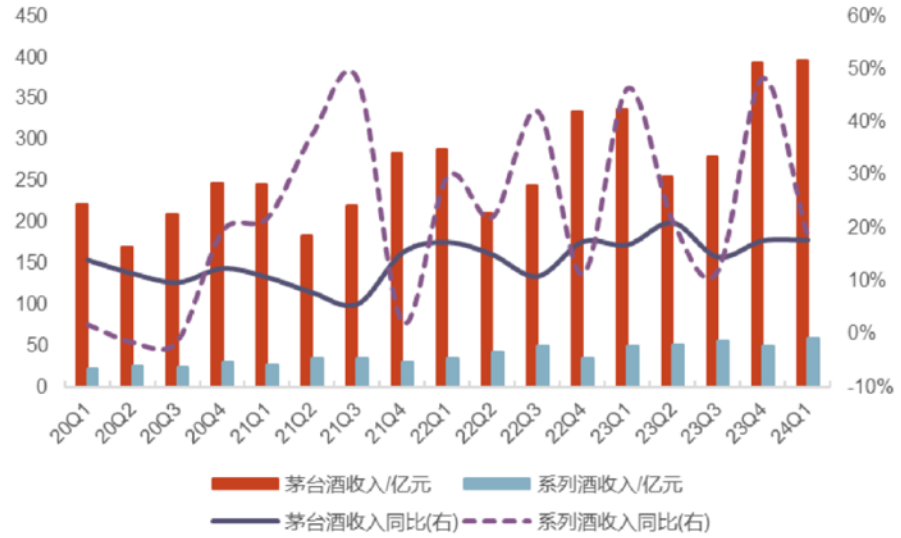
演示1: 图表理解

System Prompt

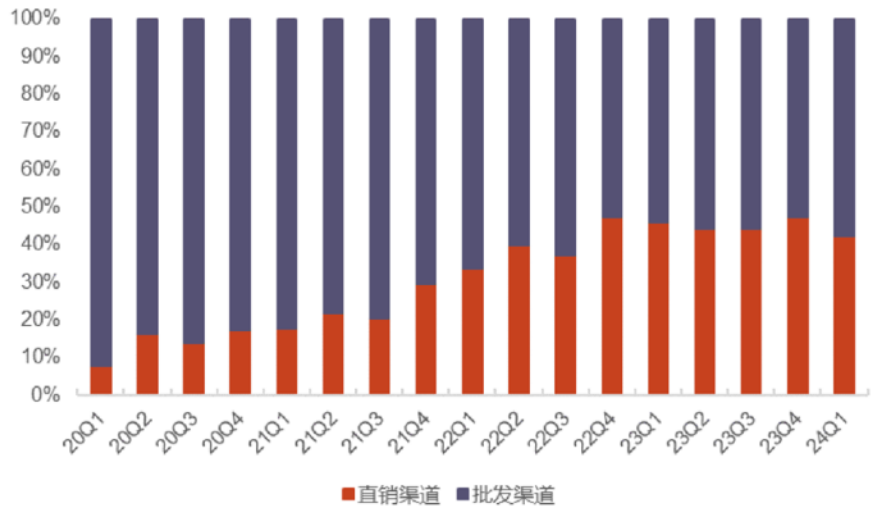
```
1  你是一个分析图表的专家，特别是图表里的柱状图，折线图，或者混合折线和柱状图在一起的图表。
2  请首先给出趋势或概览分析，之后仔细确认每一个折线的每一个波峰和每一个波谷对应的横轴和纵轴的数字（纵轴可能有两个）并列出。使用中文回复。
```

用户输入

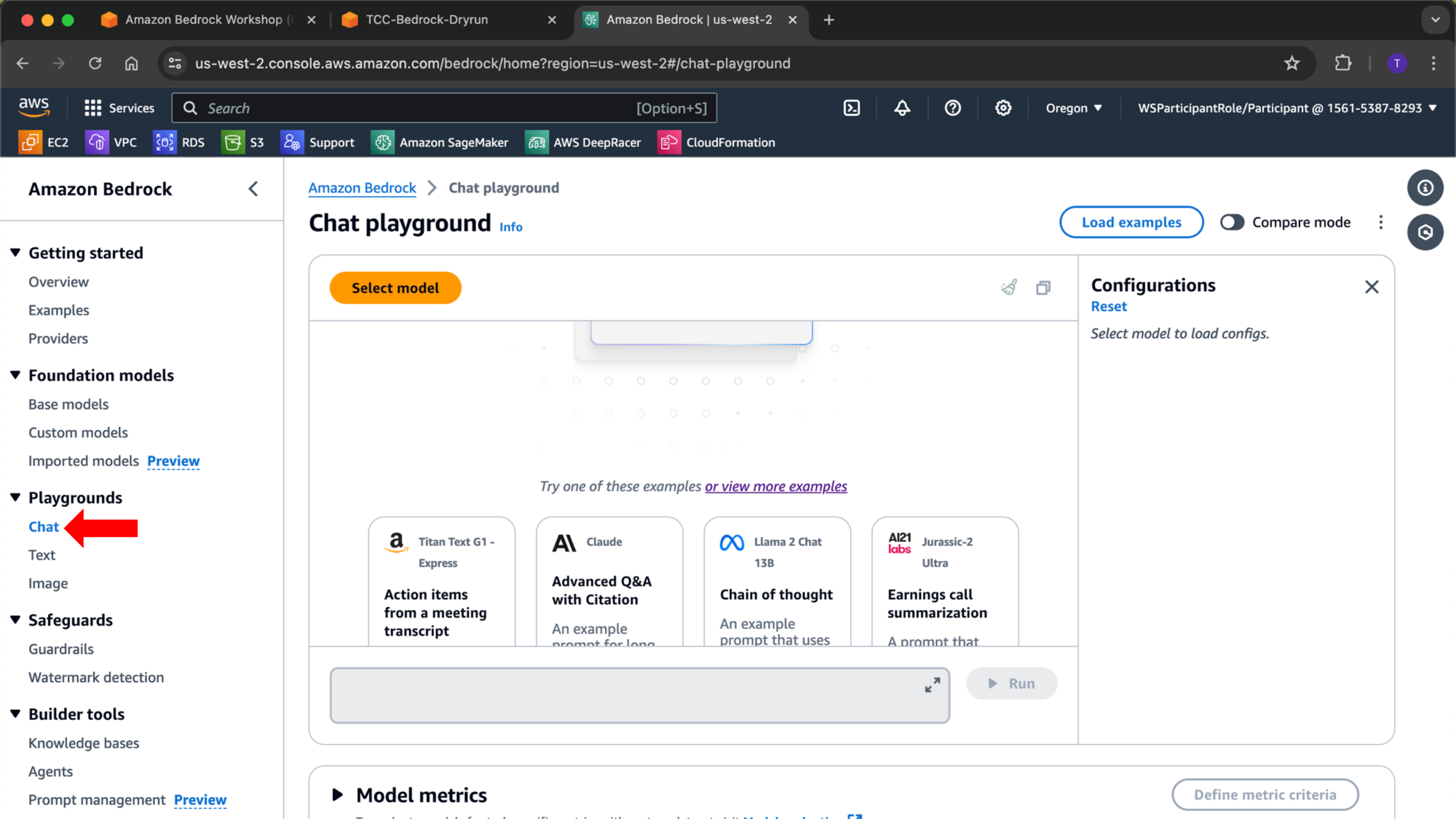
图片1

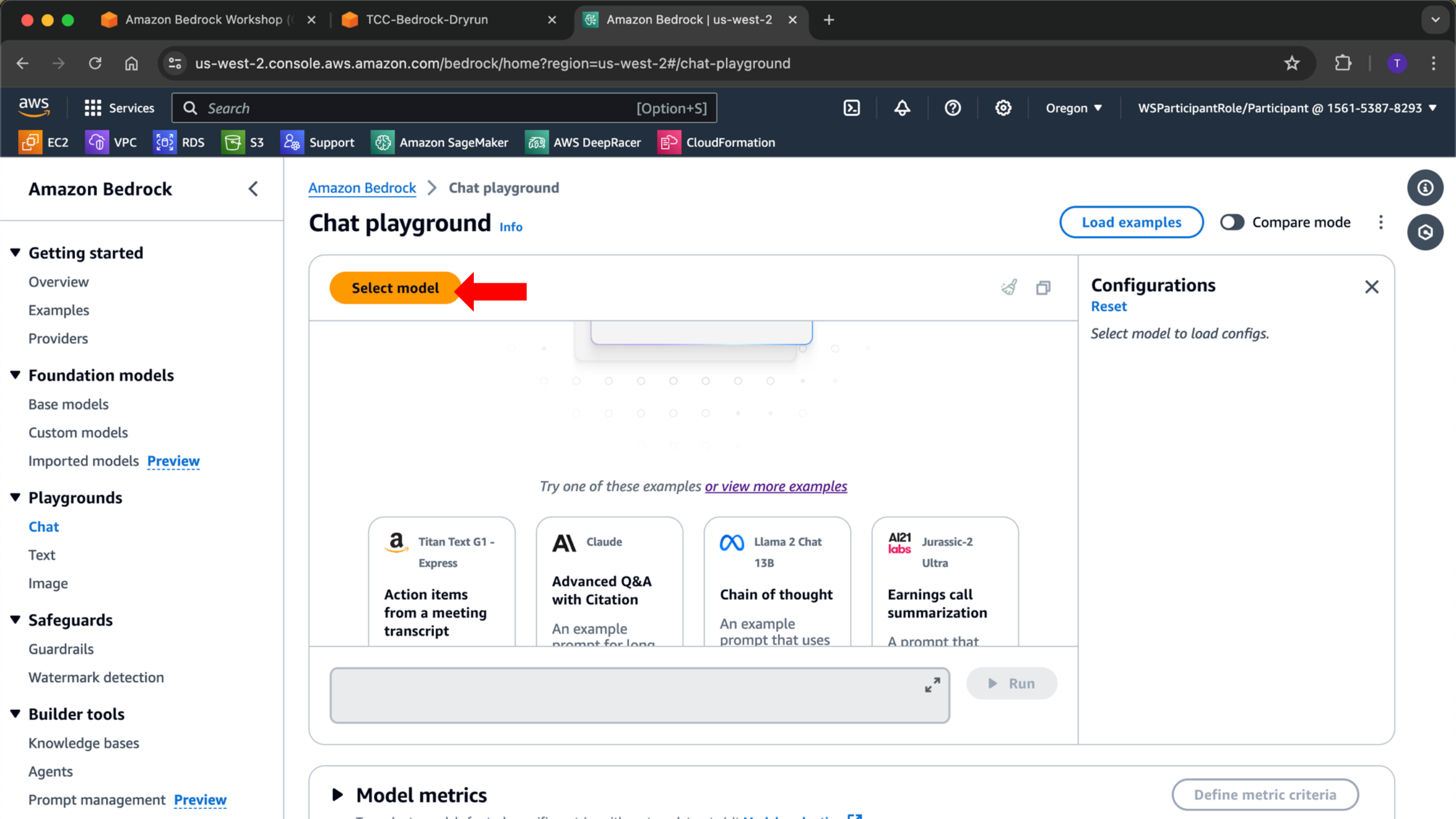


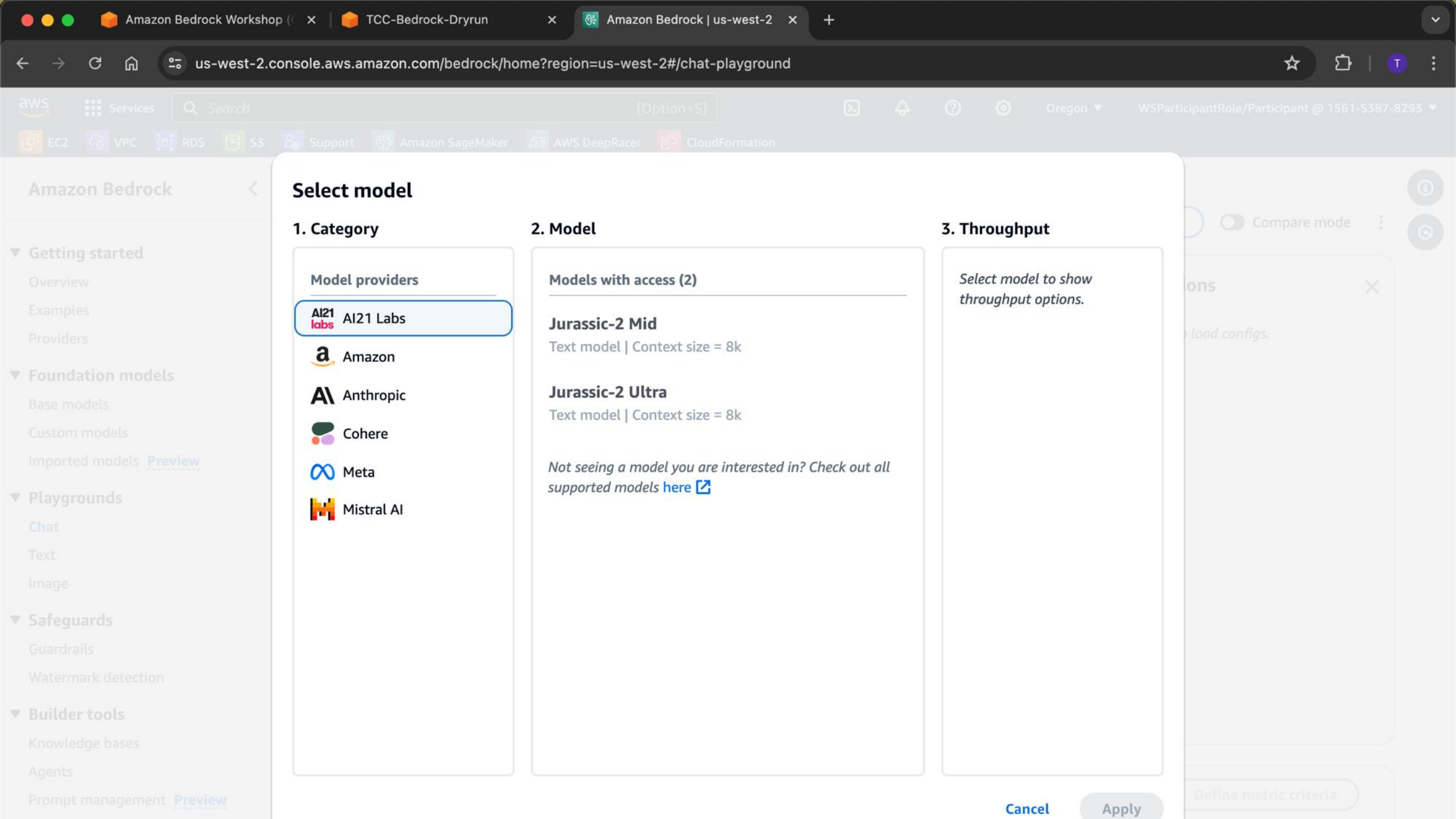
图片2

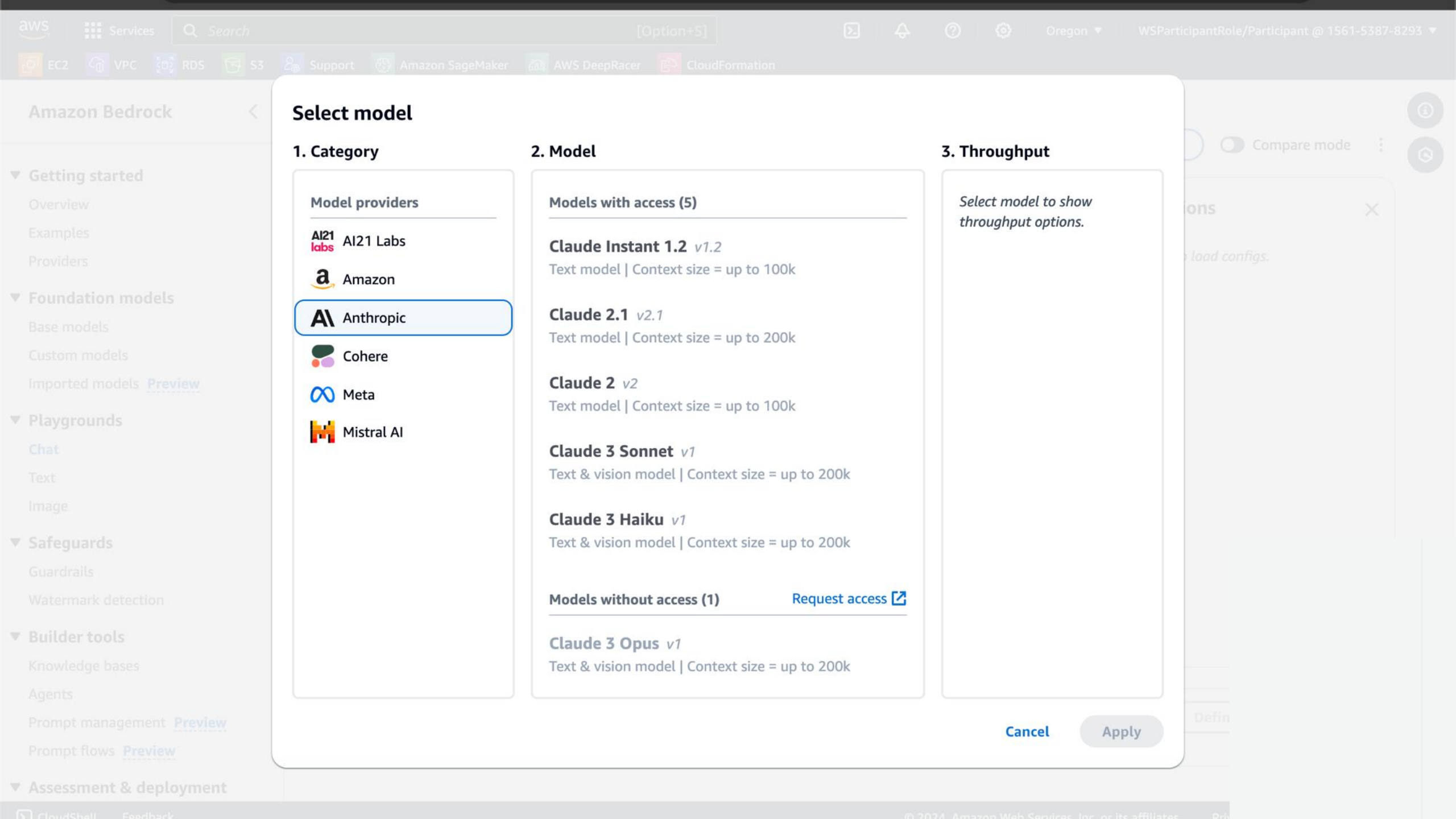


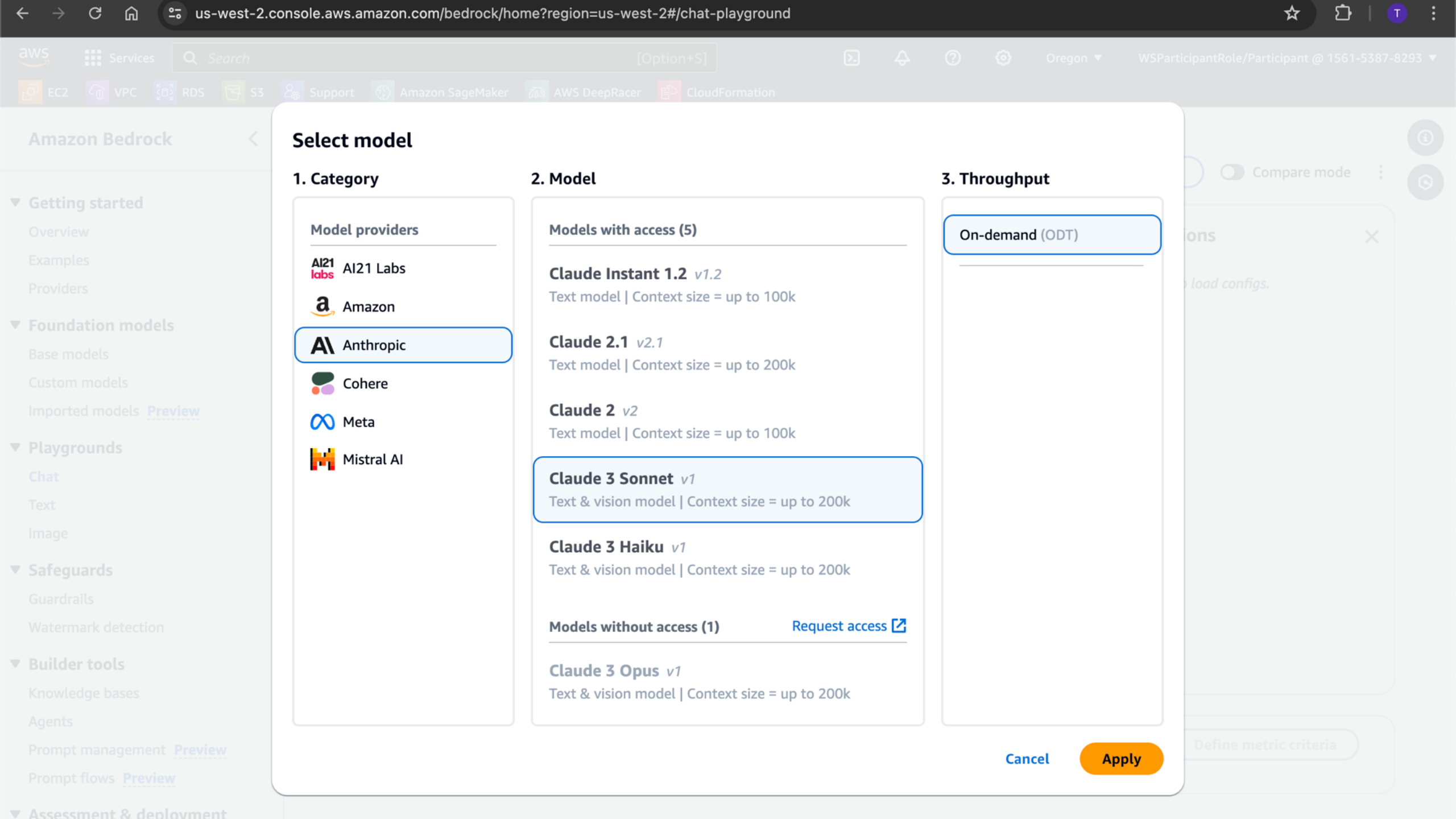
这两张图是茅台公司酒产品20Q1-24Q1的统计图，请仔细分析并给出你的分析结果：

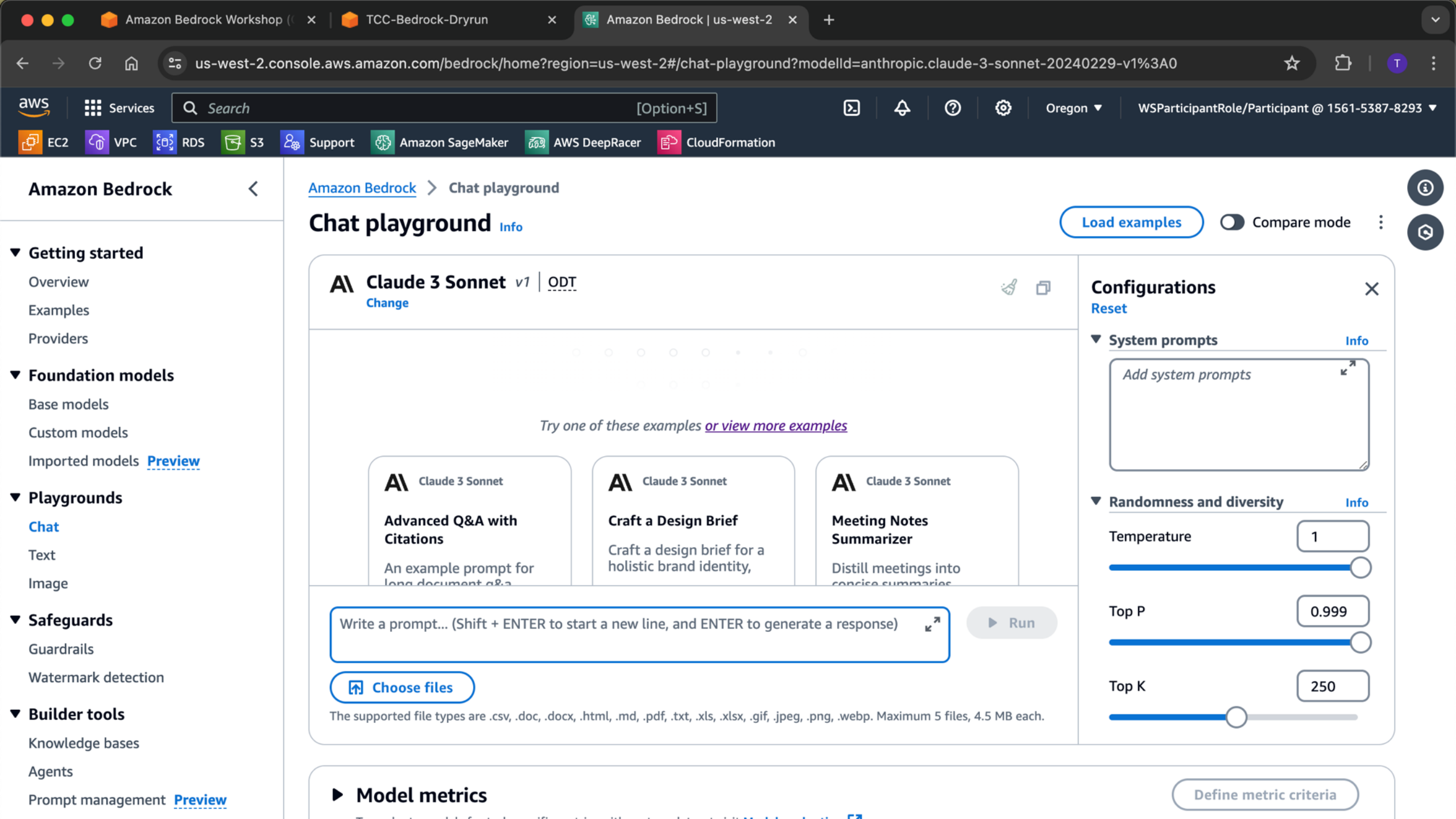

Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Amazon Bedrock
Chat playground
Chat playground
Info
Load examples
Compare mode
Select model
Configurations
Reset
Select model to load configs.
Try one of these examples or view more examples
Titan Text G1 - Express
Action items from a meeting transcript
Claude
Advanced Q&A with Citation
An example
Llama 2 Chat 13B
Chain of thought
An example prompt that uses
AI21 labs
Jurassic-2 Ultra
Earnings call summarization
A prompt that
Run
Model metrics
Define metric criteria


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Amazon Bedrock
Chat playground
Chat playground
Info
Load examples
Compare mode
Select model
Configurations
Reset
Select model to load configs.
Try one of these examples or view more examples
Titan Text G1 - Express
Action items from a meeting transcript
Claude
Advanced Q&A with Citation
An example
Llama 2 Chat 13B
Chain of thought
An example prompt that uses
AI21 labs
Jurassic-2 Ultra
Earnings call summarization
A prompt that
Run
Model metrics
Define metric criteria


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Compare mode
Define metric criteria
Select model
1. Category
Model providers
AI21 Labs
Amazon
Anthropic
Cohere
Meta
Mistral AI
2. Model
Models with access (2)
Jurassic-2 Mid
Text model | Context size = 8k
Jurassic-2 Ultra
Text model | Context size = 8k
Not seeing a model you are interested in? Check out all supported models here
3. Throughput
Select model to show throughput options.
Cancel
Apply


Select model
1. Category
Model providers
AI21 Labs
Amazon
Anthropic
Cohere
Meta
Mistral AI
2. Model
Models with access (5)
Claude Instant 1.2 v1.2
Text model | Context size = up to 100k
Claude 2.1 v2.1
Text model | Context size = up to 200k
Claude 2 v2
Text model | Context size = up to 100k
Claude 3 Sonnet v1
Text & vision model | Context size = up to 200k
Claude 3 Haiku v1
Text & vision model | Context size = up to 200k
Models without access (1)
Request access
Claude 3 Opus v1
Text & vision model | Context size = up to 200k
3. Throughput
Select model to show throughput options.
Cancel
Apply


us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Prompt flows
Preview
Assessment & deployment
Compare mode
Define metric criteria
Select model
1. Category
Model providers
AI21 Labs
Amazon
Anthropic
Cohere
Meta
Mistral AI
2. Model
Models with access (5)
Claude Instant 1.2 v1.2
Text model | Context size = up to 100k
Claude 2.1 v2.1
Text model | Context size = up to 200k
Claude 2 v2
Text model | Context size = up to 100k
Claude 3 Sonnet v1
Text & vision model | Context size = up to 200k
Claude 3 Haiku v1
Text & vision model | Context size = up to 200k
Models without access (1)
Request access
Claude 3 Opus v1
Text & vision model | Context size = up to 200k
3. Throughput
On-demand (ODT)
Cancel
Apply


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Amazon Bedrock
Chat playground
Chat playground
Info
Load examples
Compare mode
Claude 3 Sonnet
v1
ODT
Change
Try one of these examples or view more examples
Claude 3 Sonnet
Advanced Q&A with Citations
An example prompt for
Claude 3 Sonnet
Craft a Design Brief
Craft a design brief for a holistic brand identity,
Claude 3 Sonnet
Meeting Notes Summarizer
Distill meetings into
Write a prompt... (Shift + ENTER to start a new line, and ENTER to generate a response)
Run
Choose files
The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp. Maximum 5 files, 4.5 MB each.
Configurations
Reset
System prompts
Info
Add system prompts
Randomness and diversity
Info
Temperature
1
Top P
0.999
Top K
250
Model metrics
Define metric criteria

▼ Content preferences

Language

中文(繁體)

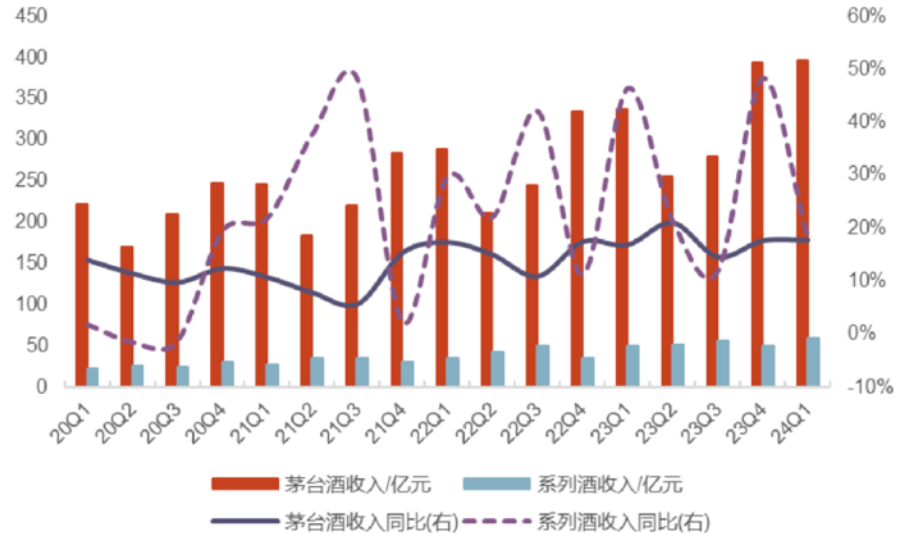
演示1: 图表理解

System Prompt

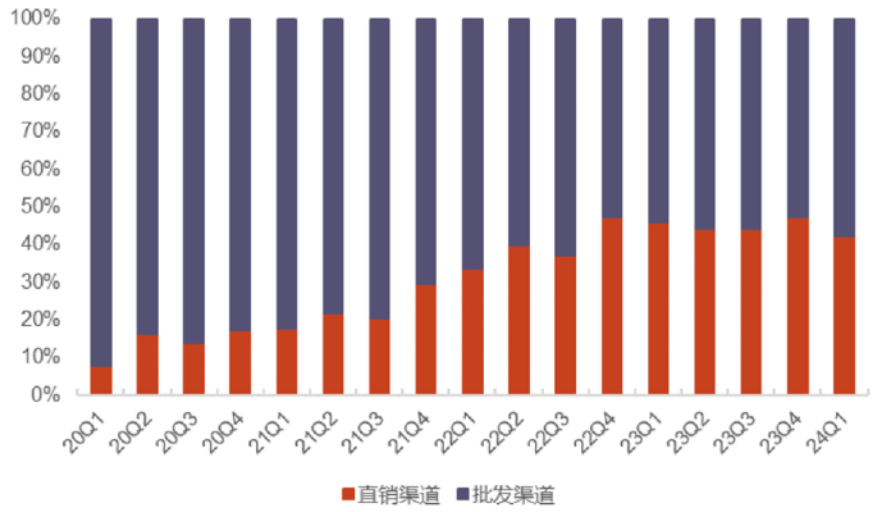
```
1  你是一个分析图表的专家，特别是图表里的柱状图，折线图，或者混合折线和柱状图在一起的图表。
2  请首先给出趋势或概览分析，之后仔细确认每一个折线的每一个波峰和每一个波谷对应的横轴和纵轴的数字（纵轴可能有两个）并列出。使用中文回复。
```

用户输入

图片1



图片2



这两张图是茅台公司酒产品20Q1-24Q1的统计图，请仔细分析并给出你的分析结果：

▼ Content preferences

Language

中文(繁體)

演示1: 图表理解

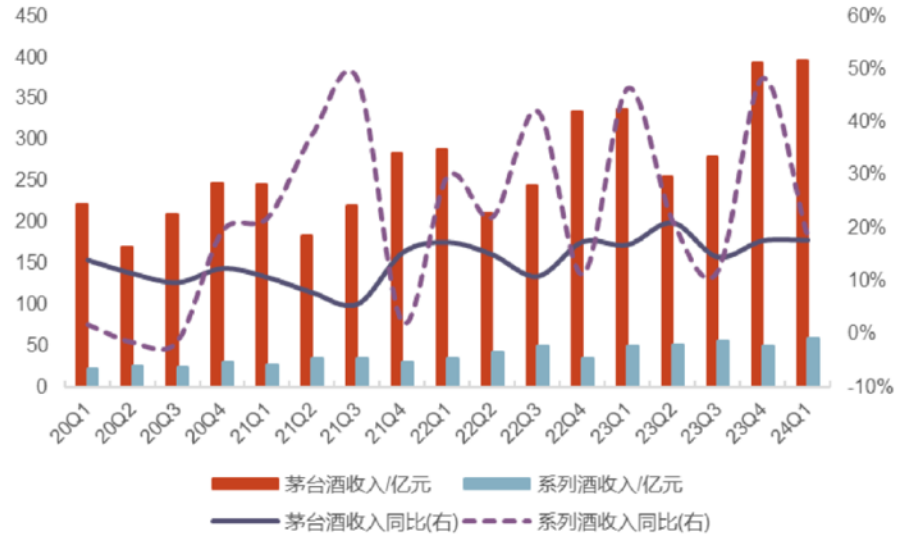
System Prompt

Copied!

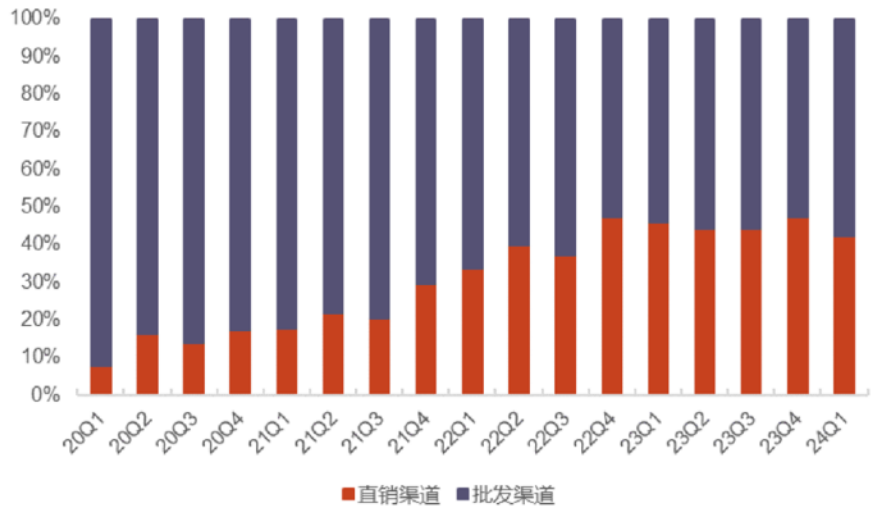
```
1  你是一个分析图表的专家，特别是图表里的柱状图，折线图，或者混合折线和柱状图在一起的图表。
2  请首先给出趋势或概览分析，之后仔细确认每一个折线的每一个波峰和每一个波谷对应的横轴和纵轴的数字（纵轴可能有两个）并列出。使用中文回复。
```

用户输入

图片1



图片2



这两张图是茅台公司酒产品20Q1-24Q1的统计图，请仔细分析并给出你的分析结果：

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Chat playground Info

Load examples | Compare mode

Claude 3 Sonnet *v1* | ODT

Change

你是一个分析图表的专家，特别是图表里的柱状图，折线图，或者混合折线和柱状图在一起的图表。
请首先给出趋势或概览分析，之后仔细确认每一个折线的每一个波峰和每一个波谷对应的横轴和纵轴的数字（纵轴可能有两个）并列出。使用中文回复。

Run

Choose files

The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp. Maximum 5 files, 4.5 MB each.

Configurations

Reset

System prompts Info

Add system prompts

Randomness and diversity Info

Temperature	1
Top P	0.999
Top K	250

Model metrics

Define metric criteria

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
  - Prompt management Preview

Amazon Bedrock > Chat playground

Chat playground Info

Load examples | Compare mode

Claude 3 Sonnet *v1* | ODT
Change

你是一个分析图表的专家，特别是图表里的柱状图，折线图，或者混合折线和柱状图在一起的图表。
请首先给出趋势或概览分析，之后仔细确认每一个折线的每一个波峰和每一个波谷对应的横轴和纵轴的数字（纵轴可能有两个）并列出。使用中文回复。

Run

Choose files

The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp. Maximum 5 files, 4.5 MB each.

Configurations
Reset

System prompts Info

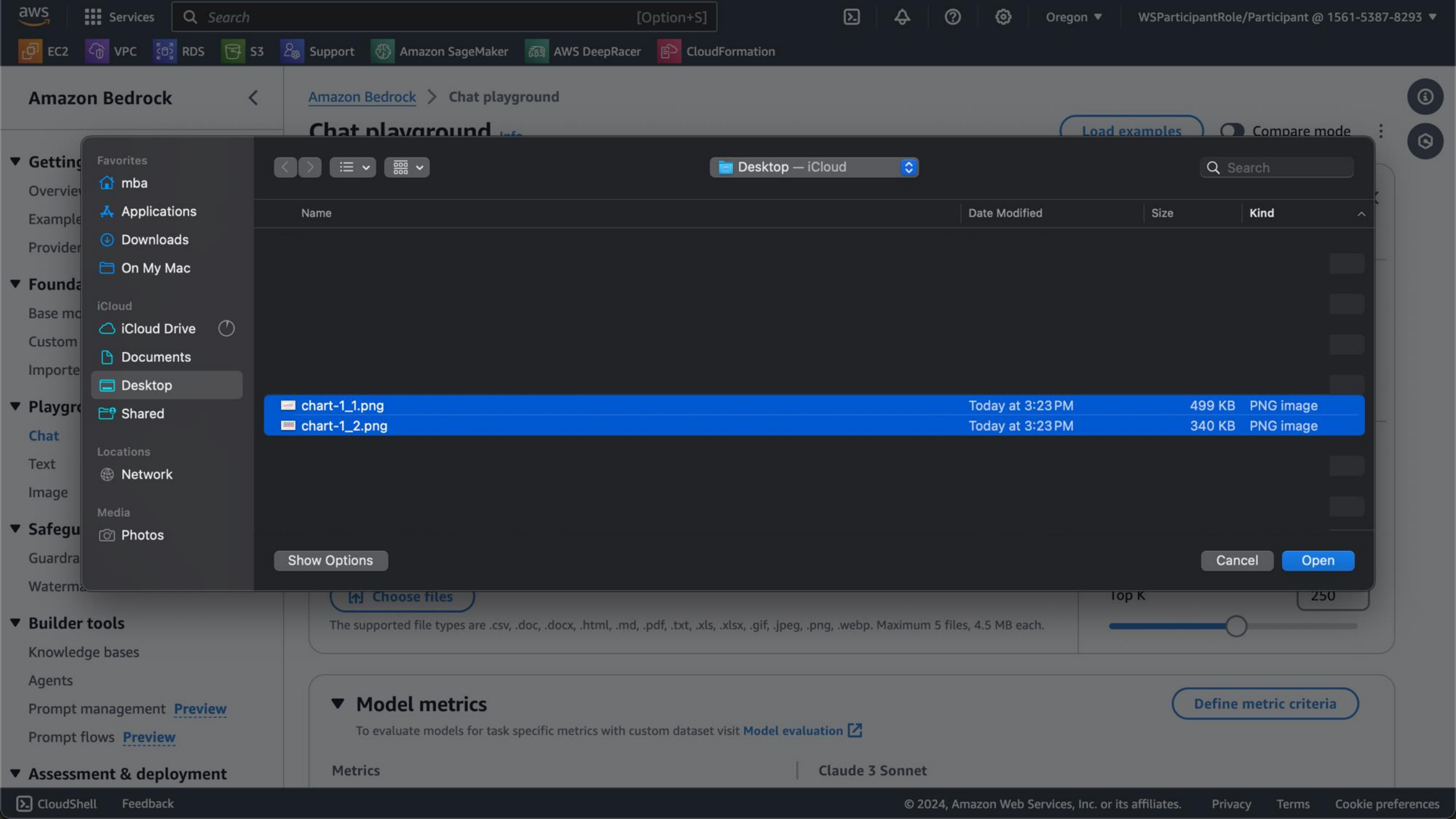
Add system prompts

Randomness and diversity Info

Temperature	1
Top P	0.999
Top K	250

Model metrics

Define metric criteria


Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Amazon Bedrock > Chat playground
Chat playground
Load examples
Compare mode
Favorites
mba
Applications
Downloads
On My Mac
iCloud
iCloud Drive
Documents
Desktop
Shared
Locations
Network
Media
Photos
Desktop — iCloud
Search
Name
Date Modified
Size
Kind
chart-1_1.png
Today at 3:23 PM
499 KB
PNG image
chart-1_2.png
Today at 3:23 PM
340 KB
PNG image
Show Options
Cancel
Open
Choose files
The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp. Maximum 5 files, 4.5 MB each.
Top K
250
Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview
Assessment & deployment
Model metrics
To evaluate models for task specific metrics with custom dataset visit Model evaluation
Define metric criteria
Metrics
Claude 3 Sonnet
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Chat playground Info

Load examples | Compare mode

Claude 3 Sonnet *v1* | ODT

Change

你是一个分析图表的专家，特别是图表里的柱状图，折线图，或者混合折线和柱状图在一起的图表。
请首先给出趋势或概览分析，之后仔细确认每一个折线的每一个波峰和每一个波谷对应的横轴和纵轴的数字（纵轴可能有两个）并列出。使用中文回复。

Run

Choose files

The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp. Maximum 5 files, 4.5 MB each.

chart-1_1.png
498.55 KB

chart-1_2.png
340.15 KB

Configurations

Reset

System prompts Info

Add system prompts

Randomness and diversity Info

Temperature 1

Top P 0.999

Top K 250

Model metrics

Define metric criteria

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Chat playground Info

Load examples | Compare mode

Claude 3 Sonnet v1 | ODT

Change

你是一个分析图表的专家，特别是图表里的柱状图，折线图，或者混合折线和柱状图在一起的图表。
请首先给出趋势或概览分析，之后仔细确认每一个折线的每一个波峰和每一个波谷对应的横轴和纵轴的数字（纵轴可能有两个）并列出。使用中文回复。

Run

Choose files

The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp. Maximum 5 files, 4.5 MB each.

chart-1_1.png
498.55 KB

chart-1_2.png
340.15 KB

Configurations

Reset

System prompts Info

Add system prompts

Randomness and diversity Info

Temperature 1

Top P 0.999

Top K 250

Model metrics

Define metric criteria

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/chat-playground?modelId=anthropic.claude-3-sonnet-20240229-v1%3A0

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Chat playground Info

Load examples | Compare mode

Claude 3 Sonnet *v1* | ODT

Change

从图表整体趋势来看,酒收入和同比变化呈现规律性的波动。在某些季节,如2009年第一季度、2011年第一季度和2013年第一季度,酒收入和同比增长都达到峰值。而在其他季节,如2010年第三季度和2012年第三季度,酒收入和同比增长则处于低谷。此外,折线图显示了百分比格式的直接渠道和批发渠道收入占比,两条折线保持相对稳定,直接渠道收入占比略高于批发渠道。

现在我详细列出每一个波峰和波谷对应的数值:

酒收入(亿元)波峰:
2009年第1季度:158.62
2011年第1季度:264.11
2013年第1季度:392.02

Write a prompt... (Shift + ENTER to start a new line, and ENTER to generate a response)

Run

Choose files

The supported file types are .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp. Maximum 5 files, 4.5 MB each.

Configurations

Reset

System prompts Info

Add system prompts

Randomness and diversity Info

Temperature 1

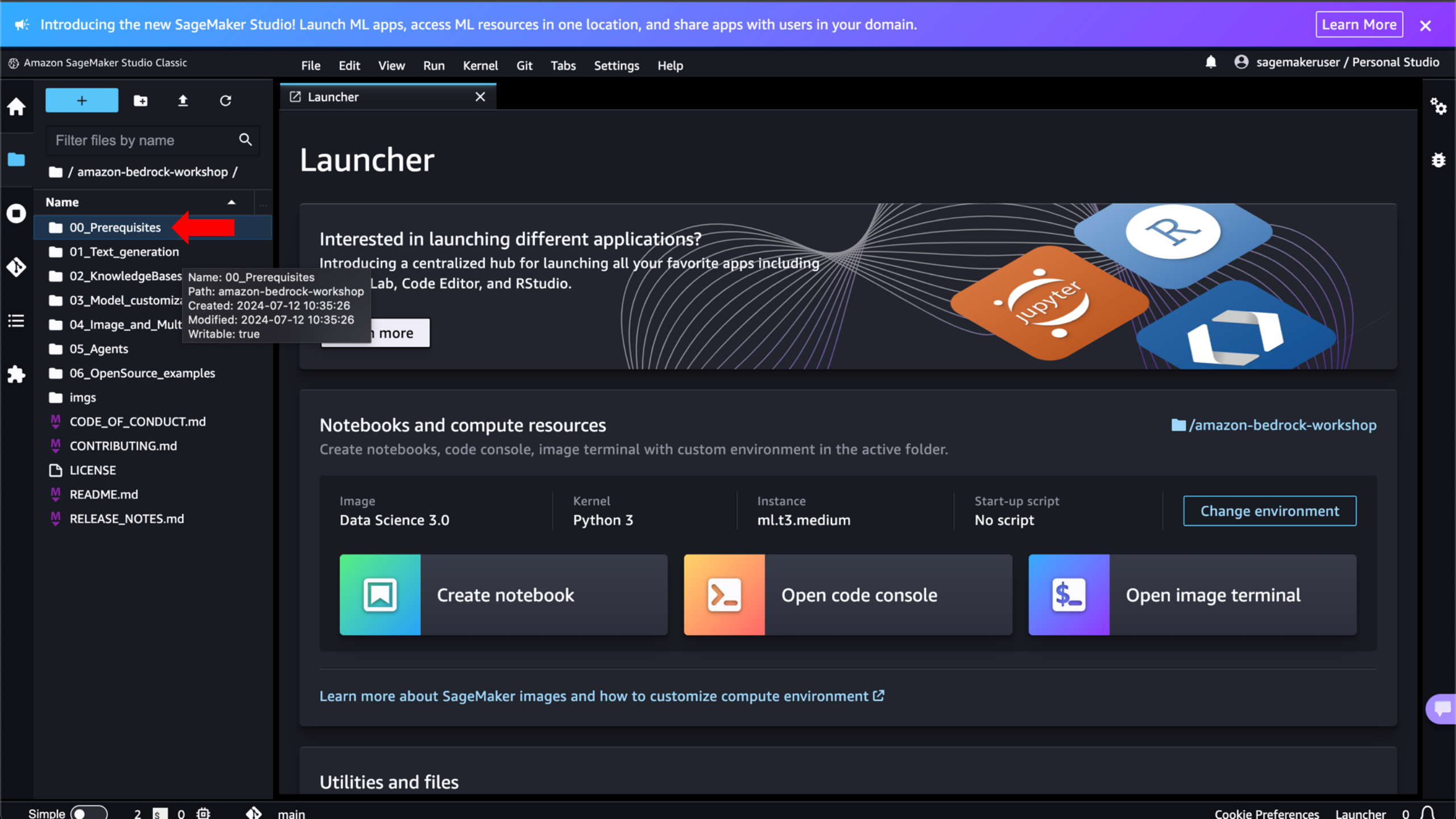
Top P 0.999

Top K 250

Model metrics

Define metric criteria

Bedrock API Hands-On Lab


Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain.
Learn More
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Launcher
Filter files by name
/ amazon-bedrock-workshop /
Name
00_Prerequisites
01_Text_generation
02_KnowledgeBases
03_Model_customiza
04_Image_and_Mult
05_Agents
06_OpenSource_examples
imgs
CODE_OF_CONDUCT.md
CONTRIBUTING.md
LICENSE
README.md
RELEASE_NOTES.md
Name: 00_Prerequisites
Path: amazon-bedrock-workshop
Created: 2024-07-12 10:35:26
Modified: 2024-07-12 10:35:26
Writable: true
Launcher
Interested in launching different applications?
Introducing a centralized hub for launching all your favorite apps including
Lab, Code Editor, and RStudio.
more
Notebooks and compute resources
/amazon-bedrock-workshop
Create notebooks, code console, image terminal with custom environment in the active folder.
Image
Data Science 3.0
Kernel
Python 3
Instance
ml.t3.medium
Start-up script
No script
Change environment
Create notebook
Open code console
Open image terminal
Learn more about SageMaker images and how to customize compute environment
Utilities and files
Simple
2
0
main
Cookie Preferences
Launcher
0

Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain. Learn More

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / Personal Studio

Launcher

Filter files by name

/ amazon-bedrock-workshop / 00_Prerequisites /

Name

bedrock_basics.ipynb

Getting_started_with_Converse...

README.md

Launcher

Interested in launching different applications?

Introducing a centralized hub for launching all your favorite apps including JupyterLab, Code Editor, and RStudio.

Learn more

Notebooks and compute resources

/amazon-bedrock-workshop/00_Prerequisites

Create notebooks, code console, image terminal with custom environment in the active folder.

Image	Kernel	Instance	Start-up script	
Data Science 3.0	Python 3	ml.t3.medium	No script	Change environment

Create notebook

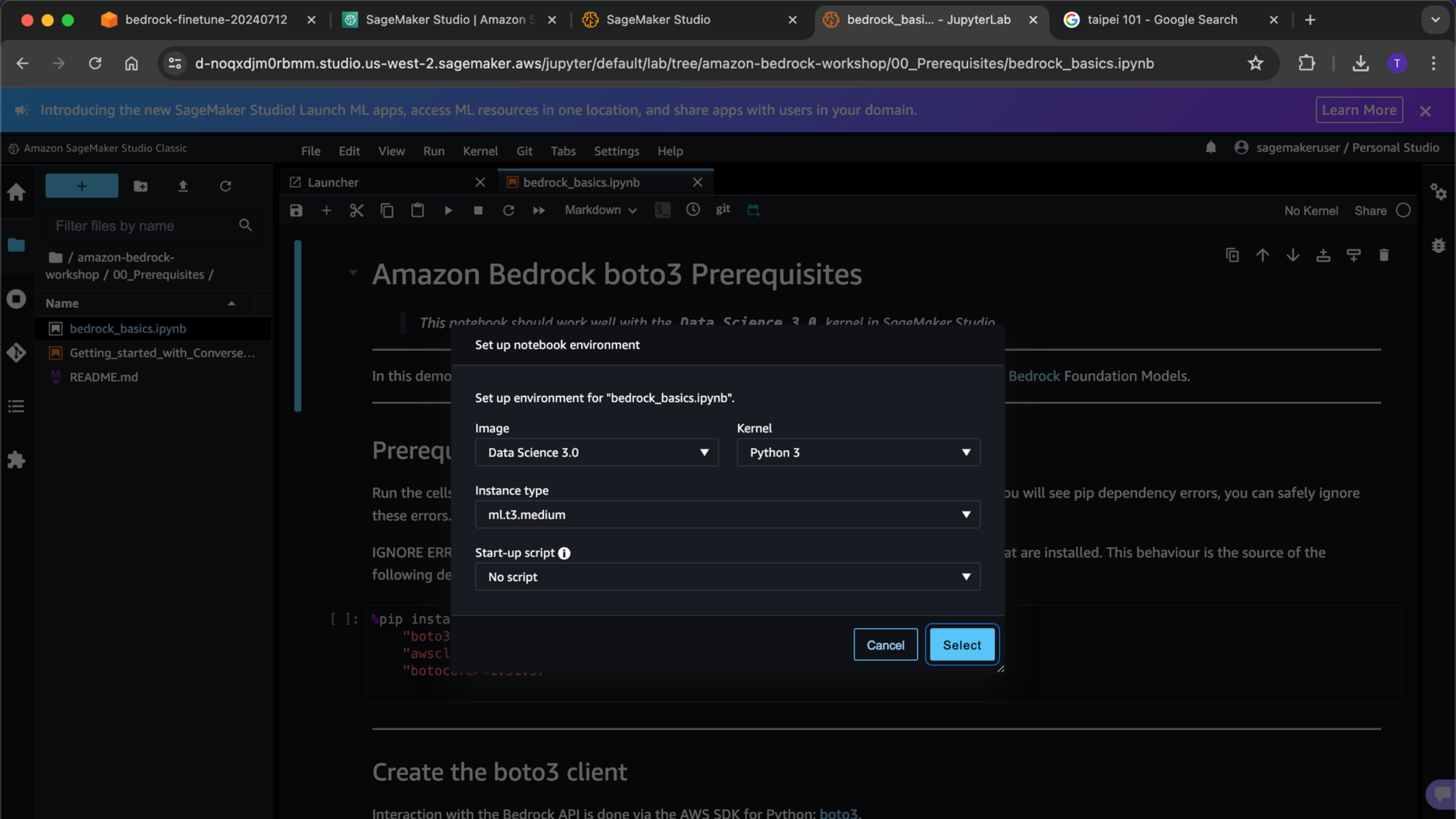
Open code console

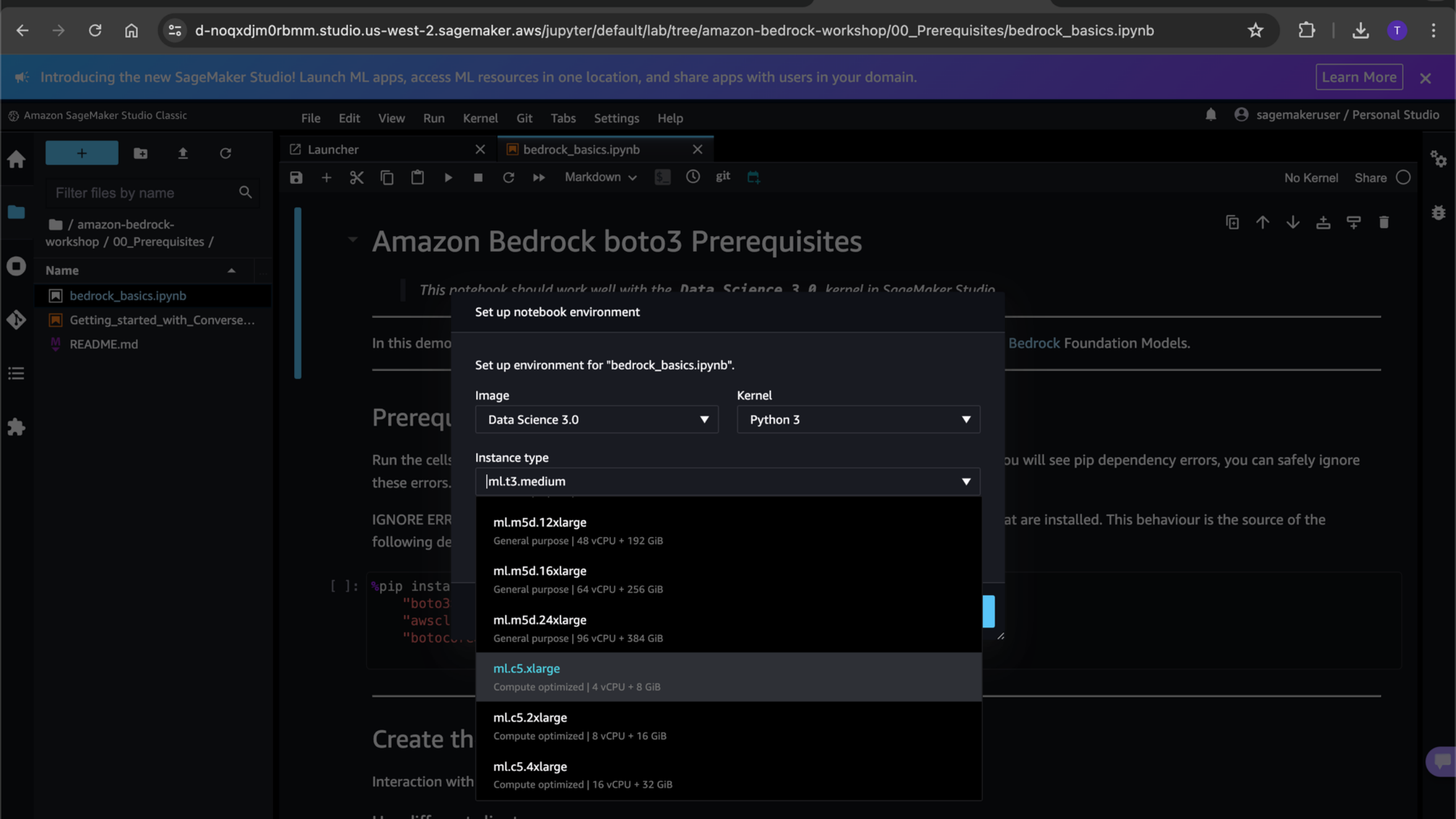
Open image terminal

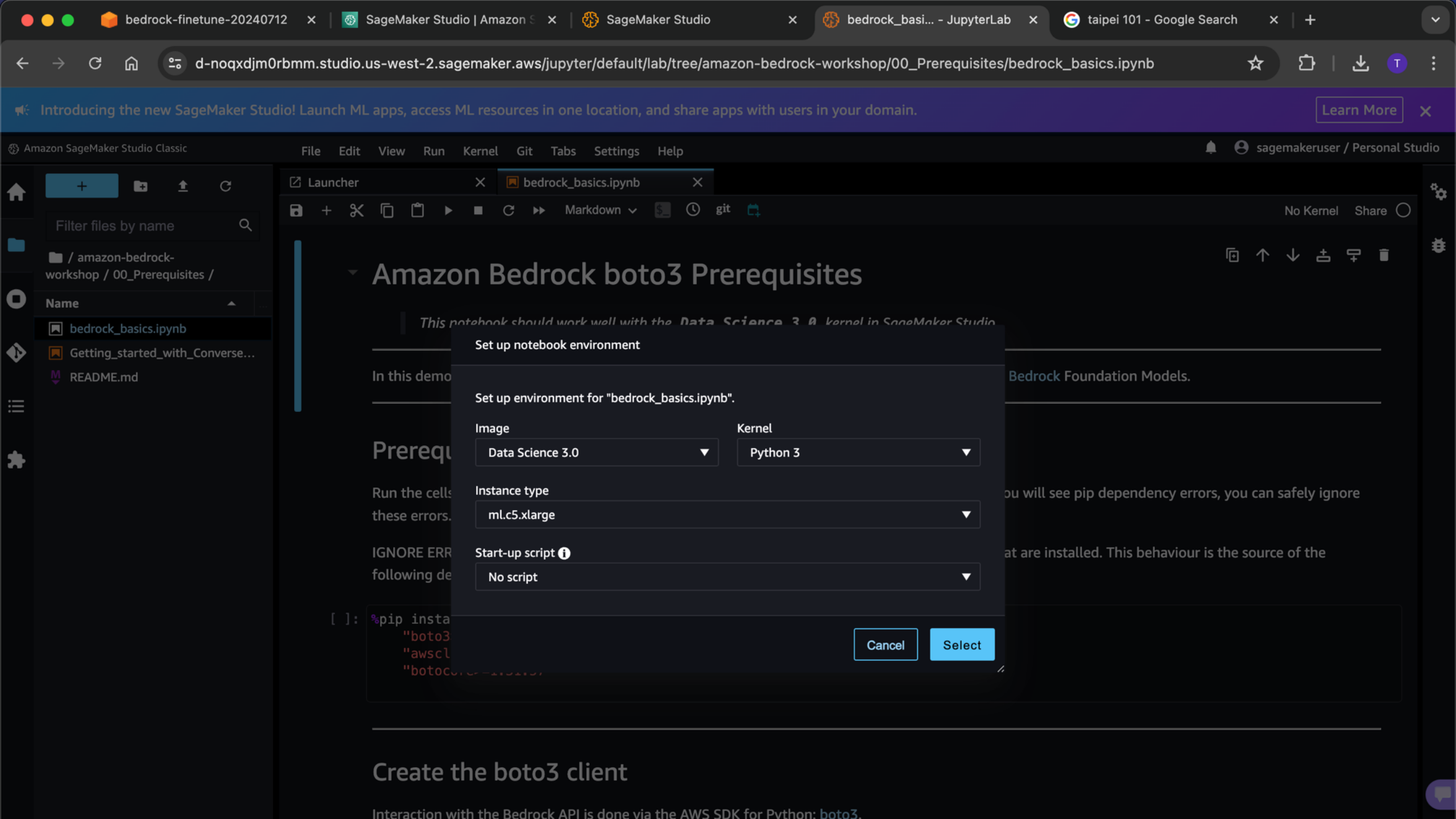
Learn more about SageMaker images and how to customize compute environment

Utilities and files

Simple main Cookie Preferences Launcher


bedrock-finetune-20240712
SageMaker Studio | Amazon
SageMaker Studio
bedrock_basi... - JupyterLab
taipei 101 - Google Search
d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb
Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain.
Learn More
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Launcher
bedrock_basics.ipynb
Markdown
git
No Kernel
Share
Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md
Amazon Bedrock boto3 Prerequisites
Set up notebook environment
Set up environment for "bedrock_basics.ipynb".
Image
Data Science 3.0
Kernel
Python 3
Instance type
ml.t3.medium
Start-up script
No script
Cancel
Select
Create the boto3 client


d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb
Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain.
Learn More
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Launcher
bedrock_basics.ipynb
Markdown
git
No Kernel
Share
Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md
Amazon Bedrock boto3 Prerequisites
This notebook should work well with the Data Science 3.0 kernel in SageMaker Studio
In this demo
Bedrock Foundation Models.
Prerequ
Run the cells
these errors.
ou will see pip dependency errors, you can safely ignore
IGNORE ERR
following de
at are installed. This behaviour is the source of the
[ ]: %pip insta
"boto3
"awscl
"botoco
Create th
Interaction with
Set up notebook environment
Set up environment for "bedrock_basics.ipynb".
Image
Data Science 3.0
Kernel
Python 3
Instance type
ml.t3.medium
ml.m5d.12xlarge
General purpose | 48 vCPU + 192 GiB
ml.m5d.16xlarge
General purpose | 64 vCPU + 256 GiB
ml.m5d.24xlarge
General purpose | 96 vCPU + 384 GiB
ml.c5.xlarge
Compute optimized | 4 vCPU + 8 GiB
ml.c5.2xlarge
Compute optimized | 8 vCPU + 16 GiB
ml.c5.4xlarge
Compute optimized | 16 vCPU + 32 GiB


bedrock-finetune-20240712
SageMaker Studio | Amazon
SageMaker Studio
bedrock_basi... - JupyterLab
taipei 101 - Google Search
d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb
Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain.
Learn More
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Launcher
bedrock_basics.ipynb
Markdown
git
No Kernel
Share
Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md
Amazon Bedrock boto3 Prerequisites
This notebook should work well with the Data Science 3.0 kernel in SageMaker Studio
In this demo
Bedrock Foundation Models.
Prerequ
Run the cells
these errors.
ou will see pip dependency errors, you can safely ignore
IGNORE ERR
following de
at are installed. This behaviour is the source of the
[ ]: %pip insta
"boto3
"awscl
"botocore
Create the boto3 client
Interaction with the Bedrock API is done via the AWS SDK for Python: boto3.
Set up notebook environment
Set up environment for "bedrock_basics.ipynb".
Image
Data Science 3.0
Kernel
Python 3
Instance type
ml.c5.xlarge
Start-up script
No script
Cancel
Select

bedrock-finetune-20240712 | SageMaker Studio | Amazon S | SageMaker Studio | bedrock_basi... - JupyterLab | taipei 101 - Google Search

d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb

Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain. Learn More

Amazon SageMaker Studio Classic | File Edit View Run Kernel Git Tabs Settings Help | sagemakeruser / Personal Studio

Launcher | bedrock_basics.ipynb

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ amazon-bedrock-workshop / 00_Prerequisites /

Name

- bedrock_basics.ipynb
- Getting_started_with_Converse...
- README.md

Amazon Bedrock boto3 Prerequisites

> *This notebook should work well with the **`Data Science 3.0`** kernel in SageMaker Studio*

In this demo notebook, we demonstrate how to use the `boto3` Python SDK to work with Amazon Bedrock Foundation Models.

Prerequisites

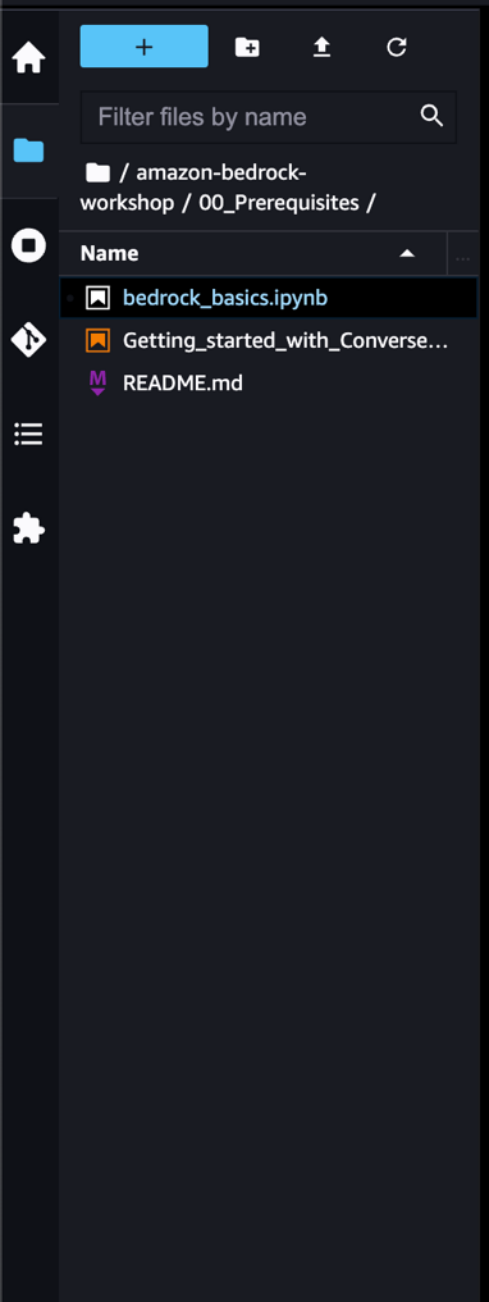
Run the cells in this section to install the packages needed by the notebooks in this workshop. ⚠️ You will see pip dependency errors, you can safely ignore these errors. ⚠️

IGNORE ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.

[ ]:
```
%pip install --no-build-isolation --force-reinstall \
    "boto3>=1.28.57" \
    "awscli>=1.29.57" \
    "botocore>=1.31.57"
```

Create the boto3 client

Interaction with the Bedrock API is done via the AWS SDK for Python: boto3.



Launcher | bedrock_basics.ipynb

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Amazon Bedrock boto3 Prerequisites

> *This notebook should work well with the* ***`Data Science 3.0`*** *kernel in SageMaker Studio*

In this demo notebook, we demonstrate how to use the `boto3` Python SDK to work with Amazon Bedrock Foundation Models.

Prerequisites

Run the cells in this section to install the packages needed by the notebooks in this workshop. ⚠️ You will see pip dependency errors, you can safely ignore these errors. ⚠️

IGNORE ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.

[2]:
```
%pip install --no-build-isolation --force-reinstall \
    "boto3>=1.28.57" \
    "awscli>=1.29.57" \
    "botocore>=1.31.57"
```

```
Collecting boto3>=1.28.57
  Downloading boto3-1.34.144-py3-none-any.whl.metadata (6.6 kB)
Collecting awscli>=1.29.57
  Downloading awscli-1.33.26-py3-none-any.whl.metadata (11 kB)
Collecting botocore>=1.31.57
  Downloading botocore-1.34.144-py3-none-any.whl.metadata (5.7 kB)
Collecting jmespath<2.0.0,>=0.7.1 (from boto3>=1.28.57)
  Downloading jmespath-1.0.1-py3-none-any.whl.metadata (7.6 kB)
Collecting s3transfer<0.11.0,>=0.10.0 (from boto3>=1.28.57)
  Downloading s3transfer-0.10.2-py3-none-any.whl.metadata (1.7 kB)
Collecting docutils<0.17,>=0.10 (from awscli>=1.29.57)
```

Launcher | bedrock_basics.ipynb
Code git Cluster Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB Share

Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md

Create the boto3 client

Interaction with the Bedrock API is done via the AWS SDK for Python: boto3.

Use different clients

The boto3 provides different clients for Amazon Bedrock to perform different actions. The actions for `InvokeModel` and `InvokeModelWithResponseStream` are supported by Amazon Bedrock Runtime where as other operations, such as ListFoundationModels, are handled via Amazon Bedrock client.

Use the default credential chain

If you are running this notebook from Amazon Sagemaker Studio and your Sagemaker Studio execution role has permissions to access Bedrock you can just run the cells below as-is. This is also the case if you are running these notebooks from a computer whose default AWS credentials have access to Bedrock.

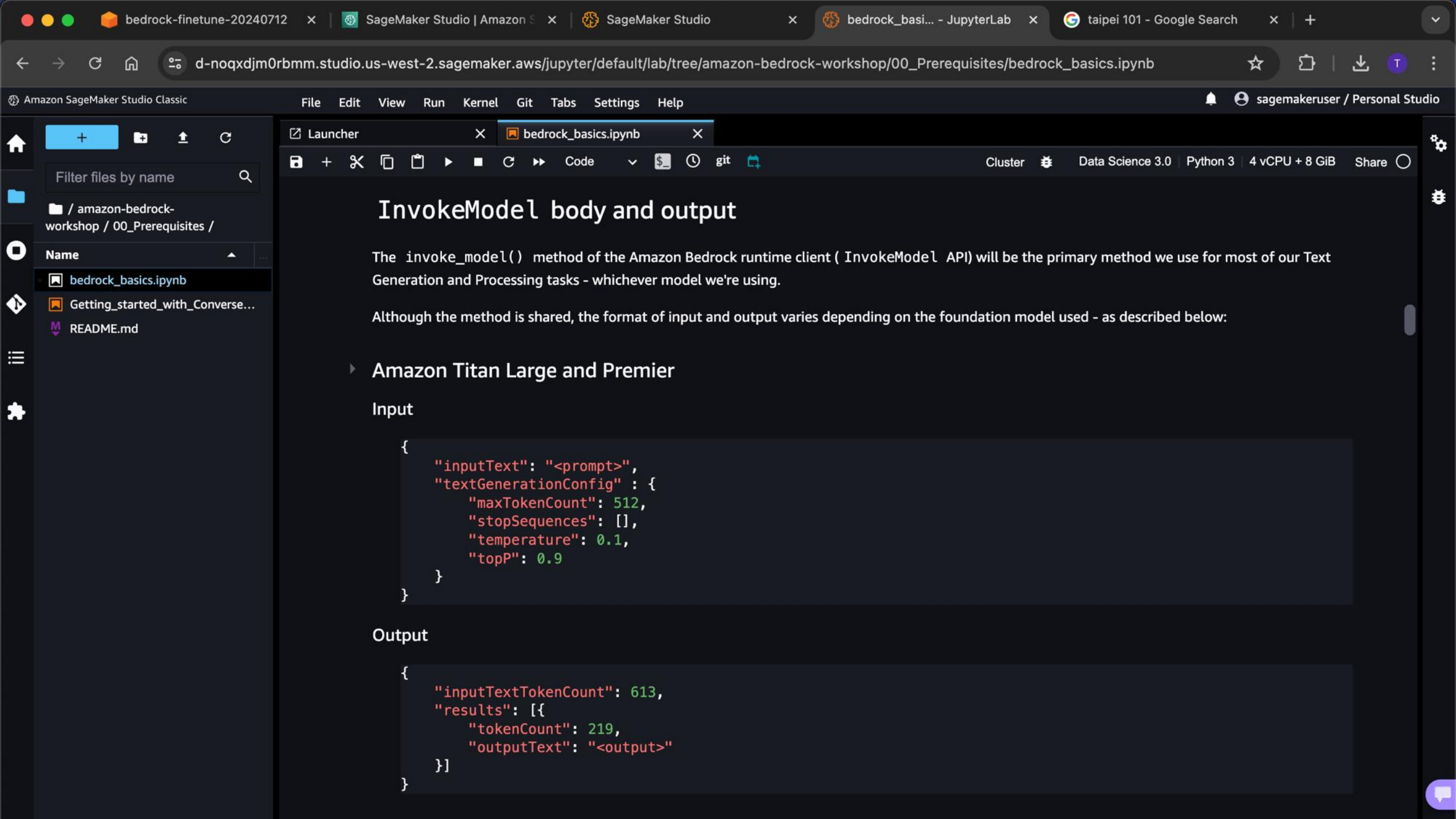
[3]:
```
import json
import os
import sys

import boto3

boto3_bedrock = boto3.client('bedrock')
```

Validate the connection

We can check the client works by trying out the `list_foundation_models()` method, which will tell us all the models available for us to use

[4]:
```
[models['modelId'] for models in boto3_bedrock.list_foundation_models()['modelSummaries']]
```

[4]:
```
['amazon.titan-tg1-large',
 'amazon.titan-embed-g1-text-02',
```


bedrock-finetune-20240712
SageMaker Studio | Amazon
SageMaker Studio
bedrock_basi... - JupyterLab
taipei 101 - Google Search
d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Launcher
bedrock_basics.ipynb
Code
git
Cluster
Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB
Share
Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md
InvokeModel body and output
The invoke_model() method of the Amazon Bedrock runtime client (InvokeModel API) will be the primary method we use for most of our Text Generation and Processing tasks - whichever model we're using.
Although the method is shared, the format of input and output varies depending on the foundation model used - as described below:
Amazon Titan Large and Premier
Input
{
"inputText": "<prompt>",
"textGenerationConfig" : {
"maxTokenCount": 512,
"stopSequences": [],
"temperature": 0.1,
"topP": 0.9
}
}
Output
{
"inputTextTokenCount": 613,
"results": [{
"tokenCount": 219,
"outputText": "<output>"
}]
}

Launcher
bedrock_basics.ipynb
Code git Cluster Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB Share

Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md

Anthropic Claude

Input

```
{
    "prompt": "\n\nHuman:<prompt>\n\nAnswer:",
    "max_tokens_to_sample": 300,
    "temperature": 0.5,
    "top_k": 250,
    "top_p": 1,
    "stop_sequences": ["\n\nHuman:"]
}
```

Output

```
{
    "completion": "<output>",
    "stop_reason": "stop_sequence"
}
```

Stability AI Stable Diffusion XL

Input

```
{
    "text_prompts": [
        {"text": "this is where you place your input text"}
    ],
    "cfg_scale": 10,
    "seed": 0,
    "steps": 50
}
```

Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md

Common inference parameter definitions

Randomness and Diversity

Foundation models generally support the following parameters to control randomness and diversity in the response.

Temperature – Large language models use probability to construct the words in a sequence. For any given next word, there is a probability distribution of options for the next word in the sequence. When you set the temperature closer to zero, the model tends to select the higher-probability words. When you set the temperature further away from zero, the model may select a lower-probability word.

In technical terms, the temperature modulates the probability density function for the next tokens, implementing the temperature sampling technique. This parameter can deepen or flatten the density function curve. A lower value results in a steeper curve with more deterministic responses, and a higher value results in a flatter curve with more random responses.

Top K – Temperature defines the probability distribution of potential words, and Top K defines the cut off where the model no longer selects the words. For example, if K=50, the model selects from 50 of the most probable words that could be next in a given sequence. This reduces the probability that an unusual word gets selected next in a sequence. In technical terms, Top K is the number of the highest-probability vocabulary tokens to keep for Top- K-filtering - This limits the distribution of probable tokens, so the model chooses one of the highest- probability tokens.

Top P – Top P defines a cut off based on the sum of probabilities of the potential choices. If you set Top P below 1.0, the model considers the most probable options and ignores less probable ones. Top P is similar to Top K, but instead of capping the number of choices, it caps choices based on the sum of their probabilities. For the example prompt "I hear the hoof beats of ," you may want the model to provide "horses," "zebras" or "unicorns" as the next word. If you set the temperature to its maximum, without capping Top K or Top P, you increase the probability of getting unusual results such as "unicorns." If you set the temperature to 0, you increase the probability of "horses." If you set a high temperature and set Top K or Top P to the maximum, you increase the probability of "horses" or "zebras," and decrease the probability of "unicorns."

Length

The following parameters control the length of the generated response.

Response length – Configures the minimum and maximum number of tokens to use in the generated response.

Length penalty – Length penalty optimizes the model to be more concise in its output by penalizing longer responses. Length penalty differs from response

bedrock-finetune-20240712 | SageMaker Studio | Amazon S | SageMaker Studio | bedrock_basi... - JupyterLab | taipei 101 - Google Search

d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / Personal Studio

Launcher | bedrock_basics.ipynb

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ amazon-bedrock-workshop / 00_Prerequisites /

Name

- bedrock_basics.ipynb
- Getting_started_with_Converse...
- README.md

Try out the models

With some theory out of the way, let's see the models in action! Run the cells below to see basic, synchronous example invocations for each model:

[5]:
```
import boto3
import botocore
import json

bedrock_runtime = boto3.client('bedrock-runtime')
```

Amazon Titan Text Premier

[6]:
```
# If you'd like to try your own prompt, edit this parameter!
prompt_data = """Command: Write me a blog about making strong business decisions as a leader.

Blog:
"""
```

Next, we will construct the body with the `prompt_data` above, and add a optional parameters like `topP` and `temperature`:

[7]:
```
try:

    body = json.dumps({"inputText": prompt_data, "textGenerationConfig" : {"topP":0.95, "temperature":0.2}})
    modelId = "amazon.titan-tg1-large" #
    accept = "application/json"
    contentType = "application/json"

    response = bedrock_runtime.invoke_model(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )
    response_body = json.loads(response.get("body").read())
```

bedrock-finetune-20240712 | SageMaker Studio | Amazon S | SageMaker Studio | bedrock_basi... - JupyterLab | taipei 101 - Google Search

d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / Personal Studio

Launcher | bedrock_basics.ipynb

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ amazon-bedrock-workshop / 00_Prerequisites /

Name

- bedrock_basics.ipynb
- Getting_started_with_Converse...
- README.md

```
As a leader, making strong business decisions is crucial for the success of your organization. Here are some key factors to consid
er when making business decisions:

Define your goals and objectives: Clearly define what you want to achieve as a leader and for your organization. This will help yo
u make decisions that are aligned with your vision and mission.

Gather relevant information: Collect and analyze relevant information about the situation, market, industry, and your competitors.
This will help you make informed decisions that are based on facts and data.

Consider multiple perspectives: Seek input and perspectives from your team, stakeholders, and experts in the field
```

Anthropic Claude

[8]:
```
# If you'd like to try your own prompt, edit this parameter!
prompt_data = """Human: Write me a blog about making strong business decisions as a leader.

Assistant:
"""
```

[9]:
```
body = json.dumps({"prompt": prompt_data, "max_tokens_to_sample": 500})
modelId = "anthropic.claude-instant-v1"  # change this to use a different version from the model provider
accept = "application/json"
contentType = "application/json"

try:

    response = bedrock_runtime.invoke_model(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )
    response_body = json.loads(response.get("body").read())

    print(response_body.get("completion"))

except botocore.exceptions.ClientError as error:
```

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / Personal Studio

Launcher | bedrock_basics.ipynb

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ amazon-bedrock-workshop / 00_Prerequisites /

Name

- bedrock_basics.ipynb
- Getting_started_with_Converse...
- README.md

Anthropic Claude

[8]:
```
# If you'd like to try your own prompt, edit this parameter!
prompt_data = """Human: Write me a blog about making strong business decisions as a leader.

Assistant:
"""
```

[9]:
```
body = json.dumps({"prompt": prompt_data, "max_tokens_to_sample": 500})
modelId = "anthropic.claude-instant-v1"  # change this to use a different version from the model provider
accept = "application/json"
contentType = "application/json"

try:

    response = bedrock_runtime.invoke_model(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )
    response_body = json.loads(response.get("body").read())

    print(response_body.get("completion"))

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")


    else:
        raise error
```

```
 Here is a draft blog post on making strong business decisions as a leader:

## How to Make Strong Business Decisions as a Leader
```

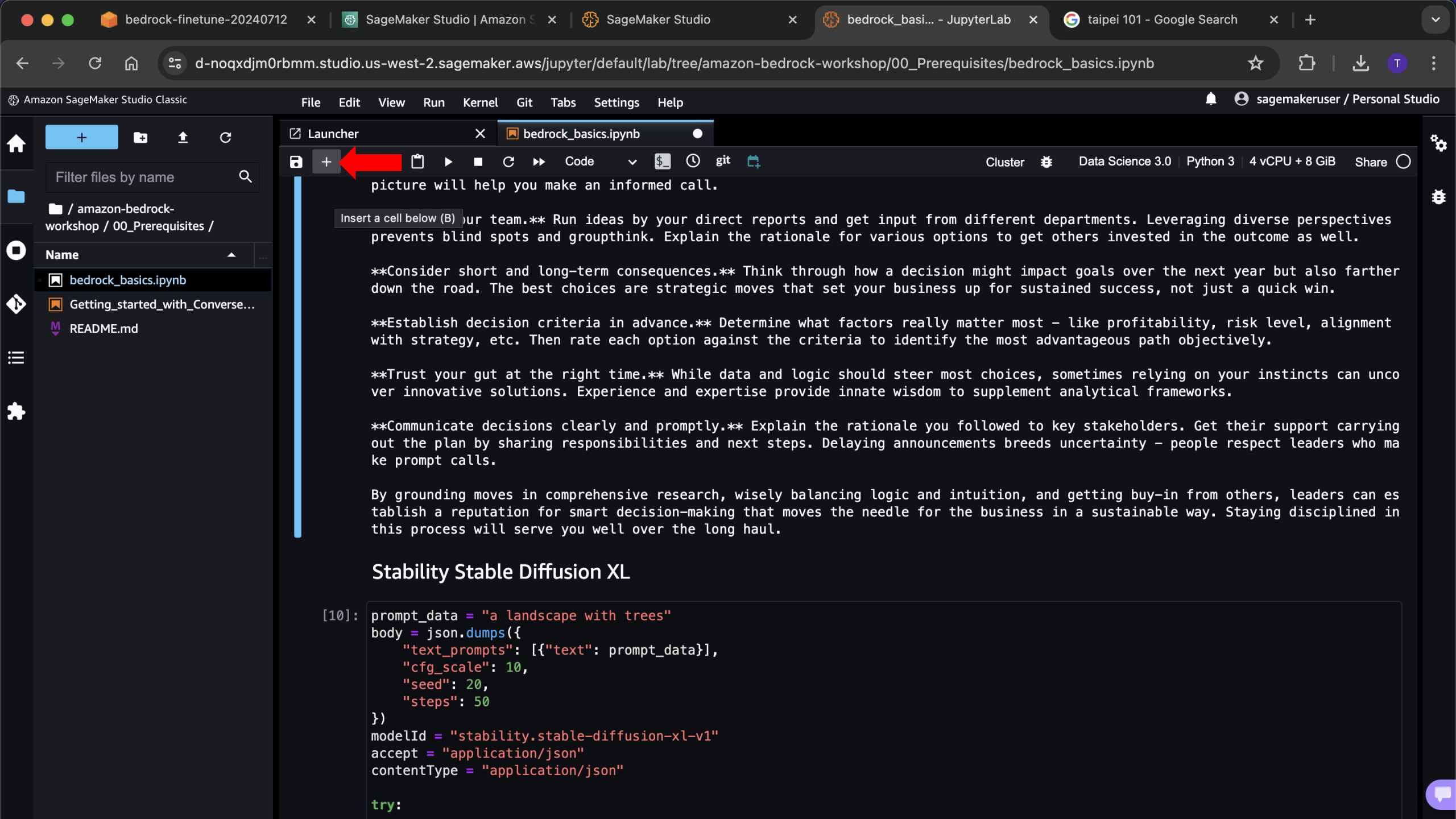
Launcher
bedrock_basics.ipynb

Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md

```
 Here is a draft blog post on making strong business decisions as a leader:

## How to Make Strong Business Decisions as a Leader

As a leader, one of your most important responsibilities is making strategic decisions that will move your business in the right d
irection. However, decision-making can also be one of the most difficult parts of being a leader. There are so many factors to con
sider and potential risks involved with any choice.

The key is making decisions thoughtfully and deliberately through a structured process. Here are some tips for making strong, impa
ctful choices as the leader of your organization:

**Gather comprehensive data and insights.** Before committing to a course of action, take time to research the issue from all angl
es. Look at trends in your industry, get customer and employee feedback, analyze financial reports and key metrics. Having a full
picture will help you make an informed call.

**Consult your team.** Run ideas by your direct reports and get input from different departments. Leveraging diverse perspectives
prevents blind spots and groupthink. Explain the rationale for various options to get others invested in the outcome as well.

**Consider short and long-term consequences.** Think through how a decision might impact goals over the next year but also farther
down the road. The best choices are strategic moves that set your business up for sustained success, not just a quick win.

**Establish decision criteria in advance.** Determine what factors really matter most - like profitability, risk level, alignment
with strategy, etc. Then rate each option against the criteria to identify the most advantageous path objectively.

**Trust your gut at the right time.** While data and logic should steer most choices, sometimes relying on your instincts can unco
ver innovative solutions. Experience and expertise provide innate wisdom to supplement analytical frameworks.

**Communicate decisions clearly and promptly.** Explain the rationale you followed to key stakeholders. Get their support carrying
out the plan by sharing responsibilities and next steps. Delaying announcements breeds uncertainty - people respect leaders who ma
ke prompt calls.

By grounding moves in comprehensive research, wisely balancing logic and intuition, and getting buy-in from others, leaders can es
tablish a reputation for smart decision-making that moves the needle for the business in a sustainable way. Staying disciplined in
this process will serve you well over the long haul.
```

Stability Stable Diffusion XL

[10]:
```
prompt_data = "a landscape with trees"
body = json.dumps({
```


bedrock-finetune-20240712
SageMaker Studio | Amazon
SageMaker Studio
bedrock_basi... - JupyterLab
taipei 101 - Google Search
d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Launcher
bedrock_basics.ipynb
Code
git
Cluster
Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB
Share
Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md
Insert a cell below (B)
picture will help you make an informed call.
ur team.** Run ideas by your direct reports and get input from different departments. Leveraging diverse perspectives prevents blind spots and groupthink. Explain the rationale for various options to get others invested in the outcome as well.
Consider short and long-term consequences. Think through how a decision might impact goals over the next year but also farther down the road. The best choices are strategic moves that set your business up for sustained success, not just a quick win.
Establish decision criteria in advance. Determine what factors really matter most - like profitability, risk level, alignment with strategy, etc. Then rate each option against the criteria to identify the most advantageous path objectively.
Trust your gut at the right time. While data and logic should steer most choices, sometimes relying on your instincts can uncover innovative solutions. Experience and expertise provide innate wisdom to supplement analytical frameworks.
Communicate decisions clearly and promptly. Explain the rationale you followed to key stakeholders. Get their support carrying out the plan by sharing responsibilities and next steps. Delaying announcements breeds uncertainty - people respect leaders who make prompt calls.
By grounding moves in comprehensive research, wisely balancing logic and intuition, and getting buy-in from others, leaders can establish a reputation for smart decision-making that moves the needle for the business in a sustainable way. Staying disciplined in this process will serve you well over the long haul.
Stability Stable Diffusion XL
[10]:
prompt_data = "a landscape with trees"
body = json.dumps({
"text_prompts": [{"text": prompt_data}],
"cfg_scale": 10,
"seed": 20,
"steps": 50
})
modelId = "stability.stable-diffusion-xl-v1"
accept = "application/json"
contentType = "application/json"
try:

Launcher | bedrock_basics.ipynb
Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
bedrock_basics.ipynb
Getting_started_with_Converse...
README.md

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Trust your gut at the right time. While data and logic should steer most choices, sometimes relying on your instincts can uncover innovative solutions. Experience and expertise provide innate wisdom to supplement analytical frameworks.

Communicate decisions clearly and promptly. Explain the rationale you followed to key stakeholders. Get their support carrying out the plan by sharing responsibilities and next steps. Delaying announcements breeds uncertainty – people respect leaders who make prompt calls.

By grounding moves in comprehensive research, wisely balancing logic and intuition, and getting buy-in from others, leaders can establish a reputation for smart decision-making that moves the needle for the business in a sustainable way. Staying disciplined in this process will serve you well over the long haul.

[ ]:

[ ]:
```
# If you'd like to try your own prompt, edit this parameter!
prompt_data = """Human: Write me a blog about making strong business decisions as a leader.

Assistant:
"""
```

[ ]:
```
body = json.dumps({"prompt": prompt_data, "max_tokens_to_sample": 500})
modelId = "anthropic.claude-instant-v1"  # change this to use a different version from the model provider
accept = "application/json"
contentType = "application/json"

try:

    response = bedrock_runtime.invoke_model(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )
    response_body = json.loads(response.get("body").read())

    print(response_body.get("completion"))

except botocore.exceptions.ClientError as error:

    if error.response['Error']['Code'] == 'AccessDeniedException':
```

bedrock-finetune-20240712 | SageMaker Studio | Amazon S | SageMaker Studio | bedrock_basi... - JupyterLab | taipei 101 - Google Search

d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / Personal Studio

Launcher | bedrock_basics.ipynb

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ amazon-bedrock-workshop / 00_Prerequisites /

Name

- bedrock_basics.ipynb
- Getting_started_with_Converse...
- README.md

[ ]:
```
# If you'd like to try your own prompt, edit this parameter!
prompt_data = """Human: 請寫一篇關於到動物園旅遊的作文。500字為限。通俗易懂。

Assistant:
"""
```

[ ]:
```
body = json.dumps({"prompt": prompt_data, "max_tokens_to_sample": 500})
modelId = "anthropic.claude-instant-v1"  # change this to use a different version from the model provider
accept = "application/json"
contentType = "application/json"

try:

    response = bedrock_runtime.invoke_model(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )
    response_body = json.loads(response.get("body").read())

    print(response_body.get("completion"))

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")

    else:
        raise error
```

[ ]:

[ ]:

bedrock-finetune-20240712 | SageMaker Studio | Amazon S | SageMaker Studio | bedrock_basi... - JupyterLab | taipei 101 - Google Search

d-noqxdjm0rbmm.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/00_Prerequisites/bedrock_basics.ipynb

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / Personal Studio

Launcher | bedrock_basics.ipynb

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ amazon-bedrock-workshop / 00_Prerequisites /

Name

- bedrock_basics.ipynb
- Getting_started_with_Converse...
- README.md

```
    )
    response_body = json.loads(response.get("body").read())

    print(response_body.get("completion"))

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")

    else:
        raise error
```

到動物園的一天

上周六，我乘車和家人一起去了附近的動物園。出發前我們研究了一下動物園的地圖，劃定好要先看哪些動物區。

到達動物園後，我們先看了獅子區。幾隻大獅子正躺在陽光下打盹，看起來很懶洋洋的。接著我們去了猩猩區，那裡有幾隻猩猩在樹枝間跳躍，動作非常活潑。看到牠們用手指抓東西吃的樣子，我還以為牠們在玩耍。

中午時我們去了附近的小餐廳，吃過飯後繼續參觀。下午我最喜歡看的就是海豹表演了。海豹在水中游來游去，還會應聲做出各種動作，看起來很聰明。除此之外，我也第一次看到鬣狗和短吻鱷，牠們的外表真的很奇特。

一天很快就過去了。回家的路上，我們分享看見的動物，並討論下次還要去看哪些動物。這次在動物園的旅行，讓我更加了解不同動物，也玩得很開心。我希望下次能再去動物園，繼續發現更多新奇的動物。

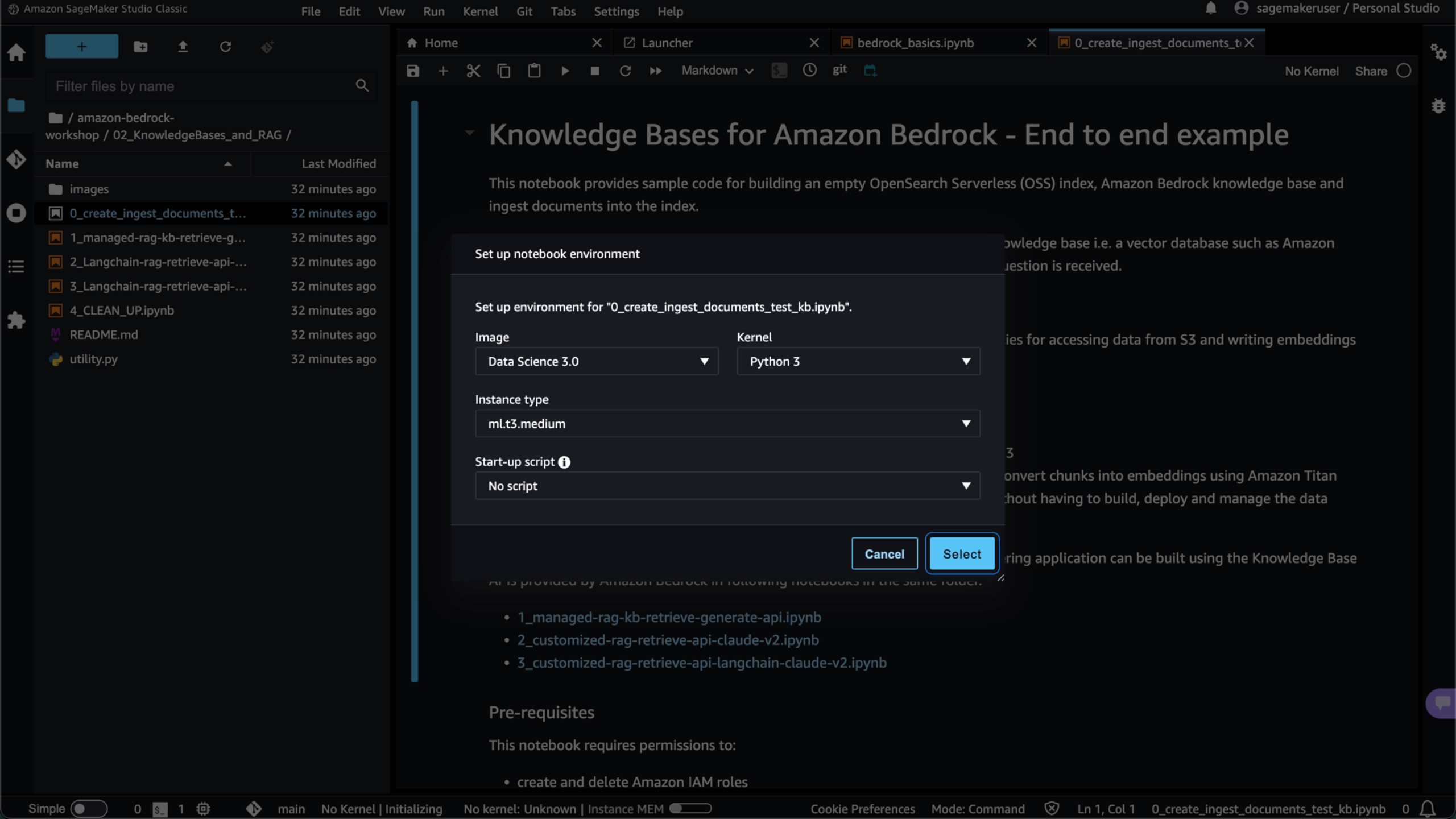
[ ]:

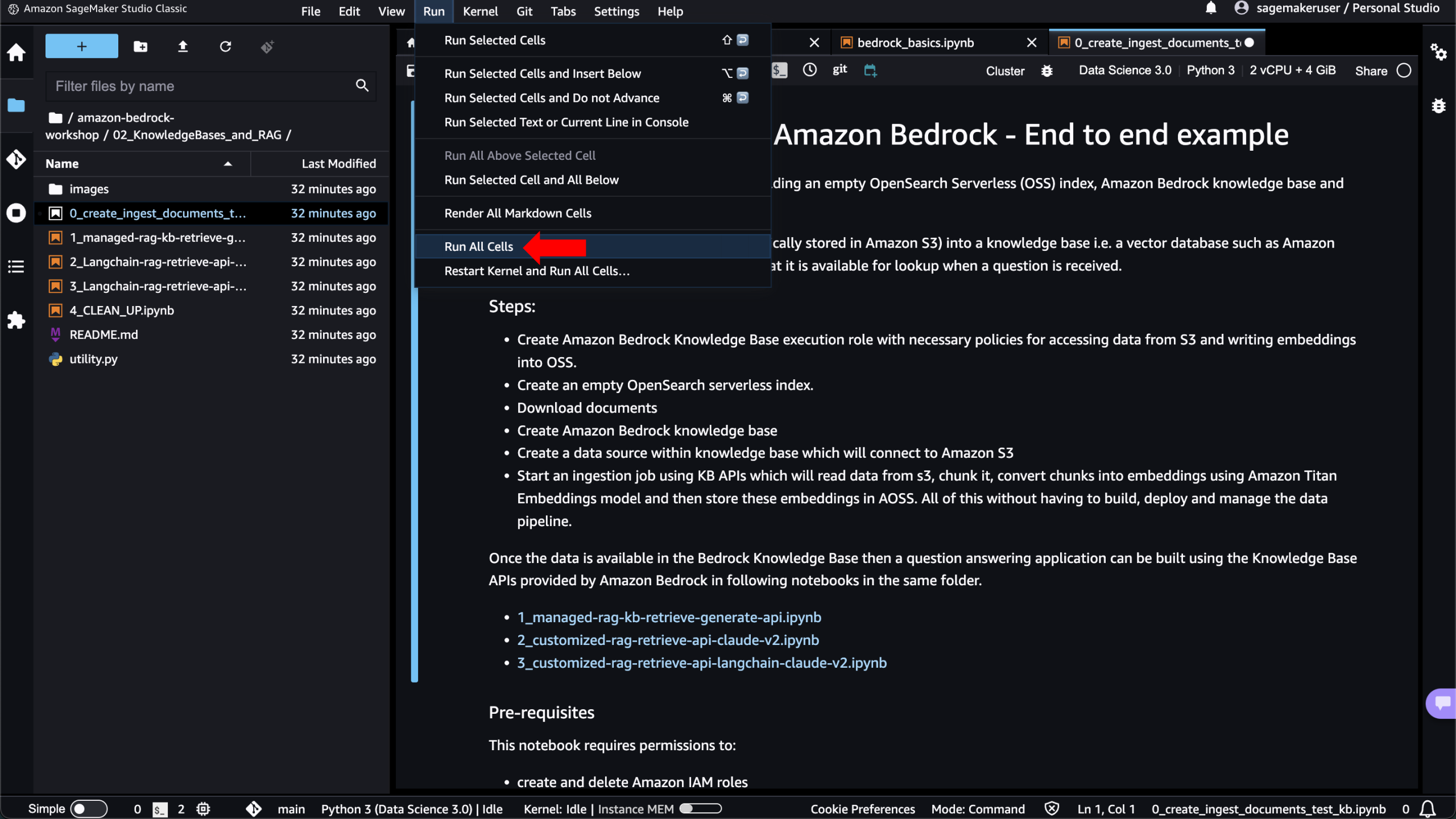
[ ]:

[ ]:

[ ]:

Bedrock KB Hands-On Lab


Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Filter files by name
/ amazon-bedrock-workshop / 02_KnowledgeBases_and_RAG /
Name
Last Modified
images 32 minutes ago
0_create_ingest_documents_t... 32 minutes ago
1_managed-rag-kb-retrieve-g... 32 minutes ago
2_Langchain-rag-retrieve-api-... 32 minutes ago
3_Langchain-rag-retrieve-api-... 32 minutes ago
4_CLEAN_UP.ipynb 32 minutes ago
README.md 32 minutes ago
utility.py 32 minutes ago
Home
Launcher
bedrock_basics.ipynb
0_create_ingest_documents_t
Markdown
git
No Kernel
Share
Knowledge Bases for Amazon Bedrock - End to end example
This notebook provides sample code for building an empty OpenSearch Serverless (OSS) index, Amazon Bedrock knowledge base and ingest documents into the index.
Set up notebook environment
Set up environment for "0_create_ingest_documents_test_kb.ipynb".
Image
Data Science 3.0
Kernel
Python 3
Instance type
ml.t3.medium
Start-up script
No script
Cancel
Select
1_managed-rag-kb-retrieve-generate-api.ipynb
2_customized-rag-retrieve-api-claude-v2.ipynb
3_customized-rag-retrieve-api-langchain-claude-v2.ipynb
Pre-requisites
This notebook requires permissions to:
create and delete Amazon IAM roles
Simple
main
No Kernel | Initializing
No kernel: Unknown | Instance MEM
Cookie Preferences
Mode: Command
Ln 1, Col 1
0_create_ingest_documents_test_kb.ipynb


Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Run Selected Cells
Run Selected Cells and Insert Below
Run Selected Cells and Do not Advance
Run Selected Text or Current Line in Console
Run All Above Selected Cell
Run Selected Cell and All Below
Render All Markdown Cells
Run All Cells
Restart Kernel and Run All Cells...
Filter files by name
/ amazon-bedrock-workshop / 02_KnowledgeBases_and_RAG /
Name
Last Modified
images 32 minutes ago
0_create_ingest_documents_t... 32 minutes ago
1_managed-rag-kb-retrieve-g... 32 minutes ago
2_Langchain-rag-retrieve-api-... 32 minutes ago
3_Langchain-rag-retrieve-api-... 32 minutes ago
4_CLEAN_UP.ipynb 32 minutes ago
README.md 32 minutes ago
utility.py 32 minutes ago
bedrock_basics.ipynb
0_create_ingest_documents_t
Cluster
Data Science 3.0 | Python 3 | 2 vCPU + 4 GiB
Share
Amazon Bedrock - End to end example
ding an empty OpenSearch Serverless (OSS) index, Amazon Bedrock knowledge base and
cally stored in Amazon S3) into a knowledge base i.e. a vector database such as Amazon
at it is available for lookup when a question is received.
Steps:
Create Amazon Bedrock Knowledge Base execution role with necessary policies for accessing data from S3 and writing embeddings into OSS.
Create an empty OpenSearch serverless index.
Download documents
Create Amazon Bedrock knowledge base
Create a data source within knowledge base which will connect to Amazon S3
Start an ingestion job using KB APIs which will read data from s3, chunk it, convert chunks into embeddings using Amazon Titan Embeddings model and then store these embeddings in AOSS. All of this without having to build, deploy and manage the data pipeline.
Once the data is available in the Bedrock Knowledge Base then a question answering application can be built using the Knowledge Base APIs provided by Amazon Bedrock in following notebooks in the same folder.
1_managed-rag-kb-retrieve-generate-api.ipynb
2_customized-rag-retrieve-api-claude-v2.ipynb
3_customized-rag-retrieve-api-langchain-claude-v2.ipynb
Pre-requisites
This notebook requires permissions to:
create and delete Amazon IAM roles
Simple 0 2 main Python 3 (Data Science 3.0) | Idle Kernel: Idle | Instance MEM Cookie Preferences Mode: Command Ln 1, Col 1 0_create_ingest_documents_test_kb.ipynb 0

Knowledge Bases for Amazon Bedrock - End to end example

This notebook provides sample code for building an empty OpenSearch Serverless (OSS) index, Amazon Bedrock knowledge base and ingest documents into the index.

A data pipeline that ingests documents (typically stored in Amazon S3) into a knowledge base i.e. a vector database such as Amazon OpenSearch Service Serverless (AOSS) so that it is available for lookup when a question is received.

Steps:

- Create Amazon Bedrock Knowledge Base execution role with necessary policies for accessing data from S3 and writing embeddings into OSS.
- Create an empty OpenSearch serverless index.
- Download documents
- Create Amazon Bedrock knowledge base
- Create a data source within knowledge base which will connect to Amazon S3
- Start an ingestion job using KB APIs which will read data from s3, chunk it, convert chunks into embeddings using Amazon Titan Embeddings model and then store these embeddings in AOSS. All of this without having to build, deploy and manage the data pipeline.

Once the data is available in the Bedrock Knowledge Base then a question answering application can be built using the Knowledge Base APIs provided by Amazon Bedrock in following notebooks in the same folder.

- 1_managed-rag-kb-retrieve-generate-api.ipynb
- 2_customized-rag-retrieve-api-claude-v2.ipynb
- 3_customized-rag-retrieve-api-langchain-claude-v2.ipynb

Pre-requisites

This notebook requires permissions to:

- create and delete Amazon IAM roles
- create, update and delete Amazon S3 buckets
- access Amazon Bedrock
- access to Amazon OpenSearch Serverless

Setup

Before running the rest of this notebook, you'll need to run the cells below to (ensure necessary libraries are installed and) connect to Bedrock.

[2]:
```
%pip install -U opensearch-py==2.3.1
%pip install -U boto3==1.33.2
%pip install -U retrying==1.3.4
```

```
Collecting opensearch-py==2.3.1
  Downloading opensearch_py-2.3.1-py2.py3-none-any.whl.metadata (6.9 kB)
Collecting urllib3<2,>=1.21.1 (from opensearch-py==2.3.1)
  Downloading urllib3-1.26.18-py2.py3-none-any.whl.metadata (48 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 48.9/48.9 kB 511.6 kB/s eta 0:00:00a 0:00:01
Requirement already satisfied: requests<3.0.0,>=2.4.0 in /opt/conda/lib/python3.10/site-packages (from opensearch-py==2.3.1) (2.31.0)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages (from opensearch-py==2.3.1) (1.16.0)
Requirement already satisfied: python-dateutil in /opt/conda/lib/python3.10/site-packages (from opensearch-py==2.3.1) (2.9.0.post0)
Requirement already satisfied: certifi>=2022.12.07 in /opt/conda/lib/python3.10/site-packages (from opensearch-py==2.3.1) (2023.11.17)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/python3.10/site-packages (from requests<3.0.0,>=2.4.0->opensearch-py==2.3.1) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-packages (from requests<3.0.0,>=2.4.0->opensearch-py==2.3.1) (3.3)
Downloading opensearch_py-2.3.1-py2.py3-none-any.whl (327 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 327.3/327.3 kB 3.1 MB/s eta 0:00:00ta 0:00:01
Downloading urllib3-1.26.18-py2.py3-none-any.whl (143 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 143.8/143.8 kB 1.7 MB/s eta 0:00:00:00:01
Installing collected packages: urllib3, opensearch-py
  Attempting uninstall: urllib3
    Found existing installation: urllib3 2.2.1
    Uninstalling urllib3-2.2.1:
      Successfully uninstalled urllib3-2.2.1
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.
distributed 2022.7.0 requires tornado<6.2,>=6.0.3, but you have tornado 6.4 which is incompatible.
Successfully installed opensearch-py-2.3.1 urllib3-1.26.18
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv

[notice] A new release of pip is available: 23.3.1 -> 24.0
[notice] To update, run: pip install --upgrade pip
Note: you may need to restart the kernel to use updated packages.
```

[3]:
```
# restart kernel
from IPython.core.display import HTML
HTML("<script>Jupyter.notebook.kernel.restart()</script>")
```

[3]:

[4]:
```
import warnings
warnings.filterwarnings('ignore')
```

[5]:
```
import json
import os
import boto3
import pprint
from utility import create_bedrock_execution_role, create_oss_policy_attach_bedrock_execution_role, create_policies_in_oss
import random
from retrying import retry
suffix = random.randrange(200, 900)

sts_client = boto3.client('sts')
boto3_session = boto3.session.Session()
region_name = boto3_session.region_name
bedrock_agent_client = boto3_session.client('bedrock-agent', region_name=region_name)
service = 'aoss'
s3_client = boto3.client('s3')
account_id = sts_client.get_caller_identity()["Account"]
s3_suffix = f"{region_name}-{account_id}"
bucket_name = f'bedrock-kb-{s3_suffix}' # replace it with your bucket name.
pp = pprint.PrettyPrinter(indent=2)
```

[6]:
```
# Create S3 bucket for knowledge base data source
s3bucket = s3_client.create_bucket(
    Bucket=bucket_name,
    CreateBucketConfiguration={ 'LocationConstraint': region_name }
)
```

Create a vector store - OpenSearch Serverless index

Create a vector store - OpenSearch Serverless index

Step 1 - Create OSS policies and collection

Firt of all we have to create a vector store. In this section we will use *Amazon OpenSerach serverless.*

Amazon OpenSearch Serverless is a serverless option in Amazon OpenSearch Service. As a developer, you can use OpenSearch Serverless to run petabyte-scale workloads without configuring, managing, and scaling OpenSearch clusters. You get the same interactive millisecond response times as OpenSearch Service with the simplicity of a serverless environment. Pay only for what you use by automatically scaling resources to provide the right amount of capacity for your application—without impacting data ingestion.

[7]:
```
import boto3
import time
vector_store_name = f'bedrock-sample-rag-{suffix}'
index_name = f"bedrock-sample-rag-index-{suffix}"
aoss_client = boto3_session.client('opensearchserverless')
bedrock_kb_execution_role = create_bedrock_execution_role(bucket_name=bucket_name)
bedrock_kb_execution_role_arn = bedrock_kb_execution_role['Role']['Arn']
```

[8]:
```
# create security, network and data access policies within OSS
encryption_policy, network_policy, access_policy = create_policies_in_oss(vector_store_name=vector_store_name,
                      aoss_client=aoss_client,
                      bedrock_kb_execution_role_arn=bedrock_kb_execution_role_arn)
collection = aoss_client.create_collection(name=vector_store_name,type='VECTORSEARCH')
```

[9]:
```
pp.pprint(collection)
```

```
{ 'ResponseMetadata': { 'HTTPHeaders': { 'connection': 'keep-alive',
                                         'content-length': '314',
                                         'content-type': 'application/x-amz-json-1.0',
                                         'date': 'Tue, 26 Mar 2024 06:14:53 '
                                                 'GMT',
                                         'x-amzn-requestid': '6e67c99d-c29e-414b-9257-2b11ea0b1389'},
                        'HTTPStatusCode': 200,
                        'RequestId': '6e67c99d-c29e-414b-9257-2b11ea0b1389',
                        'RetryAttempts': 0},
  'createCollectionDetail': { 'arn': 'arn:aws:aoss:us-west-2:655380892071:collection/hnbf0ml4o1djqrjy1rcg',
                              'createdDate': 1711433692851,
```

[10]:
```
collection_id = collection['createCollectionDetail']['id']
host = collection_id + '.' + region_name + '.aoss.amazonaws.com'
print(host)
```

```
hnbf0ml4o1djqrjy1rcg.us-west-2.aoss.amazonaws.com
```

[11]:
```
# wait for collection creation
response = aoss_client.batch_get_collection(names=[vector_store_name])
# Periodically check collection status
while (response['collectionDetails'][0]['status']) == 'CREATING':
    print('Creating collection...')
    time.sleep(30)
    response = aoss_client.batch_get_collection(names=[vector_store_name])
print('\nCollection successfully created:')
print(response["collectionDetails"])
```

```
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...
Creating collection...

Collection successfully created:
[{'arn': 'arn:aws:aoss:us-west-2:655380892071:collection/hnbf0ml4o1djqrjy1rcg', 'collectionEndpoint': 'https://hnbf0ml4o1djqrjy1rcg.us-west-2.aoss.amazonaws.com', 'createdDate': 1711433692851, 'dashboardEndpoint': 'https://hnbf0ml4o1djqrjy1rcg.us-west-2.aoss.amazonaws.com/_dashboards', 'id': 'hnbf0ml4o1djqrjy1rcg', 'kmsKeyArn': 'auto', 'lastModifiedDate': 1711434081655, 'name': 'bedrock-sample-rag-408', 'standbyReplicas': 'ENABLED', 'status': 'ACTIVE', 'type': 'VECTORSEARCH'}]
```

[12]:
```
# create oss policy and attach it to Bedrock execution role
create_oss_policy_attach_bedrock_execution_role(collection_id=collection_id,
                                                bedrock_kb_execution_role=bedrock_kb_execution_role)
```

```
Opensearch serverless arn:  arn:aws:iam::655380892071:policy/AmazonBedrockOSSPolicyForKnowledgeBase_207
```

Step 2 - Create vector index

[13]:

```
from opensearchpy import OpenSearch, RequestsHttpConnection, AWSV4SignerAuth
credentials = boto3.Session().get_credentials()
awsauth = auth = AWSV4SignerAuth(credentials, region_name, service)

index_name = f"bedrock-sample-index-{suffix}"
body_json = {
   "settings": {
      "index.knn": "true",
       "number_of_shards": 1,
       "knn.algo_param.ef_search": 512,
       "number_of_replicas": 0,
   },
   "mappings": {
      "properties": {
         "vector": {
            "type": "knn_vector",
            "dimension": 1536,
             "method": {
                 "name": "hnsw",
                 "engine": "nmslib",
                 "space_type": "cosinesimil",
                 "parameters": {
                     "ef_construction": 512,
                     "m": 16
                 },
             },
         },
         "text": {
            "type": "text"
         },
         "text-metadata": {
            "type": "text"         }
      }
   }
}
# Build the OpenSearch client
oss_client = OpenSearch(
```

[14]:
```
# Create index
response = oss_client.indices.create(index=index_name, body=json.dumps(body_json))
print('\nCreating index:')
print(response)
time.sleep(60) # index creation can take up to a minute
```

```
Creating index:
{'acknowledged': True, 'shards_acknowledged': True, 'index': 'bedrock-sample-index-408'}
```

Download data

[15]:
```
# Download and prepare dataset
!mkdir -p ./data

from urllib.request import urlretrieve
urls = [
    'https://s2.q4cdn.com/299287126/files/doc_financials/2023/ar/2022-Shareholder-Letter.pdf',
    'https://s2.q4cdn.com/299287126/files/doc_financials/2022/ar/2021-Shareholder-Letter.pdf',
    'https://s2.q4cdn.com/299287126/files/doc_financials/2021/ar/Amazon-2020-Shareholder-Letter-and-1997-Shareholder-Letter.pdf',
    'https://s2.q4cdn.com/299287126/files/doc_financials/2020/ar/2019-Shareholder-Letter.pdf'
]

filenames = [
    'AMZN-2022-Shareholder-Letter.pdf',
    'AMZN-2021-Shareholder-Letter.pdf',
    'AMZN-2020-Shareholder-Letter.pdf',
    'AMZN-2019-Shareholder-Letter.pdf'
]

data_root = "./data/"

for idx, url in enumerate(urls):
    file_path = data_root + filenames[idx]
    urlretrieve(url, file_path)
```

Upload data to S3 Bucket

Upload data to S3 Bucket

[16]:
```
# Upload data to s3
s3_client = boto3.client("s3")
def uploadDirectory(path,bucket_name):
        for root,dirs,files in os.walk(path):
            for file in files:
                s3_client.upload_file(os.path.join(root,file),bucket_name,file)

uploadDirectory(data_root, bucket_name)
```

Create Knowledge Base ¶

Steps:

- initialize Open search serverless configuration which will include collection ARN, index name, vector field, text field and metadata field.
- initialize chunking strategy, based on which KB will split the documents into pieces of size equal to the chunk size mentioned in the `chunkingStrategyConfiguration` .
- initialize the s3 configuration, which will be used to create the data source object later.
- initialize the Titan embeddings model ARN, as this will be used to create the embeddings for each of the text chunks.

[17]:
```
opensearchServerlessConfiguration = {
            "collectionArn": collection["createCollectionDetail"]['arn'],
            "vectorIndexName": index_name,
            "fieldMapping": {
                "vectorField": "vector",
                "textField": "text",
                "metadataField": "text-metadata"
            }
        }

chunkingStrategyConfiguration = {
    "chunkingStrategy": "FIXED_SIZE",
    "fixedSizeChunkingConfiguration": {
        "maxTokens": 512,
        "overlapPercentage": 20
    }
```

Create Knowledge Base

Steps:

- initialize Open search serverless configuration which will include collection ARN, index name, vector field, text field and metadata field.
- initialize chunking strategy, based on which KB will split the documents into pieces of size equal to the chunk size mentioned in the `chunkingStrategyConfiguration` .
- initialize the s3 configuration, which will be used to create the data source object later.
- initialize the Titan embeddings model ARN, as this will be used to create the embeddings for each of the text chunks.

[17]:
```
opensearchServerlessConfiguration = {
            "collectionArn": collection["createCollectionDetail"]['arn'],
            "vectorIndexName": index_name,
            "fieldMapping": {
                "vectorField": "vector",
                "textField": "text",
                "metadataField": "text-metadata"
            }
        }

chunkingStrategyConfiguration = {
    "chunkingStrategy": "FIXED_SIZE",
    "fixedSizeChunkingConfiguration": {
        "maxTokens": 512,
        "overlapPercentage": 20
    }
}

s3Configuration = {
    "bucketArn": f"arn:aws:s3:::{bucket_name}",
    # "inclusionPrefixes":["*.*"] # you can use this if you want to create a KB using data within s3 prefixes.
}

embeddingModelArn = f"arn:aws:bedrock:{region_name}::foundation-model/amazon.titan-embed-text-v1"

name = f"bedrock-sample-knowledge-base-{suffix}"
description = "Amazon shareholder letter knowledge base."
roleArn = bedrock_kb_execution_role_arn
```

[18]:
```
# Create a KnowledgeBase
from retrying import retry

@retry(wait_random_min=1000, wait_random_max=2000,stop_max_attempt_number=7)
def create_knowledge_base_func():
    create_kb_response = bedrock_agent_client.create_knowledge_base(
        name = name,
        description = description,
        roleArn = roleArn,
        knowledgeBaseConfiguration = {
            "type": "VECTOR",
            "vectorKnowledgeBaseConfiguration": {
                "embeddingModelArn": embeddingModelArn
            }
        },
        storageConfiguration = {
            "type": "OPENSEARCH_SERVERLESS",
            "opensearchServerlessConfiguration":opensearchServerlessConfiguration
        }
    )
    return create_kb_response["knowledgeBase"]
```

[19]:
```
try:
    kb = create_knowledge_base_func()
except Exception as err:
    print(f"{err=}, {type(err)=}")
```

[20]:
```
pp.pprint(kb)
```

```
{ 'createdAt': datetime.datetime(2024, 3, 26, 6, 23, 28, 437073, tzinfo=tzlocal()),
  'description': 'Amazon shareholder letter knowledge base.',
  'knowledgeBaseArn': 'arn:aws:bedrock:us-west-2:655380892071:knowledge-base/CRXSFPF4H8',
  'knowledgeBaseConfiguration': { 'type': 'VECTOR',
                                  'vectorKnowledgeBaseConfiguration': { 'embeddingModelArn': 'arn:aws:bedrock:us-west-2::foundation-model/amazon.titan-embed-tex
t-v1'}},
  'knowledgeBaseId': 'CRXSFPF4H8',
  'name': 'bedrock-sample-knowledge-base-408',
  'roleArn': 'arn:aws:iam::655380892071:role/AmazonBedrockExecutionRoleForKnowledgeBase_207',
  'status': 'CREATING',
```

[21]:
```
# Get KnowledgeBase
get_kb_response = bedrock_agent_client.get_knowledge_base(knowledgeBaseId = kb['knowledgeBaseId'])
```

Next we need to create a data source, which will be associated with the knowledge base created above. Once the data source is ready, we can then start to ingest the documents.

[22]:
```
# Create a DataSource in KnowledgeBase
create_ds_response = bedrock_agent_client.create_data_source(
    name = name,
    description = description,
    knowledgeBaseId = kb['knowledgeBaseId'],
    dataSourceConfiguration = {
        "type": "S3",
        "s3Configuration":s3Configuration
    },
    vectorIngestionConfiguration = {
        "chunkingConfiguration": chunkingStrategyConfiguration
    }
)
ds = create_ds_response["dataSource"]
pp.pprint(ds)
```

```
{ 'createdAt': datetime.datetime(2024, 3, 26, 6, 23, 29, 351908, tzinfo=tzlocal()),
  'dataSourceConfiguration': { 's3Configuration': { 'bucketArn': 'arn:aws:s3:::bedrock-kb-us-west-2-655380892071'},
                               'type': 'S3'},
  'dataSourceId': 'CQKVIMQBBD',
  'description': 'Amazon shareholder letter knowledge base.',
  'knowledgeBaseId': 'CRXSFPF4H8',
  'name': 'bedrock-sample-knowledge-base-408',
  'status': 'AVAILABLE',
  'updatedAt': datetime.datetime(2024, 3, 26, 6, 23, 29, 351908, tzinfo=tzlocal()),
  'vectorIngestionConfiguration': { 'chunkingConfiguration': { 'chunkingStrategy': 'FIXED_SIZE',
                                                               'fixedSizeChunkingConfiguration': { 'maxTokens': 512,
                                                                                                   'overlapPercentage': 20}}}}
```

[23]:
```
# Get DataSource
bedrock_agent_client.get_data_source(knowledgeBaseId = kb['knowledgeBaseId'], dataSourceId = ds["dataSourceId"])
```

[23]:
```
{'ResponseMetadata': {'RequestId': '146c0a9e-b512-4e0c-853e-24a2c22ab687',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Tue, 26 Mar 2024 06:23:29 GMT',
```

[23]:
```
# Get DataSource
bedrock_agent_client.get_data_source(knowledgeBaseId = kb['knowledgeBaseId'], dataSourceId = ds["dataSourceId"])
```

[23]:
```
{'ResponseMetadata': {'RequestId': '146c0a9e-b512-4e0c-853e-24a2c22ab687',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Tue, 26 Mar 2024 06:23:29 GMT',
   'content-type': 'application/json',
   'content-length': '573',
   'connection': 'keep-alive',
   'x-amzn-requestid': '146c0a9e-b512-4e0c-853e-24a2c22ab687',
   'x-amz-apigw-id': 'VOV7RFSDPHcEgXg=',
   'x-amzn-trace-id': 'Root=1-660269e1-14b9e1aa5ac67dd0751f7794'},
  'RetryAttempts': 0},
 'dataSource': {'knowledgeBaseId': 'CRXSFPF4H8',
  'dataSourceId': 'CQKVIMQBBD',
  'name': 'bedrock-sample-knowledge-base-408',
  'status': 'AVAILABLE',
  'description': 'Amazon shareholder letter knowledge base.',
  'dataSourceConfiguration': {'type': 'S3',
   's3Configuration': {'bucketArn': 'arn:aws:s3:::bedrock-kb-us-west-2-655380892071'}},
  'vectorIngestionConfiguration': {'chunkingConfiguration': {'chunkingStrategy': 'FIXED_SIZE',
    'fixedSizeChunkingConfiguration': {'maxTokens': 512,
     'overlapPercentage': 20}}},
  'createdAt': datetime.datetime(2024, 3, 26, 6, 23, 29, 351908, tzinfo=tzlocal()),
  'updatedAt': datetime.datetime(2024, 3, 26, 6, 23, 29, 351908, tzinfo=tzlocal())}}
```

Start ingestion job

Once the KB and data source is created, we can start the ingestion job. During the ingestion job, KB will fetch the documents in the data source, pre-process it to extract text, chunk it based on the chunking size provided, create embeddings of each chunk and then write it to the vector database, in this case OSS.

[24]:
```
# Start an ingestion job
start_job_response = bedrock_agent_client.start_ingestion_job(knowledgeBaseId = kb['knowledgeBaseId'], dataSourceId = ds["dataSourceId"])
```

[25]:
```
job = start_job_response["ingestionJob"]
pp.pprint(job)
```

```
{ 'dataSourceId': 'CQKVIMQBBD',
```

Start ingestion job

Once the KB and data source is created, we can start the ingestion job. During the ingestion job, KB will fetch the documents in the data source, pre-process it to extract text, chunk it based on the chunking size provided, create embeddings of each chunk and then write it to the vector database, in this case OSS.

[24]:
```
# Start an ingestion job
start_job_response = bedrock_agent_client.start_ingestion_job(knowledgeBaseId = kb['knowledgeBaseId'], dataSourceId = ds["dataSourceId"])
```

[25]:
```
job = start_job_response["ingestionJob"]
pp.pprint(job)
```

```
{ 'dataSourceId': 'CQKVIMQBBD',
  'ingestionJobId': 'KLQYCO0QKN',
  'knowledgeBaseId': 'CRXSFPF4H8',
  'startedAt': datetime.datetime(2024, 3, 26, 6, 23, 31, 575806, tzinfo=tzlocal()),
  'statistics': { 'numberOfDocumentsDeleted': 0,
                  'numberOfDocumentsFailed': 0,
                  'numberOfDocumentsScanned': 0,
                  'numberOfModifiedDocumentsIndexed': 0,
                  'numberOfNewDocumentsIndexed': 0},
  'status': 'STARTING',
  'updatedAt': datetime.datetime(2024, 3, 26, 6, 23, 31, 575806, tzinfo=tzlocal())}
```

[26]:
```
# Get job
while(job['status']!='COMPLETE' ):
  get_job_response = bedrock_agent_client.get_ingestion_job(
      knowledgeBaseId = kb['knowledgeBaseId'],
        dataSourceId = ds["dataSourceId"],
        ingestionJobId = job["ingestionJobId"]
  )
  job = get_job_response["ingestionJob"]
pp.pprint(job)
time.sleep(40)
```

```
{ 'dataSourceId': 'CQKVIMQBBD',
  'ingestionJobId': 'KLQYCO0QKN',
  'knowledgeBaseId': 'CRXSFPF4H8',
  'startedAt': datetime.datetime(2024, 3, 26, 6, 23, 31, 575806, tzinfo=tzlocal()),
  'statistics': { 'numberOfDocumentsDeleted': 0,
                  'numberOfDocumentsFailed': 0,
```

[27]:
```
kb_id = kb["knowledgeBaseId"]
pp.pprint(kb_id)
```

```
'CRXSFPF4H8'
```

[28]:
```
%store kb_id
```

```
Stored 'kb_id' (str)
```

Test the knowledge base

Using RetrieveAndGenerate API

Behind the scenes, RetrieveAndGenerate API converts queries into embeddings, searches the knowledge base, and then augments the foundation model prompt with the search results as context information and returns the FM-generated response to the question. For multi-turn conversations, Knowledge Bases manage short-term memory of the conversation to provide more contextual results.

The output of the RetrieveAndGenerate API includes the generated response, source attribution as well as the retrieved text chunks.

[34]:
```
# try out KB using RetrieveAndGenerate API
bedrock_agent_runtime_client = boto3.client("bedrock-agent-runtime", region_name=region_name)
model_id = "anthropic.claude-instant-v1" # try with both claude instant as well as claude-v2. for claude v2 - "anthropic.claude-v2"
model_arn = f'arn:aws:bedrock:{region_name}::foundation-model/{model_id}'
```

[35]:
```
time.sleep(5)
query = "What is Amazon's doing in the field of generative AI?"
response = bedrock_agent_runtime_client.retrieve_and_generate(
    input={
        'text': query
    },
    retrieveAndGenerateConfiguration={
        'type': 'KNOWLEDGE_BASE',
        'knowledgeBaseConfiguration': {
            'knowledgeBaseId': kb_id,
            'modelArn': model_arn
        }
    },
```

Test the knowledge base

Using RetrieveAndGenerate API

Behind the scenes, RetrieveAndGenerate API converts queries into embeddings, searches the knowledge base, and then augments the foundation model prompt with the search results as context information and returns the FM-generated response to the question. For multi-turn conversations, Knowledge Bases manage short-term memory of the conversation to provide more contextual results.

The output of the RetrieveAndGenerate API includes the generated response, source attribution as well as the retrieved text chunks.

[34]:
```
# try out KB using RetrieveAndGenerate API
bedrock_agent_runtime_client = boto3.client("bedrock-agent-runtime", region_name=region_name)
model_id = "anthropic.claude-instant-v1" # try with both claude instant as well as claude-v2. for claude v2 - "anthropic.claude-v2"
model_arn = f'arn:aws:bedrock:{region_name}::foundation-model/{model_id}'
```

[35]:
```
time.sleep(5)
query = "What is Amazon's doing in the field of generative AI?"
response = bedrock_agent_runtime_client.retrieve_and_generate(
    input={
        'text': query
    },
    retrieveAndGenerateConfiguration={
        'type': 'KNOWLEDGE_BASE',
        'knowledgeBaseConfiguration': {
            'knowledgeBaseId': kb_id,
            'modelArn': model_arn
        }
    },
)

generated_text = response['output']['text']
pp.pprint(generated_text)
```

```
('Amazon has been working on their own large language models (LLMs) for '
 'generative AI and believes it will transform and improve virtually every '
 'customer experience. They are continuing to invest substantially in these '
 'models across all of their consumer, seller, brand, and creator experiences. '
 "Additionally, as they've done for years in AWS, they're democratizing this "
```

[36]:
```
## print out the source attribution/citations from the original documents to see if the response generated belongs to the context.
citations = response["citations"]
contexts = []
for citation in citations:
    retrievedReferences = citation["retrievedReferences"]
    for reference in retrievedReferences:
        contexts.append(reference["content"]["text"])

pp.pprint(contexts)
```

```
[ 'This shift was driven by several factors, including access to higher '
  'volumes of compute capacity at lower prices than was ever available. Amazon '
  'has been using machine learning extensively for 25 years, employing it in '
  'everything from personalized ecommerce recommendations, to fulfillment '
  'center pick paths, to drones for Prime Air, to Alexa, to the many machine '
  'learning services AWS offers (where AWS has the broadest machine learning '
  'functionality and customer base of any cloud provider). More recently, a '
  'newer form of machine learning, called Generative AI, has burst onto the '
  'scene and promises to significantly accelerate machine learning adoption. '
  'Generative AI is based on very Large Language Models (trained on up to '
  'hundreds of billions of parameters, and growing), across expansive '
  'datasets, and has radically general and broad recall and learning '
  'capabilities. We have been working on our own LLMs for a while now, believe '
  'it will transform and improve virtually every customer experience, and will '
  'continue to invest substantially in these models across all of our '
  'consumer, seller, brand, and creator experiences. Additionally, as we've '
  'done for years in AWS, we're democratizing this technology so companies of '
  'all sizes can leverage Generative AI. AWS is offering the most '
  'price-performant machine learning chips in Trainium and Inferentia so small '
  'and large companies can afford to train and run their LLMs in production. '
  'We enable companies to choose from various LLMs and build applications with '
  'all of the AWS security, privacy and other features that customers are '
  'accustomed to using. And, we're delivering applications like AWS's '
  'CodeWhisperer, which revolutionizes       developer productivity by '
  'generating code suggestions in real time. I could write an entire letter on '
  'LLMs and Generative AI as I think they will be that transformative, but '
  'I'll leave that for a future letter. Let's just say that LLMs and '
  'Generative AI are going to be a big deal for customers, our shareholders, '
  'and Amazon.   So, in closing, I'm optimistic that we'll emerge from this '
  'challenging macroeconomic time in a stronger position than when we entered '
  'it. There are several reasons for it and I've mentioned many of them above. '
```

Retrieve API

Retrieve API converts user queries into embeddings, searches the knowledge base, and returns the relevant results, giving you more control to build custom workflows on top of the semantic search results. The output of the Retrieve API includes the the retrieved text chunks, the location type and URI of the source data, as well as the relevance scores of the retrievals.

[37]:
```
# retreive api for fetching only the relevant context.
relevant_documents = bedrock_agent_runtime_client.retrieve(
    retrievalQuery= {
        'text': query
    },
    knowledgeBaseId=kb_id,
    retrievalConfiguration= {
        'vectorSearchConfiguration': {
            'numberOfResults': 3 # will fetch top 3 documents which matches closely with the query.
        }
    }
)
```

[38]:
```
pp.pprint(relevant_documents["retrievalResults"])
```

```
[ { 'content': { 'text': 'This shift was driven by several factors, including '
                         'access to higher volumes of compute capacity at '
                         'lower prices than was ever available. Amazon has '
                         'been using machine learning extensively for 25 '
                         'years, employing it in everything from personalized '
                         'ecommerce recommendations, to fulfillment center '
                         'pick paths, to drones for Prime Air, to Alexa, to '
                         'the many machine learning services AWS offers (where '
                         'AWS has the broadest machine learning functionality '
                         'and customer base of any cloud provider). More '
                         'recently, a newer form of machine learning, called '
                         'Generative AI, has burst onto the scene and promises '
                         'to significantly accelerate machine learning '
                         'adoption. Generative AI is based on very Large '
                         'Language Models (trained on up to hundreds of '
                         'billions of parameters, and growing), across '
                         'expansive datasets, and has radically general and '
                         'broad recall and learning capabilities. We have been '
                         'working on our own LLMs for a while now, believe it '
```

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2 | New Tab

台北 捷運 辦法 pdf

台北 捷運 辦法 pdf - Google Search

台北捷運圖pdf

台北捷運路線圖pdf

台北捷運 報帳

Employee Stock P...

All Bookmarks

Search Google or type a URL

Add shortcut

台北 捷運 辦法 pdf

All Images News Videos Shopping Books Web More Tools

臺北大眾捷運股份有限公司
https://web.metro.taipei › QRCode PDF

臺北捷運系統旅客須知

Jun 19, 2024 — 同警察人員強制或護送其離開站、車或大眾捷運系統區域：. （一）違反法令、公共秩序、善良風俗或本公司旅客運送章則等各項. 規定。 （二）有明顯傷害他人 ...

臺北大眾捷運股份有限公司
https://web.metro.taipei › TaipeiMetroGuides › T... PDF

台北捷運營運資訊簡介

攜帶自行車 單程票 開放時段:假日全天,平日10至16時、22時至營運結束。 除文湖線各車站、淡水站、台北車站、忠孝新生站、忠孝 復興站、南京復興站、大安站以外,餘85個車站 ...

2 pages

臺北大眾捷運股份有限公司
https://web.metro.taipei › QRCode PDF

臺北捷運系統營運服務規約

一、營運時間：06:00~24:00. 二、服務內容. （一）平均班距. 1.主線：. （1）尖峰時段：. A.淡水站－象山站、松山站－新店站、南港展覽館站－.

臺北市政府全球資訊網
https://www.laws.taipei.gov.tw › caseatt › 臺北市... PDF

臺北市大眾捷運系統與地下街設施移設及連通申請自治條例

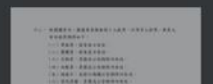
第四條本自治條例用詞定義如下：. 一移設：將原計畫(指於都市計畫變更書圖註明者)或已. 興建於都市計畫道路上之捷運或地下街設施改設於. 申請人或都市計畫規定之建築基地內 ...

7 pages


1


2


3



臺北捷運系統旅客須知

中華民國八十五年三月二十八日公告修正
中華民國八十七年十二月二十四日公告修正
中華民國八十八年十一月十日公告修正
中華民國九十一年八月二十日公告修正
中華民國九十三年十二月二十四日公告修正
中華民國九十七年三月二十八日公告修正
中華民國九十七年六月三日公告修正
中華民國九十八年三月十二日公告修正
中華民國一〇三年五月二十七日公告修正
中華民國一〇四年三月十八日公告修正
中華民國一〇四年十二月十一日公告修正
中華民國一〇六年五月十二日公告修正
中華民國一〇六年十月五日公告修正
中華民國一〇七年十月二十六日公告修正
中華民國一一一年三月二十九日公告修正
中華民國一一二年二月七日公告修正
中華民國一一二年十一月八日公告修正
中華民國一一三年六月十三日公告修正

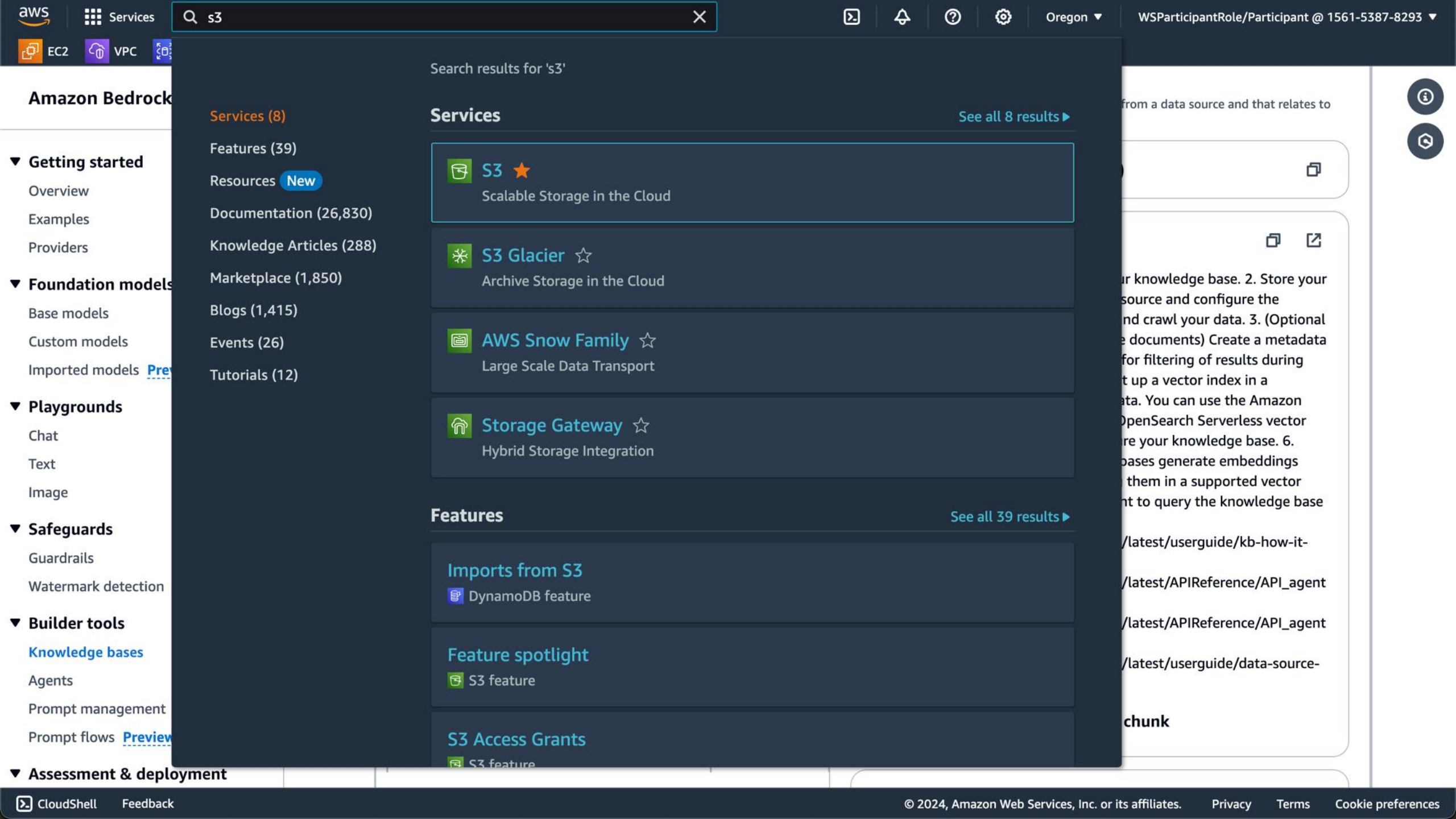
壹、一般規定

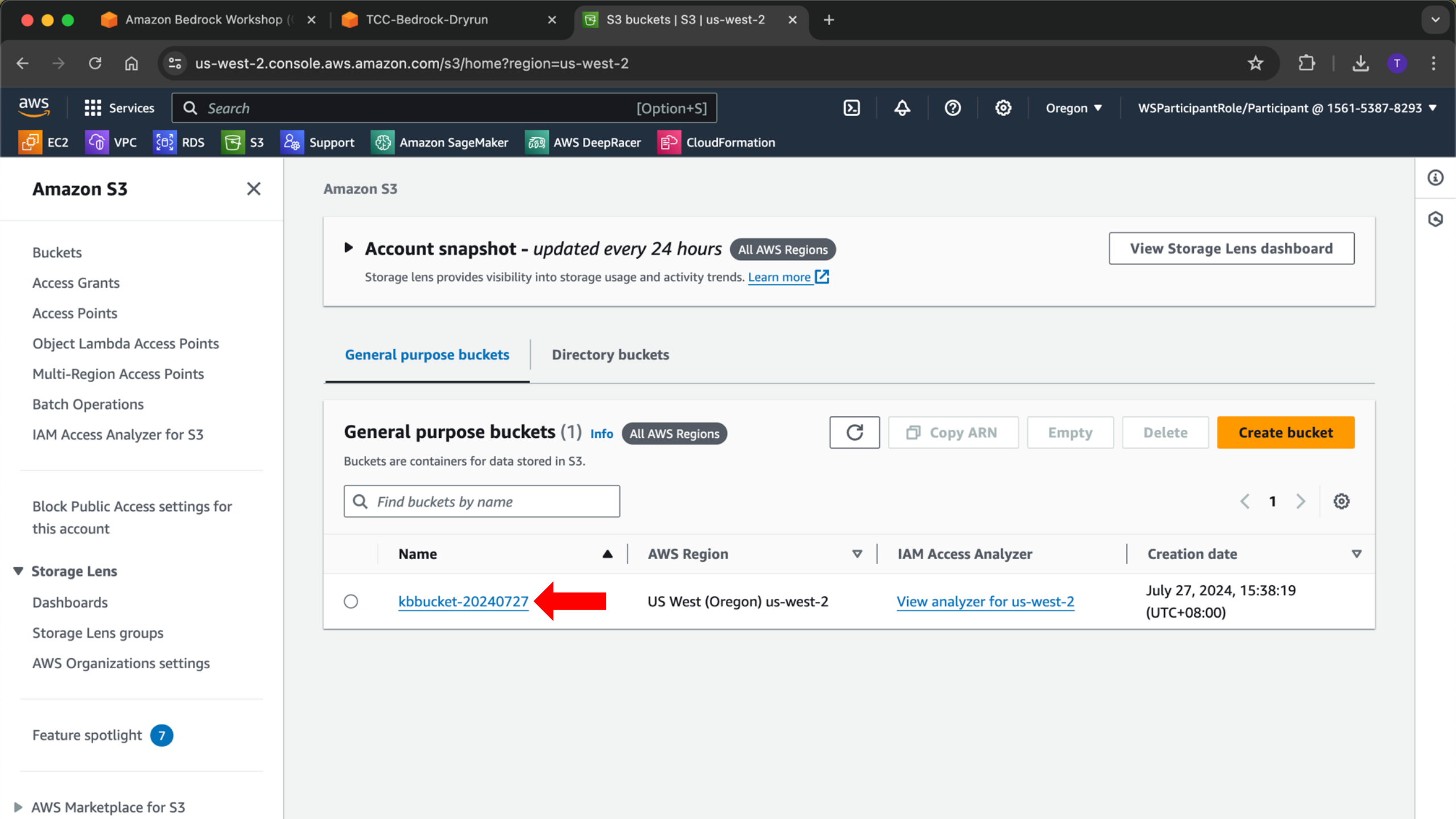
一、臺北大眾捷運股份有限公司(以下簡稱本公司)為提供捷運系統旅客安全、可靠、便捷、舒適之服務，特依「臺北市大眾捷運系統旅客運送自治條例」規定訂定本須知，並於車站公告，變更或調整時亦同。

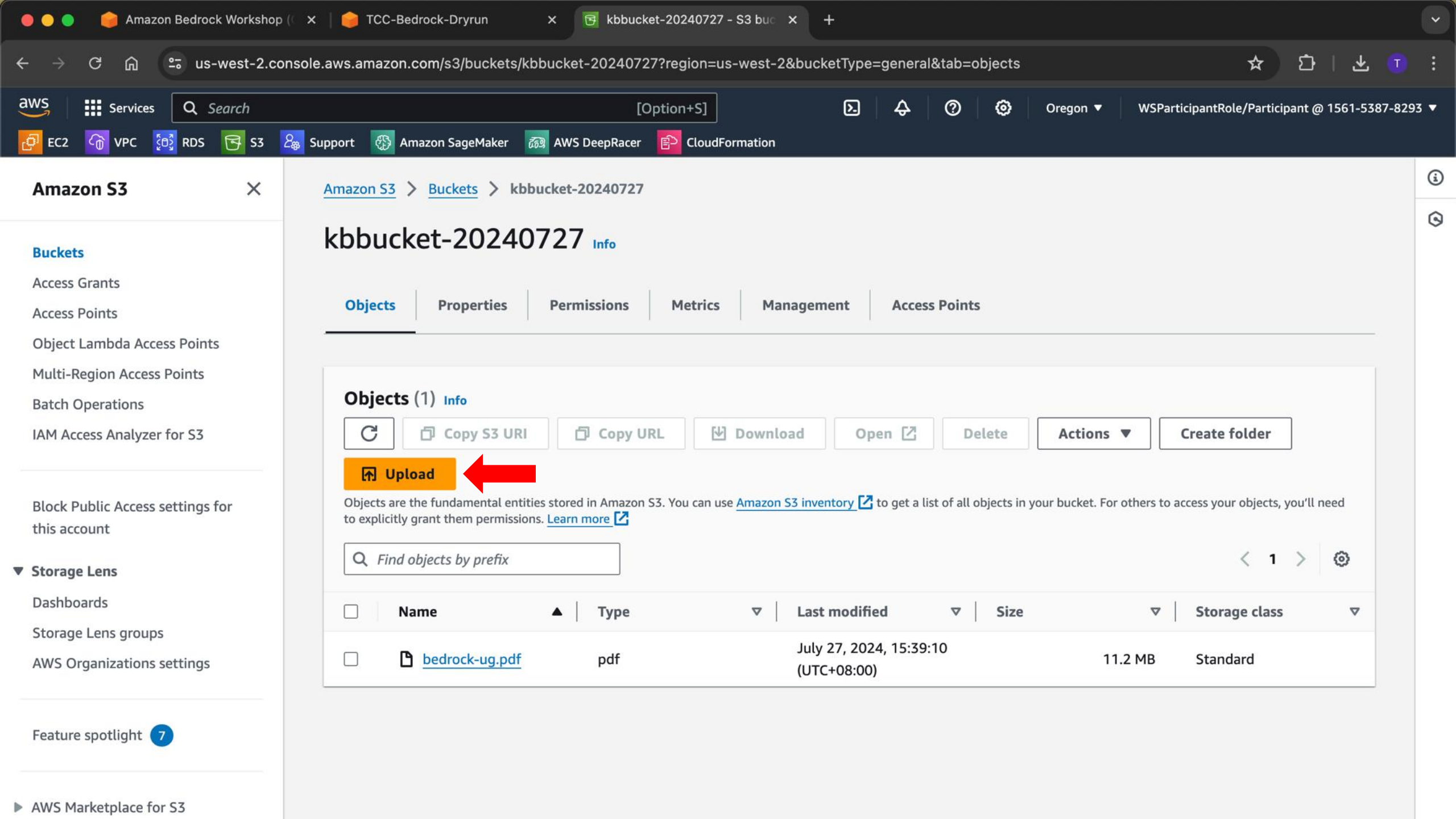
二、本須知用語定義如下：

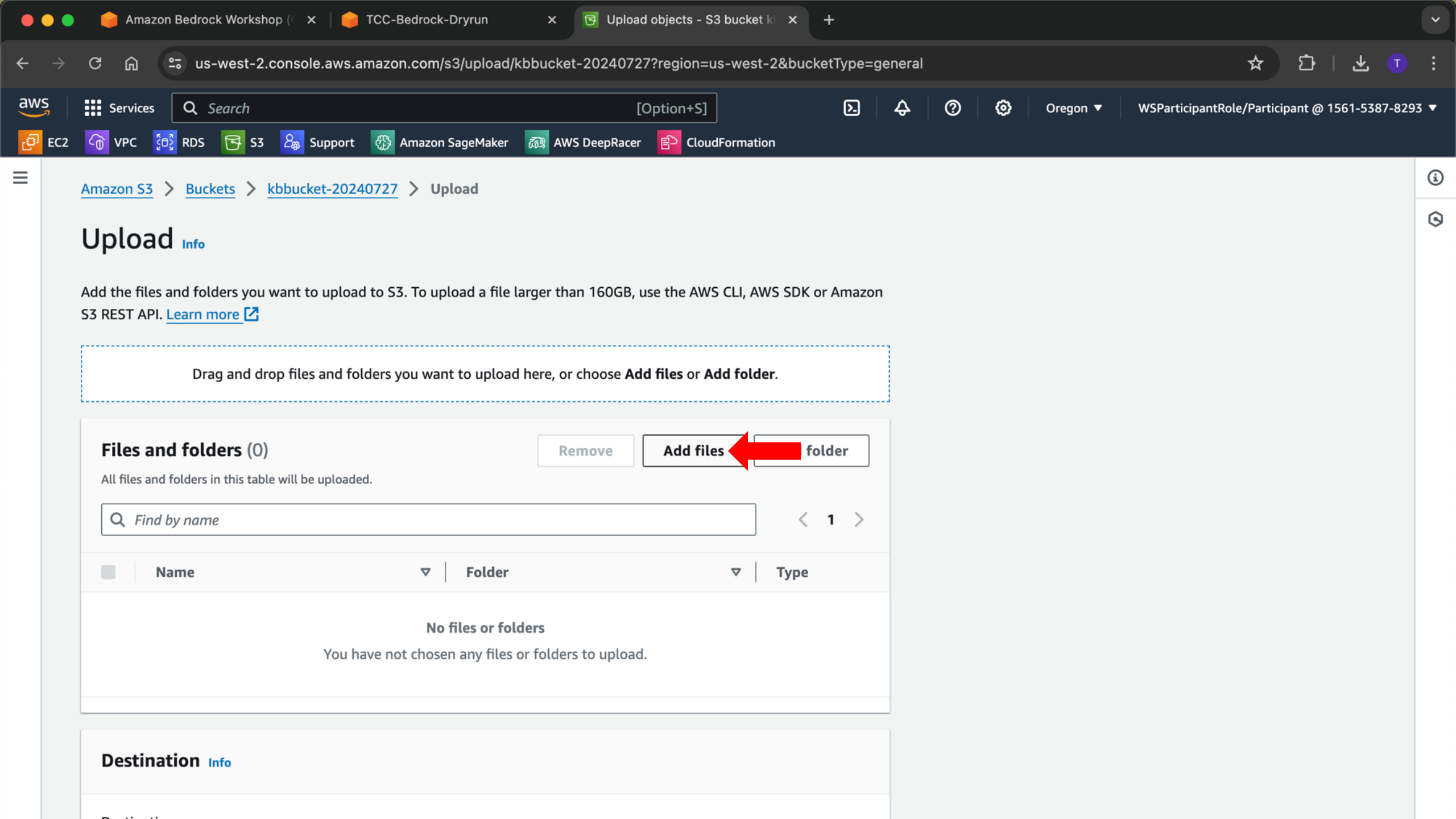
(一)捷運範圍：為本公司所經管之大眾捷運系統路網範圍內所有路線、場、站與列車等區域。

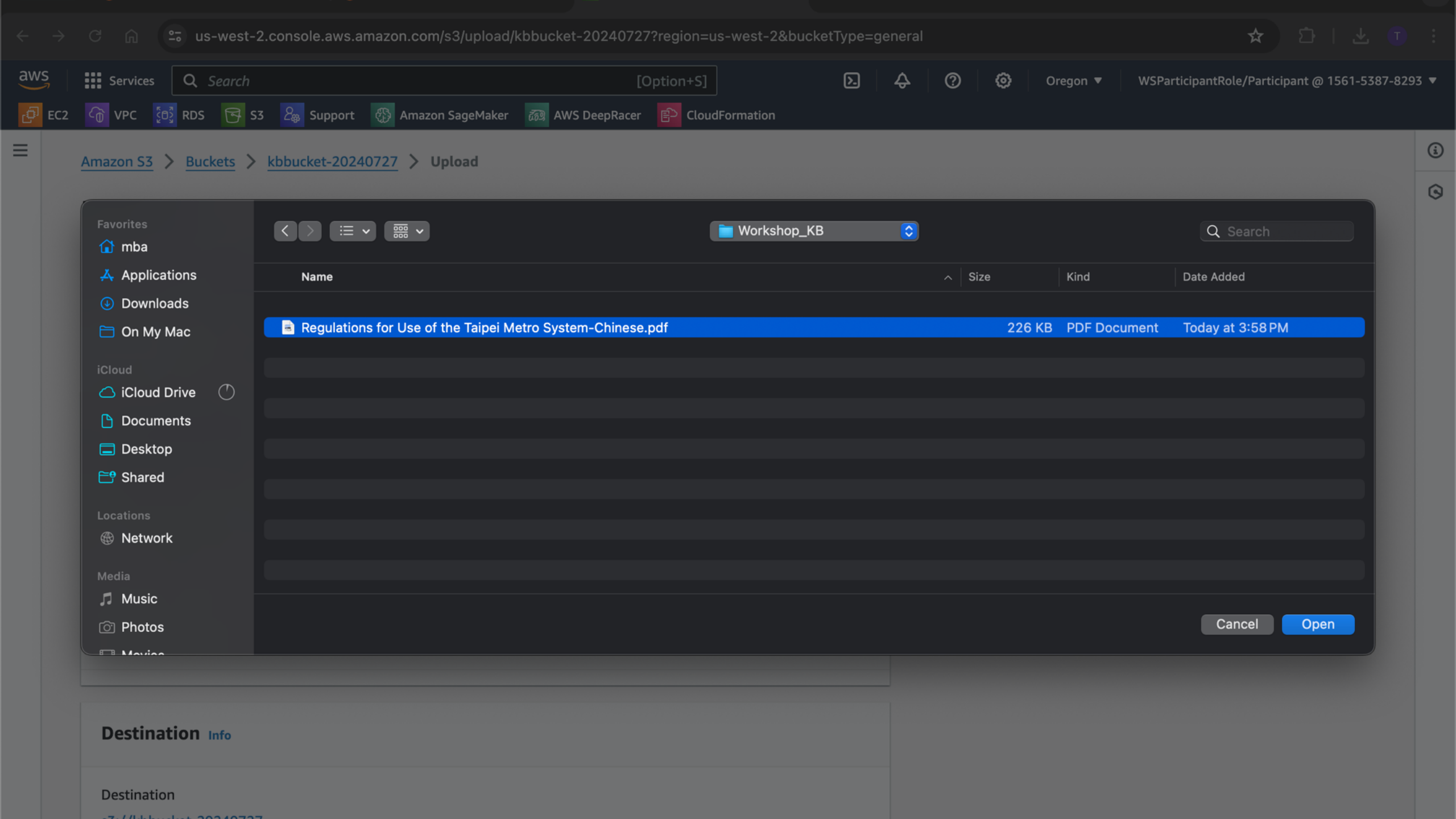
(二)旅客：指搭乘本公司列車，或持有有效車票並進出乘車處所車站大廳之人。


aws
Services
s3
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Prompt flows
Assessment & deployment
Search results for 's3'
Services (8)
Features (39)
Resources New
Documentation (26,830)
Knowledge Articles (288)
Marketplace (1,850)
Blogs (1,415)
Events (26)
Tutorials (12)
Services
See all 8 results
S3
Scalable Storage in the Cloud
S3 Glacier
Archive Storage in the Cloud
AWS Snow Family
Large Scale Data Transport
Storage Gateway
Hybrid Storage Integration
Features
See all 39 results
Imports from S3
DynamoDB feature
Feature spotlight
S3 feature
S3 Access Grants
from a data source and that relates to
r knowledge base. 2. Store your
source and configure the
nd crawl your data. 3. (Optional
documents) Create a metadata
for filtering of results during
t up a vector index in a
ta. You can use the Amazon
OpenSearch Serverless vector
re your knowledge base. 6.
bases generate embeddings
them in a supported vector
nt to query the knowledge base
/latest/userguide/kb-how-it-
/latest/APIReference/API_agent
/latest/APIReference/API_agent
/latest/userguide/data-source-
chunk
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
S3 buckets | S3 | us-west-2
us-west-2.console.aws.amazon.com/s3/home?region=us-west-2
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3
Account snapshot - updated every 24 hours
All AWS Regions
Storage lens provides visibility into storage usage and activity trends. Learn more
View Storage Lens dashboard
General purpose buckets
Directory buckets
General purpose buckets (1) Info
All AWS Regions
Copy ARN
Empty
Delete
Create bucket
Buckets are containers for data stored in S3.
Find buckets by name
1
Name
AWS Region
IAM Access Analyzer
Creation date
kbbucket-20240727
US West (Oregon) us-west-2
View analyzer for us-west-2
July 27, 2024, 15:38:19 (UTC+08:00)


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
kbbucket-20240727 - S3 buc
us-west-2.console.aws.amazon.com/s3/buckets/kbbucket-20240727?region=us-west-2&bucketType=general&tab=objects
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight 7
AWS Marketplace for S3
Amazon S3 > Buckets > kbbucket-20240727
kbbucket-20240727 Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (1) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
bedrock-ug.pdf
pdf
July 27, 2024, 15:39:10 (UTC+08:00)
11.2 MB
Standard


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Upload objects - S3 bucket
us-west-2.console.aws.amazon.com/s3/upload/kbbucket-20240727?region=us-west-2&bucketType=general
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon S3 > Buckets > kbbucket-20240727 > Upload
Upload Info
Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more
Drag and drop files and folders you want to upload here, or choose Add files or Add folder.
Files and folders (0)
Remove
Add files
folder
All files and folders in this table will be uploaded.
Find by name
1
Name
Folder
Type
No files or folders
You have not chosen any files or folders to upload.
Destination Info


us-west-2.console.aws.amazon.com/s3/upload/kbbucket-20240727?region=us-west-2&bucketType=general
aws
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon S3 > Buckets > kbbucket-20240727 > Upload
Favorites
mba
Applications
Downloads
On My Mac
iCloud
iCloud Drive
Documents
Desktop
Shared
Locations
Network
Media
Music
Photos
Workshop_KB
Search
Name
Size
Kind
Date Added
Regulations for Use of the Taipei Metro System-Chinese.pdf
226 KB
PDF Document
Today at 3:58 PM
Cancel
Open
Destination Info
Destination

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Upload objects - S3 bucket k

us-west-2.console.aws.amazon.com/s3/upload/kbbucket-20240727?region=us-west-2&bucketType=general

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (1 Total, 220.4 KB)

Remove | Add files | Add folder

All files and folders in this table will be uploaded.

Find by name

1

Name	Folder	Type
Regulations for Use of the Taipei Metro …	-	application/pdf

Destination Info

Destination

s3://kbbucket-20240727

Drag and drop files and folders you want to upload here, or choose **Add files** or **Add folder**.

Files and folders (1 Total, 220.4 KB)

Remove | Add files | Add folder

All files and folders in this table will be uploaded.

Find by name

1

Name	Folder	Type
Regulations for Use of the Taipei Metro ...	-	application/pdf

Destination Info

Destination

s3://kbbucket-20240727

▶ **Destination details**

Bucket settings that impact new objects stored in the specified destination.

▶ **Permissions**

Grant public access and access to other AWS accounts.

▶ **Properties**

Specify storage class, encryption settings, tags, and more.

Cancel **Upload**

Upload succeeded
View details below.

Upload: status

Close

The information below will no longer be available after you navigate away from this page.

Summary

Destination	Succeeded	Failed
s3://kbbucket-20240727	1 file, 220.4 KB (100.00%)	0 files, 0 B (0%)

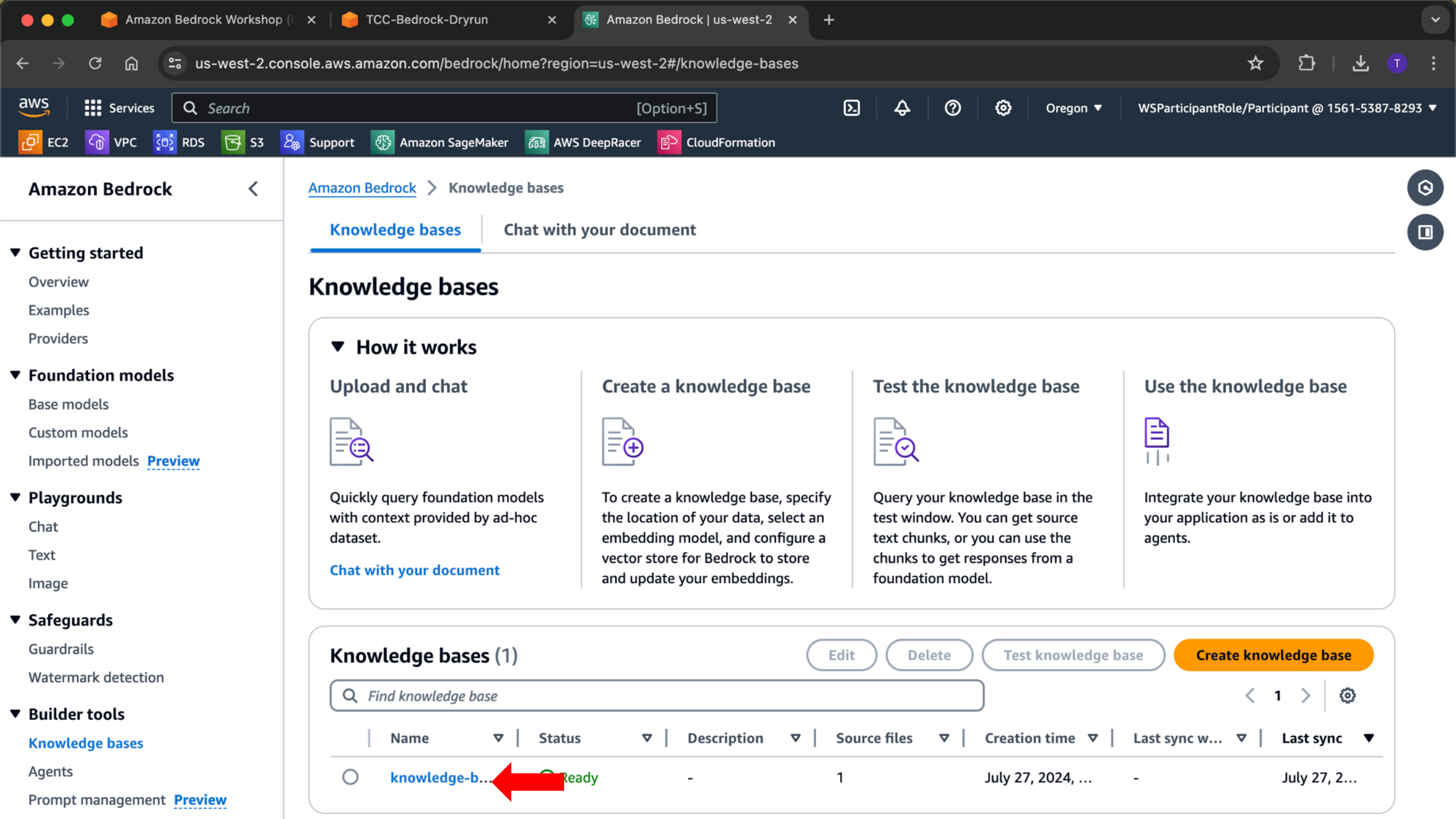
Files and folders | Configuration

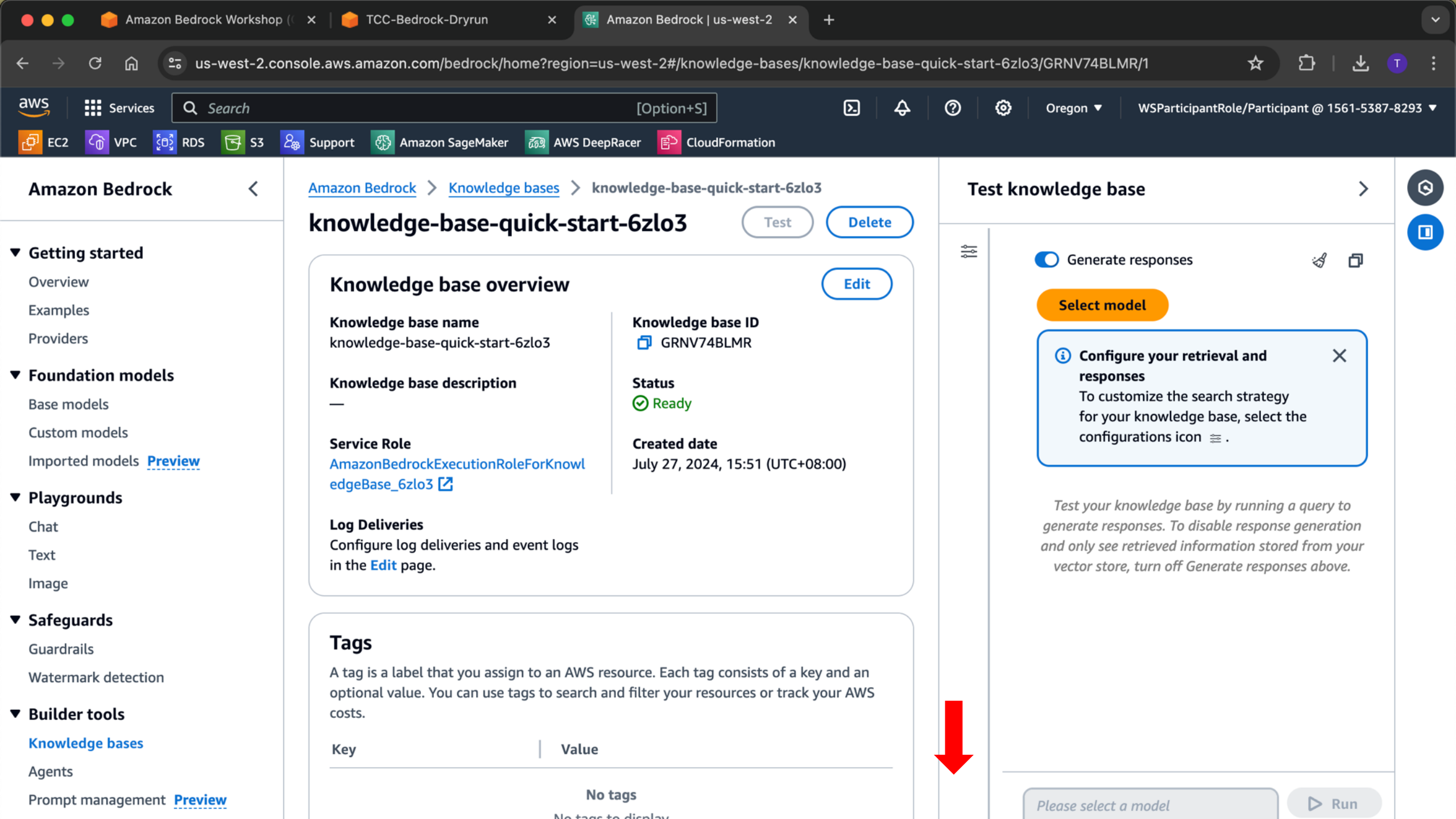
Files and folders (1 Total, 220.4 KB)

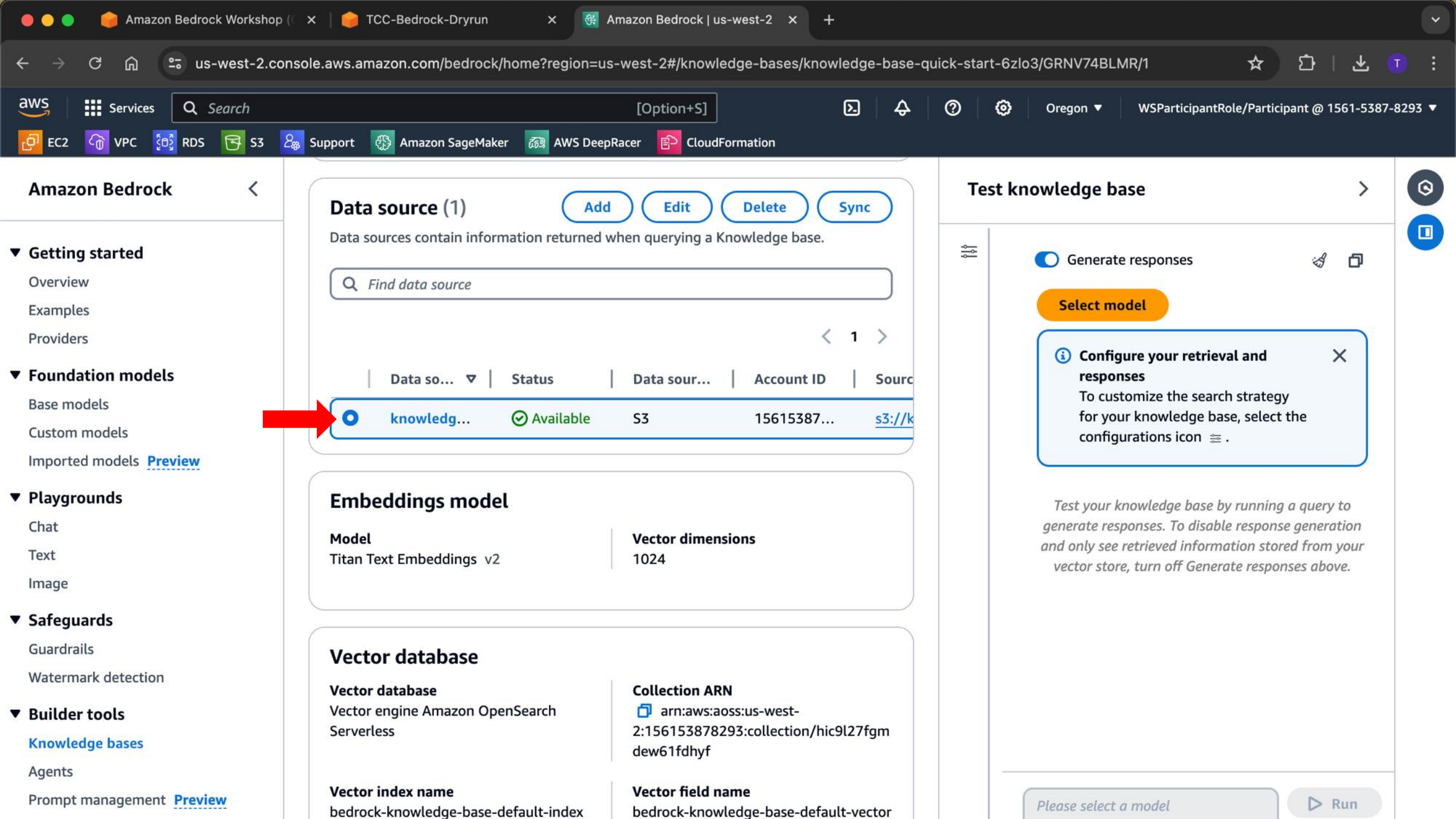
Find by name

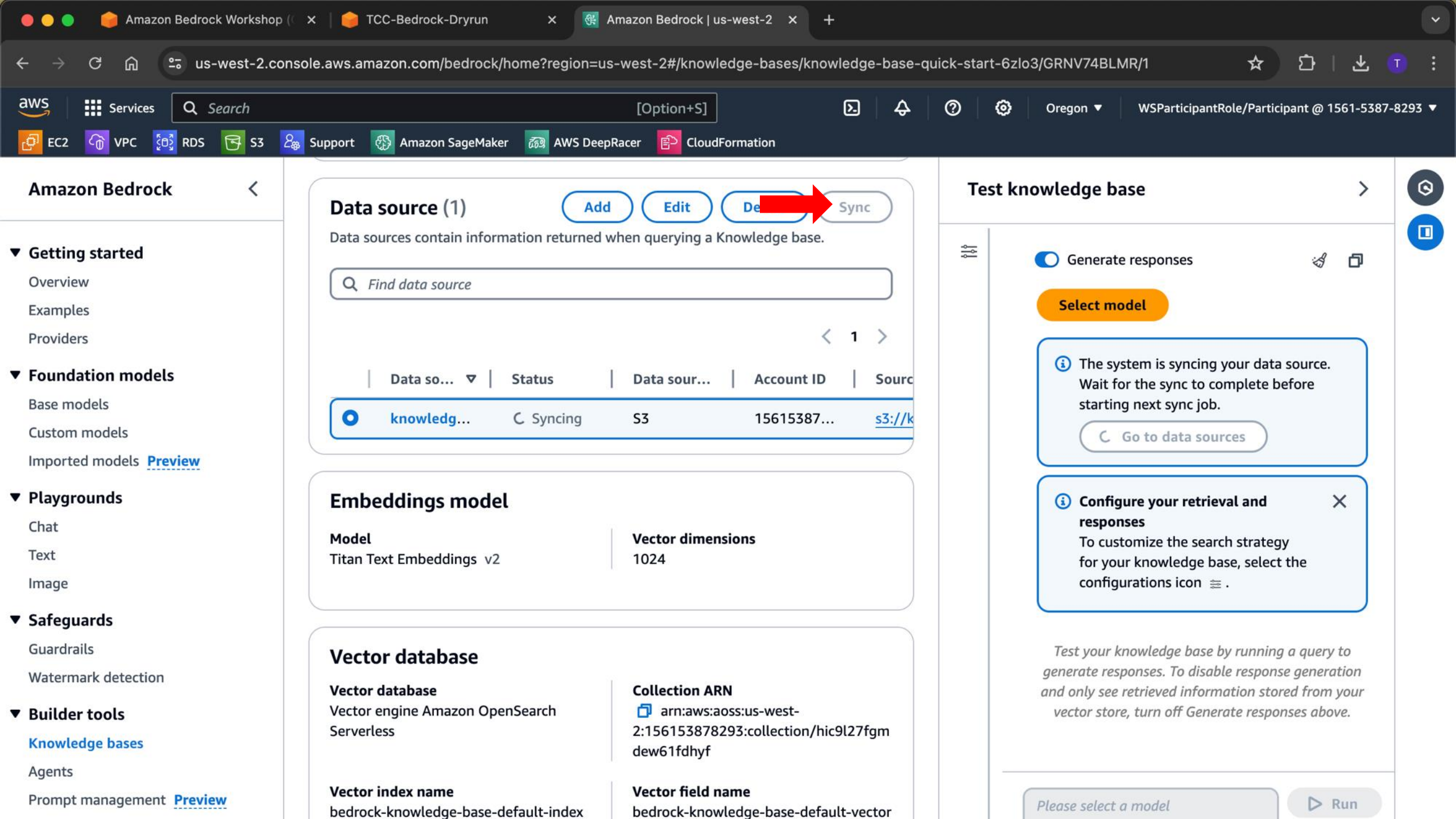
1

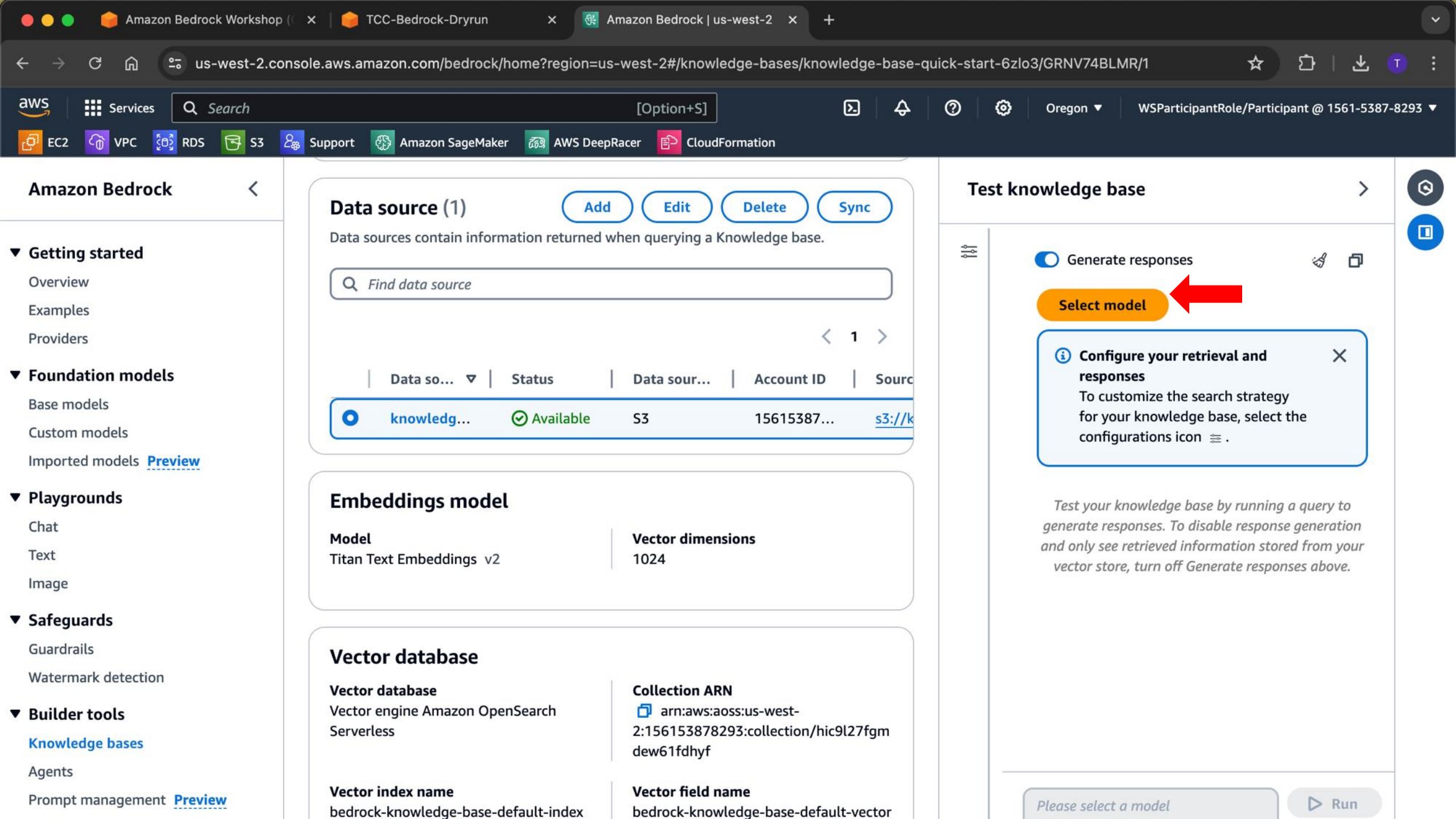
Name	Folder	Type	Size	Status	Error
Regulations	-	application/	220.4 KB	Succeeded	-

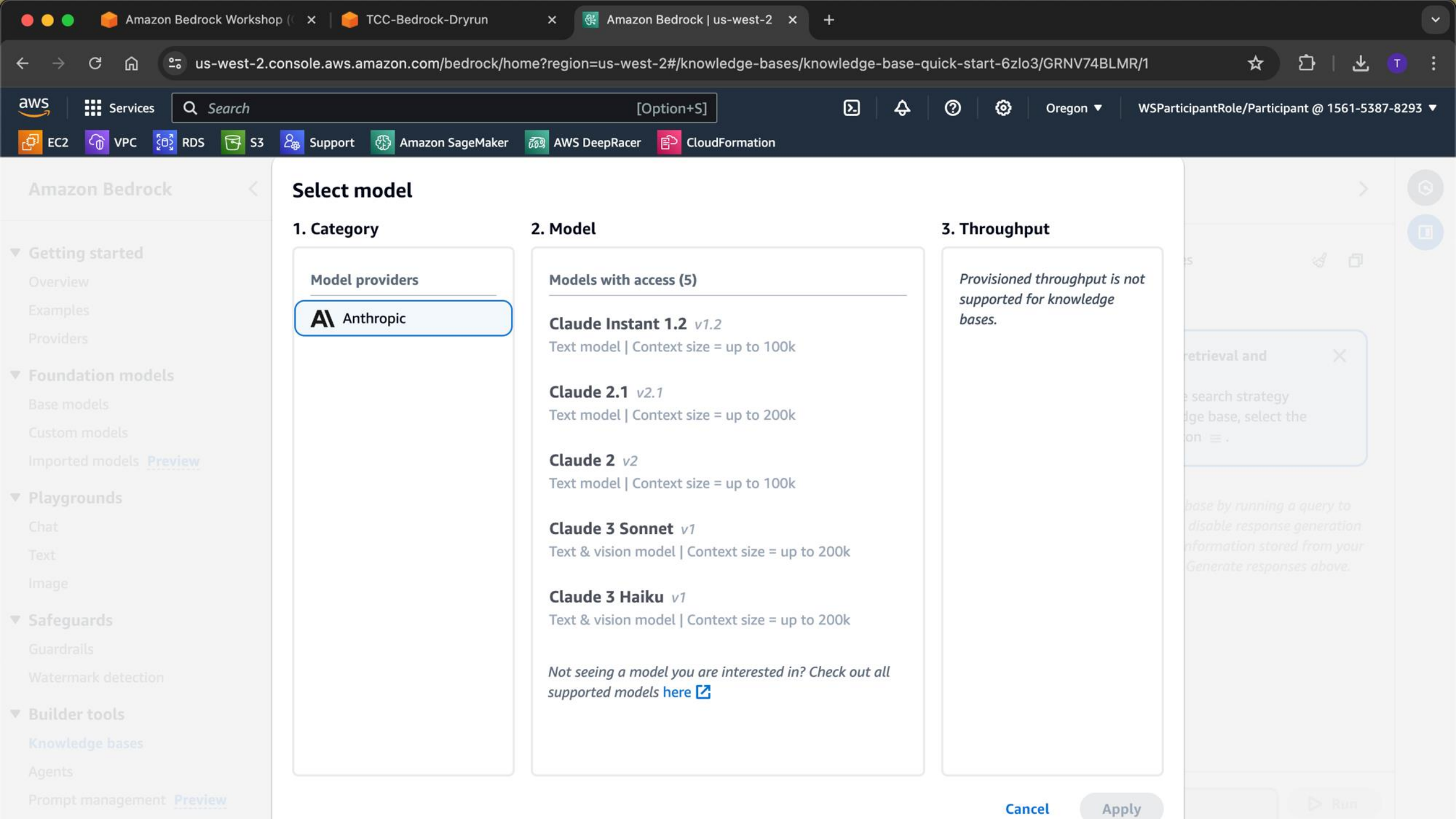

Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Amazon Bedrock > Knowledge bases
Knowledge bases
Chat with your document
Knowledge bases
How it works
Upload and chat
Quickly query foundation models with context provided by ad-hoc dataset.
Chat with your document
Create a knowledge base
To create a knowledge base, specify the location of your data, select an embedding model, and configure a vector store for Bedrock to store and update your embeddings.
Test the knowledge base
Query your knowledge base in the test window. You can get source text chunks, or you can use the chunks to get responses from a foundation model.
Use the knowledge base
Integrate your knowledge base into your application as is or add it to agents.
Knowledge bases (1)
Edit
Delete
Test knowledge base
Create knowledge base
Find knowledge base
1
Name
Status
Description
Source files
Creation time
Last sync w...
Last sync
knowledge-b...
Ready
-
1
July 27, 2024, ...
-
July 27, 2...

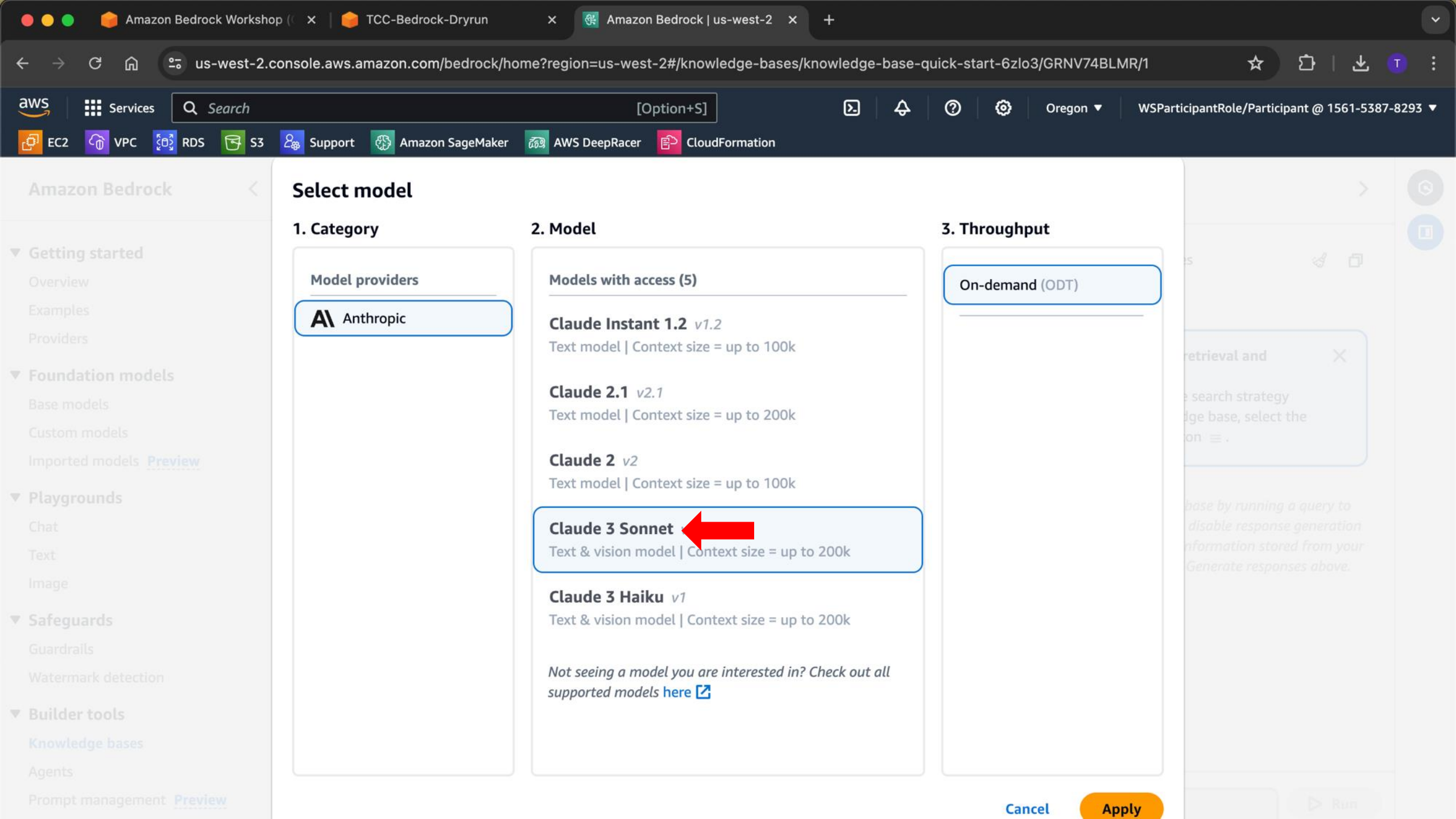

Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Amazon Bedrock > Knowledge bases > knowledge-base-quick-start-6zlo3
knowledge-base-quick-start-6zlo3
Test
Delete
Knowledge base overview
Edit
Knowledge base name
knowledge-base-quick-start-6zlo3
Knowledge base ID
GRNV74BLMR
Knowledge base description
—
Status
Ready
Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3
Created date
July 27, 2024, 15:51 (UTC+08:00)
Log Deliveries
Configure log deliveries and event logs in the Edit page.
Tags
A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.
Key
Value
No tags
No tags to display
Test knowledge base
Generate responses
Select model
Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .
Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.
Please select a model
Run


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Data source (1)
Add
Edit
Delete
Sync
Data sources contain information returned when querying a Knowledge base.
Find data source
1
Data so...
Status
Data sour...
Account ID
Sourc
knowledg...
Available
S3
15615387...
s3://k
Embeddings model
Model
Titan Text Embeddings v2
Vector dimensions
1024
Vector database
Vector database
Vector engine Amazon OpenSearch Serverless
Collection ARN
arn:aws:aoss:us-west-2:156153878293:collection/hic9l27fgmdew61fdhyf
Vector index name
bedrock-knowledge-base-default-index
Vector field name
bedrock-knowledge-base-default-vector
Test knowledge base
Generate responses
Select model
Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .
Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.
Please select a model
Run


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management Preview
Data source (1)
Add
Edit
De
Sync
Data sources contain information returned when querying a Knowledge base.
Find data source
1
Data so...
Status
Data sour...
Account ID
Sourc
knowledg...
Syncing
S3
15615387...
s3://k
Embeddings model
Model
Titan Text Embeddings v2
Vector dimensions
1024
Vector database
Vector database
Vector engine Amazon OpenSearch Serverless
Collection ARN
arn:aws:aoss:us-west-2:156153878293:collection/hic9l27fgmdew61fdhyf
Vector index name
bedrock-knowledge-base-default-index
Vector field name
bedrock-knowledge-base-default-vector
Test knowledge base
Generate responses
Select model
The system is syncing your data source. Wait for the sync to complete before starting next sync job.
Go to data sources
Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .
Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.
Please select a model
Run


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Data source (1)
Add
Edit
Delete
Sync
Data sources contain information returned when querying a Knowledge base.
Find data source
1
Data so...
Status
Data sour...
Account ID
Sourc
knowledg...
Available
S3
15615387...
s3://k
Embeddings model
Model
Titan Text Embeddings v2
Vector dimensions
1024
Vector database
Vector database
Vector engine Amazon OpenSearch Serverless
Collection ARN
arn:aws:aoss:us-west-2:156153878293:collection/hic9l27fgmdew61fdhyf
Vector index name
bedrock-knowledge-base-default-index
Vector field name
bedrock-knowledge-base-default-vector
Test knowledge base
Generate responses
Select model
Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .
Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.
Please select a model
Run


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Select model
1. Category
Model providers
Anthropic
2. Model
Models with access (5)
Claude Instant 1.2 v1.2
Text model | Context size = up to 100k
Claude 2.1 v2.1
Text model | Context size = up to 200k
Claude 2 v2
Text model | Context size = up to 100k
Claude 3 Sonnet v1
Text & vision model | Context size = up to 200k
Claude 3 Haiku v1
Text & vision model | Context size = up to 200k
Not seeing a model you are interested in? Check out all supported models here
3. Throughput
Provisioned throughput is not supported for knowledge bases.
Cancel
Apply


Amazon Bedrock Workshop
TCC-Bedrock-Dryrun
Amazon Bedrock | us-west-2
us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1
Services
Search
[Option+S]
Oregon
WSParticipantRole/Participant @ 1561-5387-8293
EC2
VPC
RDS
S3
Support
Amazon SageMaker
AWS DeepRacer
CloudFormation
Amazon Bedrock
Getting started
Overview
Examples
Providers
Foundation models
Base models
Custom models
Imported models
Preview
Playgrounds
Chat
Text
Image
Safeguards
Guardrails
Watermark detection
Builder tools
Knowledge bases
Agents
Prompt management
Preview
Select model
1. Category
Model providers
Anthropic
2. Model
Models with access (5)
Claude Instant 1.2 v1.2
Text model | Context size = up to 100k
Claude 2.1 v2.1
Text model | Context size = up to 200k
Claude 2 v2
Text model | Context size = up to 100k
Claude 3 Sonnet
Text & vision model | Context size = up to 200k
Claude 3 Haiku v1
Text & vision model | Context size = up to 200k
Not seeing a model you are interested in? Check out all supported models here
3. Throughput
On-demand (ODT)
Cancel
Apply

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
  - Prompt management Preview

knowledge-base-quick-start-6zlo3

Test Delete

Knowledge base overview

Edit

Knowledge base name
knowledge-base-quick-start-6zlo3

Knowledge base ID
GRNV74BLMR

Knowledge base description
—

Status
Ready

Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3

Created date
July 27, 2024, 15:51 (UTC+08:00)

Log Deliveries
Configure log deliveries and event logs in the Edit page.

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value

No tags

No tags to display

Test knowledge base

Generate responses

Claude 3 Sonnet *v1* | ODT
Change

Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .

Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.

Enter your message here

Run

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview

knowledge-base-quick-start-6zlo3

Test | Delete

Knowledge base overview

Edit

Knowledge base name
knowledge-base-quick-start-6zlo3

Knowledge base ID
GRNV74BLMR

Knowledge base description
—

Status
Ready

Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3

Created date
July 27, 2024, 15:51 (UTC+08:00)

Log Deliveries
Configure log deliveries and event logs in the Edit page.

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value

No tags

No tags to display

Manage tags

Test knowledge base

Generate responses

Claude 3 Sonnet *v1* | ODT

Change

Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .

Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.

在台北捷運內可不可以溜直排輪？

Run

Amazon Bedrock

Getting started
Overview
Examples
Providers

Foundation models
Base models
Custom models
Imported models Preview

Playgrounds
Chat
Text
Image

Safeguards
Guardrails
Watermark detection

Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview

knowledge-base-quick-start-6zlo3

Test Delete

Knowledge base overview

Edit

Knowledge base name
knowledge-base-quick-start-6zlo3

Knowledge base ID
GRNV74BLMR

Knowledge base description
—

Status
Ready

Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3

Created date
July 27, 2024, 15:51 (UTC+08:00)

Log Deliveries
Configure log deliveries and event logs in the Edit page.

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value

No tags
No tags to display

Manage tags

Test knowledge base

Generate responses

Claude 3 Sonnet *v1* | ODT
Change

Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .

Test your knowledge base by running a query to generate responses. To disable response generation and only see retrieved information stored from your vector store, turn off Generate responses above.

在台北捷運內可不可以溜直排輪？

Run

Amazon Bedrock Workshop | TCC-Bedrock-Dryrun | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/1?modelId=a...

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

knowledge-base-quick-start-6zlo3

Test | Delete

Knowledge base overview

Edit

Knowledge base name
knowledge-base-quick-start-6zlo3

Knowledge base ID
GRNV74BLMR

Knowledge base description
—

Status
Ready

Service Role
AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3

Created date
July 27, 2024, 15:51 (UTC+08:00)

Log Deliveries
Configure log deliveries and event logs in the Edit page.

Tags

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Key	Value

No tags

No tags to display

Test knowledge base

Generate responses

Claude 3 Sonnet v1 | ODT

Change

Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .

在台北捷運內可不可以溜直排輪？

根據台北捷運系統旅客須知的規定,在捷運範圍內是不允許使用直排輪的。具體來說,第8條明確規定"在捷運範圍內,不得有下列行為:(一)使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。"因此,在台北捷運內是禁止溜直排輪的。[1]

Show source details >

Enter your message here

Run

Amazon Bedrock

Getting started
Overview
Examples
Providers

Foundation models
Base models
Custom models
Imported models Preview

Playgrounds
Chat
Text
Image

Safeguards
Guardrails
Watermark detection

Builder tools
Knowledge bases
Agents
Prompt management Preview
Prompt flows Preview

Assessment & deployment

Test Knowledge base

Generate responses

Claude 3 Sonnet *v1* | ODT
Change

Configure your retrieval and responses
To customize the search strategy for your knowledge base, select the configurations icon .

在台北捷運內可不可以溜直排輪？

根據台北捷運系統旅客須知的規定,在捷運範圍內是不允許使用直排輪的。具體來說,第8條明確規定"在捷運範圍內,不得有下列行為:(一)使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。"因此,在台北捷運內是禁止溜直排輪的。[1]

< Hide source details

Source details (1)

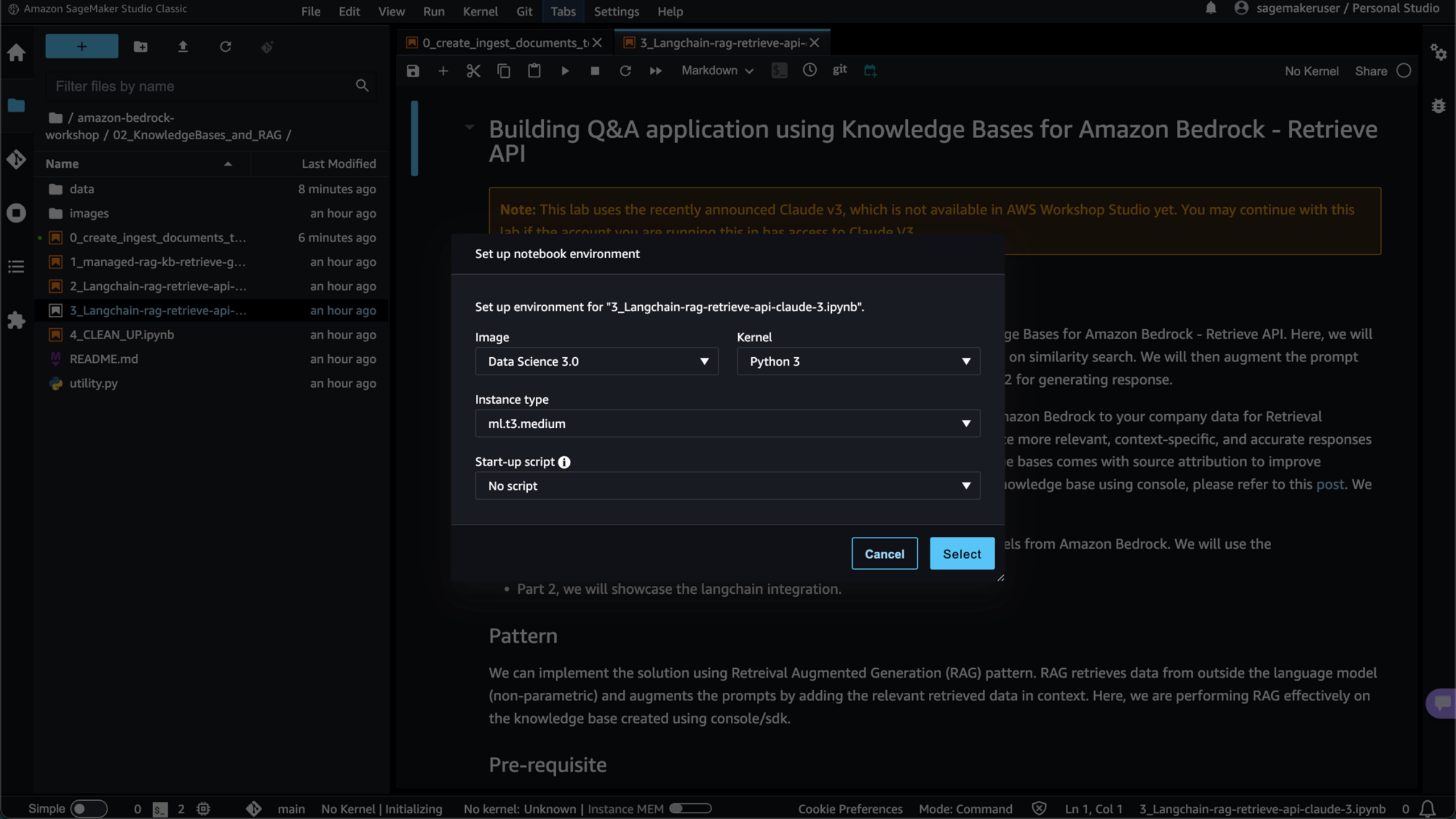
Each container below shows a chunk that comes from a data source and that relates to your query.

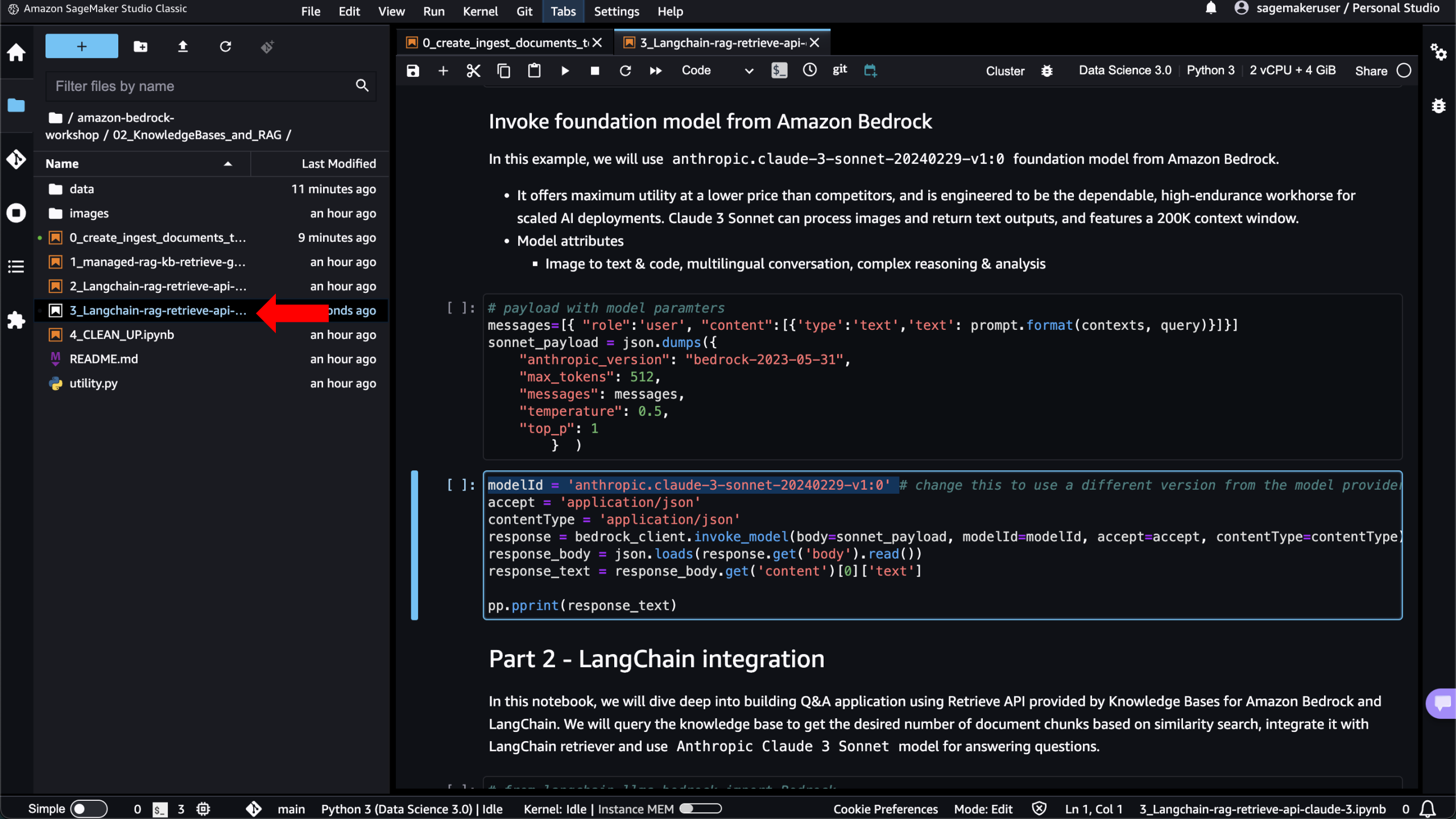
Query configurations (1)

Source chunk 1

（七）其他經本公司公告禁止之事項。八、 在捷運範圍內，不得有下列行為： （一）使用直排輪、溜冰鞋、滑板、滑板車或其他類似器材。（二）推擠或影響自己或他人安全。 （三）坐、臥於車廂、車站地板、設備或設施上。 北捷024184臺北捷運系統旅客須知 3 （四）吸菸。（五）其他行為有造成大眾捷運系統設備損壞、運轉障礙或構成危險 之虞。 九、在捷運範圍內為下列行為，應向本公司申請許可後，始得為之： （一）聚眾講演、播放音響、演奏樂器或其他干擾之行為。 （二）張貼、塗抹、刻畫任何文字、圖畫或其他類似東西於各項設施 及建築物上。 （三）於車站或車廂內，照相、拍攝或攝影，而妨礙他人或系統安全 之虞者。 （四）非營運時間內，於車站或車廂內逗留。 （五）向他人為傳教、市場調查或其他類似行為。 （六）散發報紙、傳單、廣告物或宣傳品。 （七）使用車站、車廂內未開放使用之電源插座。 十、 在捷運範圍內，電扶梯之搭乘規定及注意事項如下： （一）年長及行動不便者宜改搭電梯。 （二）須遵循電扶梯方向搭乘，握好扶手、站穩踏階，勿倚靠側板。 （三）禁止於電扶梯上奔跑、嬉戲、跳躍、跨越兩側護欄或其他危險 行為。 （四）禁止攜帶大型物品搭乘電扶梯。 貳、 車票使用規定 十一、 車票種類： （一）單程票：提供旅客單次使用之車票。 （二）團體票：提供旅客 2人以上，全程同行且起訖站相同使用 之車票。 （三）定期票：提供旅客於一定期間內使用之車票。 （四）回數票：可供旅客於一定區間或不限區間搭乘一定次 數之車票。 （五）儲值卡：各發行機

Bedrock KB + LangChain Hands-On Lab


Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
0_create_ingest_documents_t
3_Langchain-rag-retrieve-api-
Markdown
git
No Kernel
Share
Filter files by name
/ amazon-bedrock-workshop / 02_KnowledgeBases_and_RAG /
Name
Last Modified
data 8 minutes ago
images an hour ago
0_create_ingest_documents_t... 6 minutes ago
1_managed-rag-kb-retrieve-g... an hour ago
2_Langchain-rag-retrieve-api-... an hour ago
3_Langchain-rag-retrieve-api-... an hour ago
4_CLEAN_UP.ipynb an hour ago
README.md an hour ago
utility.py an hour ago
Building Q&A application using Knowledge Bases for Amazon Bedrock - Retrieve API
Note: This lab uses the recently announced Claude v3, which is not available in AWS Workshop Studio yet. You may continue with this lab if the account you are running this in has access to Claude V3
Set up notebook environment
Set up environment for "3_Langchain-rag-retrieve-api-claude-3.ipynb".
Image
Data Science 3.0
Kernel
Python 3
Instance type
ml.t3.medium
Start-up script
No script
Cancel
Select
ge Bases for Amazon Bedrock - Retrieve API. Here, we will
on similarity search. We will then augment the prompt
2 for generating response.
nazon Bedrock to your company data for Retrieval
e more relevant, context-specific, and accurate responses
e bases comes with source attribution to improve
owledge base using console, please refer to this post. We
els from Amazon Bedrock. We will use the
Part 2, we will showcase the langchain integration.
Pattern
We can implement the solution using Retreival Augmented Generation (RAG) pattern. RAG retrieves data from outside the language model (non-parametric) and augments the prompts by adding the relevant retrieved data in context. Here, we are performing RAG effectively on the knowledge base created using console/sdk.
Pre-requisite
Simple
main
No Kernel | Initializing
No kernel: Unknown | Instance MEM
Cookie Preferences
Mode: Command
Ln 1, Col 1
3_Langchain-rag-retrieve-api-claude-3.ipynb



Invoke foundation model from Amazon Bedrock

In this example, we will use `anthropic.claude-3-sonnet-20240229-v1:0` foundation model from Amazon Bedrock.

- It offers maximum utility at a lower price than competitors, and is engineered to be the dependable, high-endurance workhorse for scaled AI deployments. Claude 3 Sonnet can process images and return text outputs, and features a 200K context window.
- Model attributes
  - Image to text & code, multilingual conversation, complex reasoning & analysis

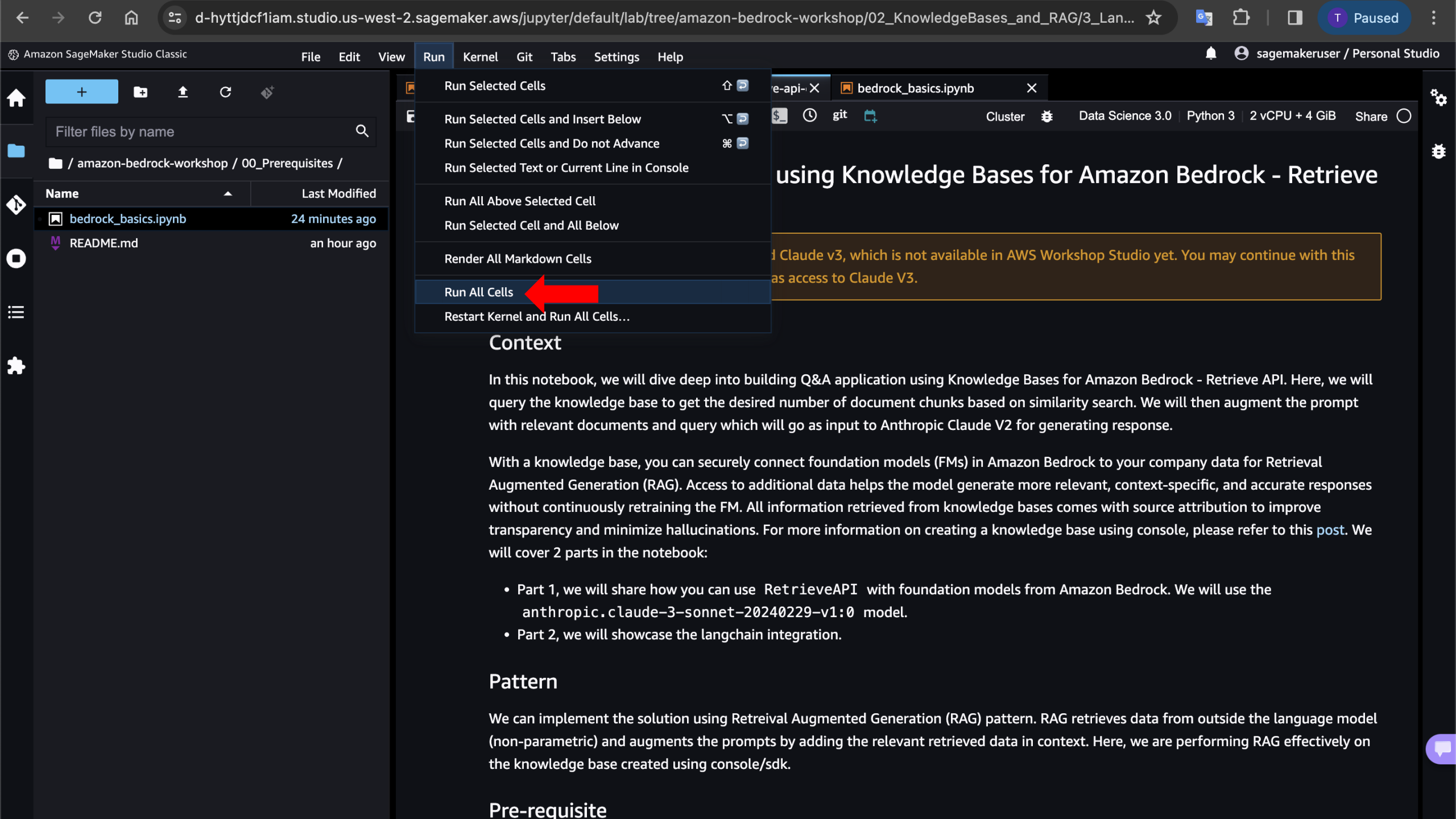
[ ]:
```
# payload with model paramters
messages=[{ "role":'user', "content":[{'type':'text','text': prompt.format(contexts, query)}]}]
sonnet_payload = json.dumps({
    "anthropic_version": "bedrock-2023-05-31",
    "max_tokens": 512,
    "messages": messages,
    "temperature": 0.5,
    "top_p": 1
        }  )
```

[ ]:
```
modelId = 'anthropic.claude-3-sonnet-20240229-v1:0' # change this to use a different version from the model provide
accept = 'application/json'
contentType = 'application/json'
response = bedrock_client.invoke_model(body=sonnet_payload, modelId=modelId, accept=accept, contentType=contentType)
response_body = json.loads(response.get('body').read())
response_text = response_body.get('content')[0]['text']

pp.pprint(response_text)
```

Part 2 - LangChain integration

In this notebook, we will dive deep into building Q&A application using Retrieve API provided by Knowledge Bases for Amazon Bedrock and LangChain. We will query the knowledge base to get the desired number of document chunks based on similarity search, integrate it with LangChain retriever and use `Anthropic Claude 3 Sonnet` model for answering questions.

[ ]:
```
# from langchain.llms.bedrock import Bedrock
```


d-hyttjdcf1iam.studio.us-west-2.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop/02_KnowledgeBases_and_RAG/3_Lan...
Paused
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sagemakeruser / Personal Studio
Run Selected Cells
Run Selected Cells and Insert Below
Run Selected Cells and Do not Advance
Run Selected Text or Current Line in Console
Run All Above Selected Cell
Run Selected Cell and All Below
Render All Markdown Cells
Run All Cells
Restart Kernel and Run All Cells...
Filter files by name
/ amazon-bedrock-workshop / 00_Prerequisites /
Name
Last Modified
bedrock_basics.ipynb
24 minutes ago
README.md
an hour ago
bedrock_basics.ipynb
git
Cluster
Data Science 3.0 | Python 3 | 2 vCPU + 4 GiB
Share
using Knowledge Bases for Amazon Bedrock - Retrieve
Claude v3, which is not available in AWS Workshop Studio yet. You may continue with this
as access to Claude V3.
Context
In this notebook, we will dive deep into building Q&A application using Knowledge Bases for Amazon Bedrock - Retrieve API. Here, we will query the knowledge base to get the desired number of document chunks based on similarity search. We will then augment the prompt with relevant documents and query which will go as input to Anthropic Claude V2 for generating response.
With a knowledge base, you can securely connect foundation models (FMs) in Amazon Bedrock to your company data for Retrieval Augmented Generation (RAG). Access to additional data helps the model generate more relevant, context-specific, and accurate responses without continuously retraining the FM. All information retrieved from knowledge bases comes with source attribution to improve transparency and minimize hallucinations. For more information on creating a knowledge base using console, please refer to this post. We will cover 2 parts in the notebook:
Part 1, we will share how you can use RetrieveAPI with foundation models from Amazon Bedrock. We will use the anthropic.claude-3-sonnet-20240229-v1:0 model.
Part 2, we will showcase the langchain integration.
Pattern
We can implement the solution using Retreival Augmented Generation (RAG) pattern. RAG retrieves data from outside the language model (non-parametric) and augments the prompts by adding the relevant retrieved data in context. Here, we are performing RAG effectively on the knowledge base created using console/sdk.
Pre-requisite

Part 2 - LangChain integration

In this notebook, we will dive deep into building Q&A application using Retrieve API provided by Knowledge Bases for Amazon Bedrock and LangChain. We will query the knowledge base to get the desired number of document chunks based on similarity search, integrate it with LangChain retriever and use `Anthropic Claude 3 Sonnet` model for answering questions.

[20]:
```
# from langchain.llms.bedrock import Bedrock
from langchain_community.chat_models.bedrock import BedrockChat
from langchain.retrievers.bedrock import AmazonKnowledgeBasesRetriever

llm = BedrockChat(model_id=modelId,
                  client=bedrock_client)
```

Create a `AmazonKnowledgeBasesRetriever` object from LangChain which will call the `Retreive API` provided by Knowledge Bases for Amazon Bedrock which converts user queries into embeddings, searches the knowledge base, and returns the relevant results, giving you more control to build custom workflows on top of the semantic search results. The output of the `Retrieve API` includes the the `retrieved text chunks` , the `location type` and `URI` of the source data, as well as the relevance `scores` of the retrievals.

[21]:
```
query = "What is Amazon doing in the field of Generative AI?"
retriever = AmazonKnowledgeBasesRetriever(
        knowledge_base_id=kb_id,
        retrieval_config={"vectorSearchConfiguration":
                          {"numberOfResults": 4,
                           'overrideSearchType': "SEMANTIC", # optional
                           }
                          },
        # endpoint_url=endpoint_url,
        # region_name=region,
        # credentials_profile_name="<profile_name>",
    )
docs = retriever.get_relevant_documents(
        query=query
    )
pp.pprint(docs)
```

```
[ Document(page_content='This shift was driven by several factors, including access to higher volumes of compute capacity at lower prices than was ever availabl
e. Amazon has been using machine learning extensively for 25 years, employing it in everything from personalized ecommerce recommendations, to fulfillment cente
r pick paths, to drones for Prime Air, to Alexa, to the many machine learning services AWS offers (where AWS has the broadest machine learning functionality and
customer base of any cloud provider). More recently, a newer form of machine learning, called Generative AI, has burst onto the scene and promises to significan
tly accelerate machine learning adoption. Generative AI is based on very Large Language Models (trained on up to hundreds of billions of parameters, and growin
```

Prompt specific to the model to personalize responses

Here, we will use the specific prompt below for the model to act as a financial advisor AI system that will provide answers to questions by using fact based and statistical information when possible. We will provide the Retrieve API responses from above as a part of the `{context}` in the prompt for the model to refer to, along with the user `query`.

[22]:
```
from langchain.prompts import PromptTemplate

PROMPT_TEMPLATE = """
Human: You are a financial advisor AI system, and provides answers to questions by using fact based and statistical information when possible.
Use the following pieces of information to provide a concise answer to the question enclosed in <question> tags.
If you don't know the answer, just say that you don't know, don't try to make up an answer.
<context>
{context}
</context>

<question>
{question}
</question>

The response should be specific and use statistics or numbers when possible.

Assistant:"""
claude_prompt = PromptTemplate(template=PROMPT_TEMPLATE,
                               input_variables=["context","question"])
```

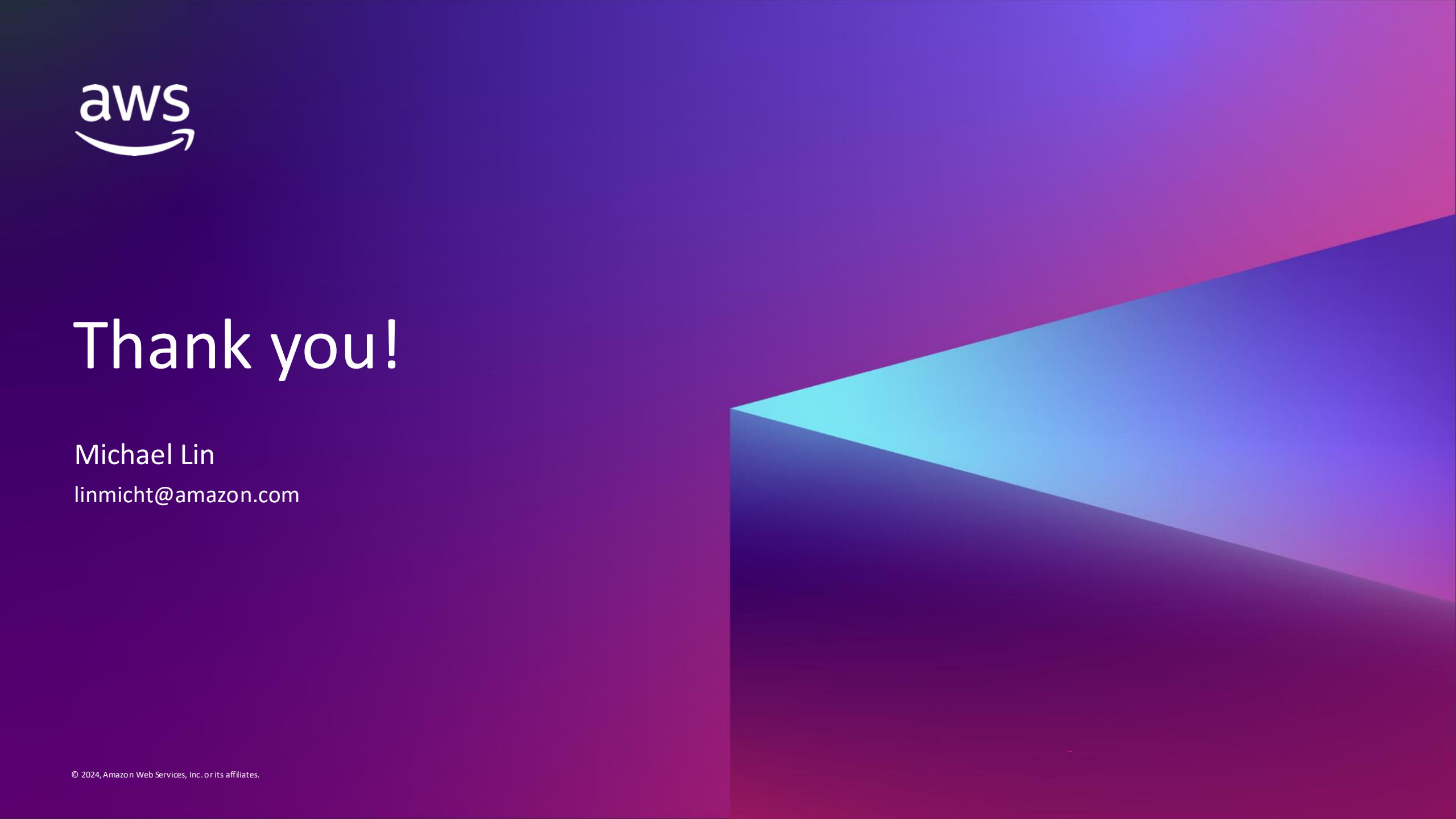
Integrating the retriever and the LLM defined above with RetrievalQA Chain to build the Q&A application.

[23]:
```
from langchain.chains import RetrievalQA

qa = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=retriever,
    return_source_documents=True,
    chain_type_kwargs={"prompt": claude_prompt}
)
```

[23]:
```
from langchain.chains import RetrievalQA

qa = RetrievalQA.from_chain_type(
    llm=llm,
    chain_type="stuff",
    retriever=retriever,
    return_source_documents=True,
    chain_type_kwargs={"prompt": claude_prompt}
)
```

[24]:
```
answer = qa.invoke(query)
pp.pprint(answer)
```

```
{ 'query': 'What is Amazon doing in the field of Generative AI?',
  'result': 'From the context provided:\n'
            '\n'
            'Amazon has been working on their own large language models (LLMs) '
            'for Generative AI for a while now. They believe Generative AI, '
            'based on LLMs trained on up to hundreds of billions of parameters '
            'across expansive datasets, will transform and improve virtually '
            'every customer experience. Amazon is continuing to invest '
            'substantially in these models across all of their consumer, '
            'seller, brand, and creator experiences. Additionally, Amazon is '
            'democratizing this technology through AWS so companies of all '
            'sizes can leverage Generative AI. AWS offers various LLMs and '
            'enables companies to build applications using Generative AI with '
            "AWS's security, privacy and other features. One example "
            "application mentioned is AWS's CodeWhisperer, which uses "
            'Generative AI to generate code suggestions in real time to '
            'revolutionize developer productivity.',
  'source_documents': [ Document(page_content='This shift was driven by several factors, including access to higher volumes of compute capacity at lower prices
than was ever available. Amazon has been using machine learning extensively for 25 years, employing it in everything from personalized ecommerce recommendation
s, to fulfillment center pick paths, to drones for Prime Air, to Alexa, to the many machine learning services AWS offers (where AWS has the broadest machine lea
rning functionality and customer base of any cloud provider). More recently, a newer form of machine learning, called Generative AI, has burst onto the scene an
d promises to significantly accelerate machine learning adoption. Generative AI is based on very Large Language Models (trained on up to hundreds of billions of
parameters, and growing), across expansive datasets, and has radically general and broad recall and learning capabilities. We have been working on our own LLMs
for a while now, believe it will transform and improve virtually every customer experience, and will continue to invest substantially in these models across all
of our consumer, seller, brand, and creator experiences. Additionally, as we've done for years in AWS, we're democratizing this technology so companies of all s
izes can leverage Generative AI. AWS is offering the most price-performant machine learning chips in Trainium and Inferentia so small and large companies can af
ford to train and run their LLMs in production. We enable companies to choose from various LLMs and build applications with all of the AWS security, privacy and
```



Thank you!

Michael Lin

linmicht@amazon.com