



AWS GenAI Workshop

Michael Lin

Sr. Solutions Architect
Amazon Web Services

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- Validating Bedrock agent
- Exploring Bedrock fine-tune
- Q & A

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- Validating Bedrock agent
- Exploring Bedrock fine-tune
- Q & A

Com

Services

Features

Resources New

Documentation

Knowledge articles

Marketplace

Blog posts

Events

Tutorials

Services

Show more



Cloud9

A Cloud IDE for Writing, Running, and Debugging Code



Amazon CodeCatalyst

Integrated DevOps Service



AWS Cloud Map

Build a dynamic map of your cloud



Features

Show more

Cloud WAN

- VPC feature

Namespaces

- AWS Cloud Map feature

Workloads

- AWS Well-Architected Tool feature

Were these results helpful?

Yes

No

Resources / for a focused search

get the most out of AWS.

Introducing resource search

\$0.03

Cost (\$)

2

1

X



Set to default layout

+ Add widgets

Create application



applications

< 1 >

Region | Originati.

ations
ng an application.

lication

lications



Cost (\$)

2

1

X



Developer Tools

AWS Cloud9

A cloud IDE for writing, running, and debugging code

AWS Cloud9 allows you to write, run, and debug your code with just a browser. With AWS Cloud9, you have immediate access to a rich code editor, integrated debugger, and built-in terminal with preconfigured AWS CLI. You can get started in minutes and no longer have to spend the time to install local applications or configure your development machine.

How it works

Create an AWS Cloud9 development environment on a new Amazon EC2 instance or connect it to your own Linux server through SSH. Once you've created an AWS Cloud9 environment, you will have immediate access to a rich code editor, integrated debugger, and built-in terminal with pre-configured AWS CLI – all within your browser.

Using the AWS Cloud9 dashboard, you can create and switch between many different AWS Cloud9 environments, each one containing the custom tools, runtimes, and files for a specific project.

[Learn more](#)

New AWS Cloud9 environment

[Create environment](#)

Getting started

[Before you start](#) (2 min read)

[Create an environment](#) (2 min read)

[Working with environments](#) (15 min read)

[Working with the IDE](#) (10 min read)

[Working with AWS Lambda](#) (5 min read)

[More resources](#)





Search

[Option+S]



United States (Oregon) ▾

WSParticipantRole/Participant @ 1225-1613-7395 ▾

☰ AWS Cloud9 > Environments > Create environment



ⓘ For capabilities similar to AWS Cloud9, explore AWS Toolkits in your own IDE and AWS CloudShell in the AWS Management Console. [Learn more](#)



ⓘ Introducing a new code editing experience.



Enjoy an updated editing experience with access to an extension library. You also have automatic access to extensions such as, Amazon Q and Amazon Toolkit. To try or use these extensions, open your environment using the 'Open in Code Editor' option.

Create environment

Details

Name

Limit of 60 characters, alphanumeric, and unique per user.

Description - optional

Limit 200 characters.

Environment type

Determines what the Cloud9 IDE will run on.

New EC2 instance

Cloud9 creates an EC2 instance in your account. The configuration of your EC2 instance cannot be changed by Cloud9 after creation.

Existing compute

You have an existing instance or server that you'd like to use.

New EC2 instance

Instance type



us-west-2.console.aws.amazon.com/cloud9control/home?region=us-west-2#/create/ Verify it's you

aws Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 1225-1613-7395

AWS Cloud9 Environments Create environment

For capabilities similar to AWS Cloud9, explore AWS Toolkits in your own IDE and AWS CloudShell in the AWS Management Console. [Learn more](#)

Introducing a new code editing experience. Enjoy an updated editing experience with access to an extension library. You also have automatic access to extensions such as, Amazon Q and Amazon Toolkit. To try or use these extensions, open your environment using the 'Open in Code Editor' option.

Create environment [Info](#)

Details

Name

claude-chat-20250321



Limit of 60 characters, alphanumeric, and unique per user.

Description - optional

Limit 200 characters.

Environment type [Info](#)

Determines what the Cloud9 IDE will run on.

New EC2 instance

Cloud9 creates an EC2 instance in your account. The configuration of your EC2 instance cannot be changed by Cloud9 after creation.

Existing compute

You have an existing instance or server that you'd like to use.

New EC2 instance

Instance type [Info](#)



us-west-2.console.aws.amazon.com/cloud9control/home?region=us-west-2#/create/ Verify it's you

AWS Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 1225-1613-7395

AWS Cloud9 Environments Create environment

New EC2 instance

Instance type Info
The memory and CPU of the EC2 instance that will be created for Cloud9 to run on.

- t2.micro (1 GiB RAM + 1 vCPU)**
Free-tier eligible. Ideal for educational users and exploration.
- t3.small (2 GiB RAM + 2 vCPU)**
Recommended for small web projects.
- m5.large (8 GiB RAM + 2 vCPU)**
Recommended for production and most general-purpose development.

Additional instance types
Explore additional instances to fit your need.

Platform Info
This will be installed on your EC2 instance. We recommend Amazon Linux 2023.

Amazon Linux 2023

Timeout
How long Cloud9 can be inactive (no user input) before auto-hibernating. This helps prevent unnecessary charges.

30 minutes

Credentials
Controls how your environment credentials are managed.

- Use AWS managed temporary credentials.**
Cloud9 manages AWS access credentials in an EC2 environment on your behalf, while also following AWS security best practices. This keeps your environment in sync with your login credentials with no additional effort from you.
- Manage your own credentials**
Store your permanent AWS access credentials within the environment or attach them to the instance using an IAM instance profile or inset them directly into your code.

Network settings Info



Search

[Option+S]



United States (Oregon) ▾

WSParticipantRole/Participant @ 1225-1613-7395 ▾



AWS Cloud9 > Environments > Create environment

New EC2 instance

Instance type Info

The memory and CPU of the EC2 instance that will be created for Cloud9 to run on.

t2.micro (1 GiB RAM + 1 vCPU)

Free-tier eligible. Ideal for educational users and exploration.

t3.small (2 GiB RAM + 2 vCPU)

Recommended for small web projects.

m5.large (8 GiB RAM + 2 vCPU)

Recommended for production and most general-purpose development.

Additional instance types

Explore additional instances to fit your need.

Platform Info

This will be installed on your EC2 instance. We recommend Amazon Linux 2023.

Amazon Linux 2023



Timeout

How long Cloud9 can be inactive (no user input) before auto-hibernating. This helps prevent unnecessary charges.

30 minutes



Credentials

Controls how your environment credentials are managed.

Use AWS managed temporary credentials.

Cloud9 manages AWS access credentials in an EC2 environment on your behalf, while also following AWS security best practices. This keeps your environment in sync with your login credentials with no additional effort from you.

Manage your own credentials

Store your permanent AWS access credentials within the environment or attach them to the instance using an IAM instance profile or inset them directly into your code.

Network settings Info



Store your permanent AWS access credentials within the environment or attach them to the instance using an IAM instance profile or inset them directly into your code.

Network settings Info

Connection

How your environment is accessed.

AWS Systems Manager (SSM)

Accesses environment via SSM without opening inbound ports (no ingress).

Secure Shell (SSH)

Accesses environment directly via SSH, opens inbound ports.

► VPC settings Info

► Tags - optional Info

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.

i The following IAM resources will be created in your account

- **AWSServiceRoleForAWSCloud9** - AWS Cloud9 creates a service-linked role for you. This allows AWS Cloud9 to call other AWS services on your behalf. You can delete the role from the AWS IAM console once you no longer have any AWS Cloud9 environments. [Learn more](#) ↗
- **AWSCloud9SSMAccessRole** and **AWSCloud9SSMInstanceProfile** - A service role and an instance profile are automatically created if Cloud9 accesses its EC2 instance through AWS Systems Manager. If your environments no longer require EC2 instances that block incoming traffic, you can delete these roles using the AWS IAM console. [Learn more](#) ↗

Cancel

Create





Search

[Option+S]



United States (Oregon) ▾

WSParticipantRole/Participant @ 1225-1613-7395 ▾



AWS Cloud9 > Environments



AWS Cloud9

My environments

Shared with me

All account environments

Documentation

Creating claudie-chat-20250321. This can take several minutes. While you wait, see [Best practices for using AWS Cloud9](#)

Environments (1)

[Delete](#)[Manage user access](#)[View details](#)[Open in Cloud9](#)[Create environment](#)

My environments

< 1 >



Name	Cloud9 IDE	Environment type	Connection	Permission	Owner ARN
claudie-chat-20250321	Open	EC2 instance	AWS Systems Manager (SSM)	Owner	arn:aws:sts::122516137395:assumed-role/WSParticipantRole/Participant



AWS Cloud9

My environments

Shared with me

All account environments

Documentation

Successfully created claude-chat-20250321. To get the most out of your environment, see [Best practices for using AWS Cloud9](#)

Environments (1)

[Delete](#)[Manage user access](#)[View details](#)[Open in Cloud9](#)[Create environment](#)

My environments

Name	Cloud9 IDE	Environment type	Connection	Permission	Owner ARN
claude-chat-20250321	Open	EC2 instance	AWS Systems Manager (SSM)	Owner	 arn:aws:sts::122516137395:assumed-role/WSParticipantRole/Participant

us-west-2.console.aws.amazon.com/cloud9control/home?region=us-west-2#/ Verify it's you

AWS Cloud9 | Search [Option+S] | United States (Oregon) | WSParticipantRole/Participant @ 1225-1613-7395

AWS Cloud9 > Environments

AWS Cloud9

- My environments
- Shared with me
- All account environments

Documentation

Successfully created claudie-chat-20250321. To get the most out of your environment, see [Best practices for using AWS Cloud9](#)

Environments (1)

Delete Manage user access View details Open in Cloud9 Create environment

My environments

Name	Cloud9 IDE	Environment type	Connection	Permission	Owner ARN
claudie-chat-20250321	Open	C2 instance	AWS Systems Manager (SSM)	Owner	arn:aws:sts::122516137395:assumed-role/WSParticipantRole/Participant

1 / 1



Verify it's you



!

Try using VIM mode!



A screenshot of the AWS Cloud9 IDE interface. At the top, there's a navigation bar with icons for File, Edit, Find, View, Go, Run, Tools, Window, Support, Preview, and Run. On the right side of the top bar are icons for Share and Settings. Below the navigation bar is a search bar labeled "Go to Anything (⌘ P)". To the left of the main content area is a sidebar with a file tree showing a folder named "claudie-chat-20250" containing a file named "README.md".

The main content area has a title "Welcome" and a "Developer Tools" section. Below this, a large heading says "AWS Cloud9" and "Welcome to your development environment". A descriptive text block explains that AWS Cloud9 allows you to write, run, and debug code with just a browser, mentioning the IDE, AWS Lambda, and Amazon API Gateway.

A central panel titled "Toolkit for AWS Cloud9" describes the AWS Toolkit for Cloud9, which simplifies interacting with AWS services like Lambda, CloudFormation, and API Gateway. It also mentions the AWS Serverless Application Model (SAM) and provides a "Learn more" link.

To the right of the toolkit panel is a "Getting started" sidebar with options: "Create File", "Upload Files...", and "Clone from GitHub". Below this is another sidebar titled "Configure AWS Cloud9". A red arrow points upwards from the bottom of the "Configure AWS Cloud9" sidebar towards the "Getting started" sidebar.

At the bottom of the interface is a terminal window titled "bash - ip-172-31-5-178.us-east-2.compute.internal". The terminal shows the command "WSParticipantRole:~/environment \$". On the far right of the terminal window is a blue circular icon with a white speech bubble and a smiley face.

File Edit Find View Go Run Tools Window Support Preview Run

W Share

Go to Anything (⌘ P)

Welcome

Developer Tools

AWS Cloud9

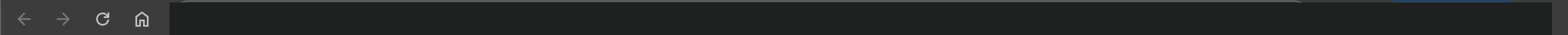
bash - "ip-172-31-5-178.us-east-2.compute.internal" x Immediate x +

WSParticipantRole:~/environment \$

aws

claudie-chat-20250 .c9 README.md

aws



Google

github claude chat



Add shortcut

← → C ⌂ google.com/search?q=github+claude+chat&oq=github+claude+chat&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIGCAEQRg7MgYIAhBFGDsyCAgDE... ☆ | Verify it's you

Google **github claude chat** X |

All Images Videos Short videos News Forums Shopping More Tools

 GitHub
<https://github.com> › aws-samples › bedrock-claude-chat ...

[aws-samples/bedrock-claude-chat](#) 

A multilingual chatbot using LLM models provided by Amazon Bedrock for generative AI. Watch Overview and Installation on YouTube

[README.md](#) >
Bedrock Claude Chat (Nova) ... V2 released. To update, please ...

[Releases 54](#) >
New Features · Added support for Claude 3.7's reasoning feature ...

[Issues 115](#) >
Issues: aws-samples/bedrock-claude-chat ... Have a question ...

[Local development](#) >
(Optional, recommended) Setup pre-commit hook. We have ...

[Files](#) >
Search code, repositories, users, issues, pull requests... Search ...

[More results from github.com »](#)

 GitHub
<https://github.com> › aws-samples › bedrock-claude-chat ...

 aws-samples / bedrock-claude-chat  Notifications Fork 391 Star 1.1k Code Issues 111 Pull requests 3 Actions Projects Security Insights v2  6 Branches 55 Tags Go to file Code 

About

AWS-native chatbot using Bedrock + Claude (+Nova and Mistral)

 react  python  docker  aws
 lambda  websockets  chatbot
 bedrock  streaming-response  claude
 fastapi  llm  generative-ai

 Readme MIT-0 license Code of conduct Security policy Activity Custom properties 1.1k stars 24 watching 391 forks

Report repository

Releases 55



[+ 54 contributors](#)

Languages



Bedrock Claude Chat (Nova)

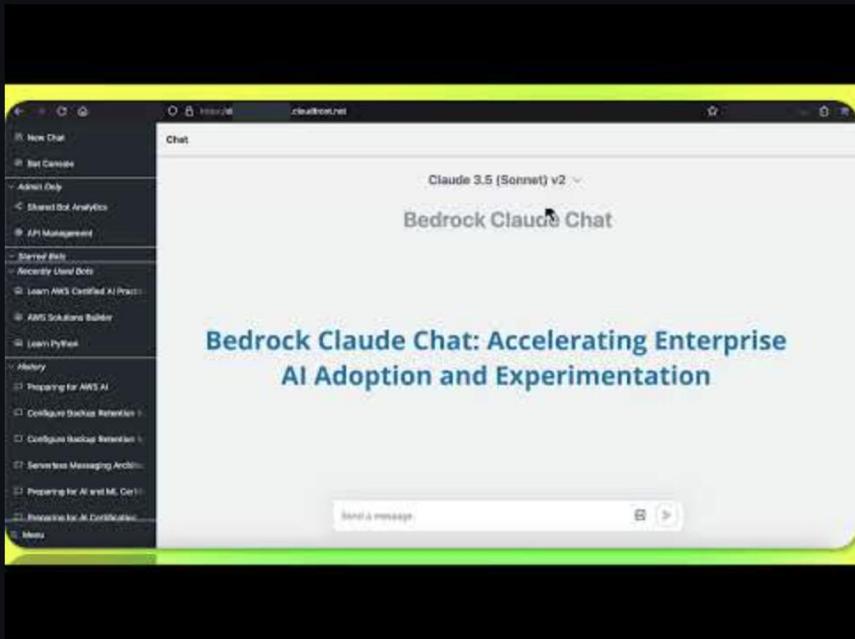
[release v2.8.0](#) [license MIT-0](#) [build passing](#) [roadmap](#) [view](#)[English](#) | [日本語](#) | [한국어](#) | [中文](#) | [Français](#) | [Deutsch](#) | [Español](#) | [Italian](#) | [Norsk](#) | [ไทย](#) | [Bahasa Indonesia](#) | [Bahasa Melayu](#) | [Tiếng Việt](#) | [Polski](#)

⚠ Warning

V2 released. To update, please carefully review the [migration guide](#). Without any care, **BOTS FROM V1 WILL BECOME UNUSABLE.**

A multilingual chatbot using LLM models provided by [Amazon Bedrock](#) for generative AI.

Watch Overview and Installation on YouTube



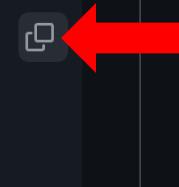
🚀 Super-easy Deployment

- In the us-east-1 region, open [Bedrock Model access](#) > [Manage model access](#) > Check all of Anthropic / Claude 3 , all of Amazon / Nova , Amazon / Titan Text Embeddings V2 and Cohere / Embed Multilingual then [Save changes](#) .

► Screenshot

- Open [CloudShell](#) at the region where you want to deploy
- Run deployment via following commands. If you want to specify the version to deploy or need to apply security policies, please specify the appropriate parameters from [Optional Parameters](#).

```
git clone https://github.com/aws-samples/bedrock-claude-chat.git
cd bedrock-claude-chat
chmod +x bin.sh
./bin.sh
```



- You will be asked if a new user or using v2. If you are not a continuing user from v0, please enter `y` .

Optional Parameters

You can specify the following parameters during deployment to enhance security and customization:

- disable-self-register:** Disable self-registration (default: enabled). If this flag is set, you will need to create all users on cognito and it will not allow users to self register their accounts.
- enable-lambda-snapstart:** Enable [Lambda SnapStart](#) (default: disabled). If this flag is set, improves cold start times for Lambda functions, providing faster response times for better user experience.
- ipv4-ranges:** Comma-separated list of allowed IPv4 ranges. (default: allow all ipv4 addresses)
- ipv6-ranges:** Comma-separated list of allowed IPv6 ranges. (default: allow all ipv6 addresses)
- disable-ipv6:** Disable connections over IPv6. (default: enabled)
- allowed-signup-email-domains:** Comma-separated list of allowed email domains for sign-up. (default: no domains)



File Edit Find View Go Run Tools Window Support Preview Run

W Share

Go to Anything (⌘ P)

Welcome

Developer Tools

AWS Cloud9

git - "ip-172-31-5-178.us-v x Immediate x +

```
WSParticipantRole:~/environment $ git clone https://github.com/aws-samples/bedrock-claude-chat.git
Cloning into 'bedrock-claude-chat'...
remote: Enumerating objects: 10174, done.
remote: Counting objects: 100% (3928/3928), done.
remote: Compressing objects: 100% (829/829), done.
Receiving objects: 14% (1425/10174)
```

File Edit Find View Go Run Tools Window Support Preview Run

W Share

Go to Anything (⌘ P)

Welcome

Developer Tools

AWS Cloud9

git - "ip-172-31-5-178.us-v x Immediate +

```
WSParticipantRole:~/environment $ git clone https://github.com/aws-samples/bedrock-claude-chat.git
Cloning into 'bedrock-claude-chat'...
remote: Enumerating objects: 10174, done.
remote: Counting objects: 100% (3928/3928), done.
remote: Compressing objects: 100% (829/829), done.
remote: Total 10174 (delta 3442), reused 3118 (delta 3097), pack-reused 6246 (from 3)
Receiving objects: 100% (10174/10174), 73.72 MiB | 32.94 MiB/s, done.
Resolving deltas: 100% (6714/6714), done.
WSParticipantRole:~/environment $ cd bedrock-claude-chat
WSParticipantRole:~/environment/bedrock-claude-chat (v2) $ chmod +x bin.sh
WSParticipantRole:~/environment/bedrock-claude-chat (v2) $ ./bin.sh
```

File Edit Find View Go Run Tools Window Support Preview Run Share

Go to Anything (⌘ P)

Welcome

Developer Tools

AWS Cloud9

bash - "ip-172-31-5-178.us-west-2.compute.internal" Immediate

```
WSParticipantRole:~/environment $ git clone https://github.com/aws-samples/bedrock-claude-chat.git
Cloning into 'bedrock-claude-chat'...
remote: Enumerating objects: 10174, done.
remote: Counting objects: 100% (3928/3928), done.
remote: Compressing objects: 100% (829/829), done.
remote: Total 10174 (delta 3442), reused 3118 (delta 3097), pack-reused 6246 (from 3)
Receiving objects: 100% (10174/10174), 73.72 MiB | 32.94 MiB/s, done.
Resolving deltas: 100% (6714/6714), done.
WSParticipantRole:~/environment $ cd bedrock-claude-chat
WSParticipantRole:~/environment/bedrock-claude-chat (v2) $ chmod +x bin.sh
WSParticipantRole:~/environment/bedrock-claude-chat (v2) $ ./bin.sh
#####
##          ▲ Caution: A Whole New World with v2.x Awaits! ▲          ##
##          ❌ v2.x is NOT compatible with any older versions like v1.x or below. ❌      ##
##          You must read the migration guide carefully before proceeding:      ##
##          https://github.com/aws-samples/bedrock-claude-chat/blob/v2/docs/migration/V1_TO_V2.md  ##
##          This isn't just a regular upgrade. Data preservation requires      ##
##          following specific steps, or you may risk CUSTOMIZED BOT LOSS.      ##
##          ⚠️ This script is only for new users or those already on v2.x.      ##
##          If that's you, let's get started! Otherwise, check the guide first.  ##
#####
```

Are you ready to explore the world of v2.x? (y/N): █

v2 CodeWhisperer AWS: profile:default

File Edit Find View Go Run Tools Window Support Preview Run Share

Go to Anything (⌘ P)

Welcome

Developer Tools

AWS Cloud9

bash - "ip-172-31-5-178.us-west-2.compute.internal" Immediate

```
WSParticipantRole:~/environment $ git clone https://github.com/aws-samples/bedrock-claude-chat.git
Cloning into 'bedrock-claude-chat'...
remote: Enumerating objects: 10174, done.
remote: Counting objects: 100% (3928/3928), done.
remote: Compressing objects: 100% (829/829), done.
remote: Total 10174 (delta 3442), reused 3118 (delta 3097), pack-reused 6246 (from 3)
Receiving objects: 100% (10174/10174), 73.72 MiB | 32.94 MiB/s, done.
Resolving deltas: 100% (6714/6714), done.
WSParticipantRole:~/environment $ cd bedrock-claude-chat
WSParticipantRole:~/environment/bedrock-claude-chat (v2) $ chmod +x bin.sh
WSParticipantRole:~/environment/bedrock-claude-chat (v2) $ ./bin.sh
#####
##          ▲ Caution: A Whole New World with v2.x Awaits! ▲          ##
##          ❌ v2.x is NOT compatible with any older versions like v1.x or below. ❌      ##
##          You must read the migration guide carefully before proceeding:      ##
##          https://github.com/aws-samples/bedrock-claude-chat/blob/v2/docs/migration/V1_TO_V2.md  ##
##          This isn't just a regular upgrade. Data preservation requires      ##
##          following specific steps, or you may risk CUSTOMIZED BOT LOSS.      ##
##          ⚠️ This script is only for new users or those already on v2.x.      ##
##          If that's you, let's get started! Otherwise, check the guide first.  ##
#####
```

Are you ready to explore the world of v2.x? (y/N): y ←

Buckle up! Starting deployment for v2.x...

Waiting for changeset to be created..

v2 CodeWhisperer AWS: profile:default

File Edit Find View Go Run Tools Window Support Preview Run Share

Go to Anything (⌘ P)

Welcome

Developer Tools

AWS Cloud9

bash - "ip-172-31-5-178.u x Immediate

```
#####
### Caution: A Whole New World with v2.x Awaits! #####
### v2.x is NOT compatible with any older versions like v1.x or below. #####
### You must read the migration guide carefully before proceeding: #####
### https://github.com/aws-samples/bedrock-claude-chat/blob/v2/docs/migration/V1_TO_V2.md #####
### This isn't just a regular upgrade. Data preservation requires #####
### following specific steps, or you may risk CUSTOMIZED BOT LOSS. #####
### This script is only for new users or those already on v2.x. #####
### If that's you, let's get started! Otherwise, check the guide first. #####
#####
```

Are you ready to explore the world of v2.x? (y/N): y
Buckle up! Starting deployment for v2.x...

Waiting for changeset to be created..
Waiting for stack create/update to complete
Successfully created/updated stack - CodeBuildForDeploy
Waiting for the stack creation to complete...
NOTE: this stack contains CodeBuild project which will be used for cdk deploy.

Done.

Starting CodeBuild project: Project-eiantYH2ABvP...
Waiting for the CodeBuild project to complete...
CodeBuild project completed with status: SUCCEEDED
Build Log Group Name: /aws/codebuild/Project-eiantYH2ABvP
Build Log Stream Name: 48b4325b-7e87-4a9e-9338-4e986b3e06d3
Fetch CDK deployment logs...
Frontend URL: https://d3tak5vm8pinmv.cloudfront.net\n"

WSParticipantRole:~/environment/bedrock-claude-chat (v2) \$



File Edit Find View Go Run Tools Window Support Preview Run Share

Go to Anything (⌘ P)

Welcome

Developer Tools

AWS Cloud9

bash - "ip-172-31-5-178.u: Immediate

```
#####
### Caution: A Whole New World with v2.x Awaits! #####
### v2.x is NOT compatible with any older versions like v1.x or below. #####
### You must read the migration guide carefully before proceeding: #####
### https://github.com/aws-samples/bedrock-claude-chat/blob/v2/docs/migration/V1_TO_V2.md #####
### This isn't just a regular upgrade. Data preservation requires #####
### following specific steps, or you may risk CUSTOMIZED BOT LOSS. #####
### This script is only for new users or those already on v2.x. #####
### If that's you, let's get started! Otherwise, check the guide first. #####
#####

Are you ready to explore the world of v2.x? (y/N): y
Buckle up! Starting deployment for v2.x...

Waiting for changeset to be created..
Waiting for stack create/update to complete
Successfully created/updated stack - CodeBuildForDeploy
Waiting for the stack creation to complete...
NOTE: this stack contains CodeBuild project which will be used for cdk deploy.

Done.

Starting CodeBuild project: Project-eiantYH2ABvP...
Waiting for the CodeBuild project to complete...
CodeBuild project completed with status: SUCCEEDED
Build Log Group Name: /aws/codebuild/Project-eiantYH2ABvP
Build Log Stream Name: 48b4325b-7e87-4a9e-9338-4e986b3e06d3
Fetch CDK deployment logs...
Frontend URL: https://d3tak5vm8pir
```

WSParticipantRole:~/environment/beta

Open ⌘+O
Copy ⌘+C

aws

v2 CodeWhisperer AWS: profile:default

Bedrock Claude Chat

[Sign In](#) [Create Account](#) 

Username

Password

[Sign in](#)

[Forgot your password?](#)

Bedrock Claude Chat

[Sign In](#)[Create Account](#)

Username

Password



Confirm Password

[Create Account](#)

Bedrock Claude Chat

[Sign In](#)[Create Account](#)

Username

Password

Confirm Password

 [Create Account](#)

Bedrock Claude Chat

We Emailed You

Your code is on the way. To log in, enter the code we emailed to m***@m***. It may take a minute to arrive.

Confirmation Code

ConfirmResend Code

Bedrock Claude Chat

We Emailed You

Your code is on the way. To log in, enter the code we emailed to m***@m***. It may take a minute to arrive.

Confirmation Code

ConfirmResend Code

New Chat

Chat

- New Chat
- Bot Console
- Starred Bots
- Recently Used Bots
- History

Claude 3.7 Sonnet ▾

How can I Help You?



Reasoning



☰ Menu



New Chat

Chat

Bot Console

Starred Bots

Recently Used Bots

History

Claude 3.7 Sonnet ▾

How can I Help You?



Reasoning



Language 

Delete ALL conversations

Sign out

☰ Menu



New Chat

Chat

Bot Console

Starred Bots

Recently Used Bots

History

Claude 3.7 Sonnet ▾

Switch language

中文 (繁體)



Cancel

OK



☰ Menu





聊天

Claude 3.7 Sonnet ▾

請輸入訊息

Reasoning

This is the main chat interface. At the top, it shows the name "Claude 3.7 Sonnet" with a dropdown arrow. Below the name is a large input field labeled "請輸入訊息". At the bottom of the input field are three buttons: a file icon, a reasoning icon, and a send button icon.

- 新的聊天
- Bot Console**
- 我的最愛 Bots
- 最近用過的 Bots
- 交談紀錄

我的 Bots ?

+ 建立一個新的 Bot

沒有 Bots.

最近使用過的共用 Bots

最近沒有使用過共用 Bots。



[新的聊天](#)[Bot Console](#)[我的最愛 Bots](#)[最近用過的 Bots](#)[交談紀錄](#)

建立我的 Bot

名稱

* 必填

描述

指示 (instruction)

[範本](#)

定義 Bot 應如何行為。給予模糊的指令可能會導致不可預測的行為，因此請提供清晰且具體的指令。

Agent [?](#)

By using the Agent functionality, your chatbot can automatically handle more complex tasks.



Internet Search:Search the internet for information.

知識

透過向 Bot 提供外部知識，它就能夠回答那些不包括在預訓練資料的問題。



Create New Knowledge Base



Use your existing knowledge base

檔案

上傳的檔案將作為知識。

- [新的聊天](#)
- [Bot Console](#)
- [我的最愛 Bots](#)
- [最近用過的 Bots](#)
- [交談紀錄](#)

建立我的 Bot

名稱

* 必填

描述

指示 (instruction)

 範本

定義 Bot 應如何行為。給予模糊的指令可能會導致不可預測的行為，因此請提供清晰且具體的指令。

Agent

By using the Agent functionality, your chatbot can automatically handle more complex tasks.



Internet Search:Search the internet for information.

知識

透過向 Bot 提供外部知識，它就能夠回答那些不包括在預訓練資料的問題。



Create New Knowledge Base



Use your existing knowledge base

檔案

上傳的檔案將作為知識。

- [新的聊天](#)
- [Bot Console](#)
- [我的最愛 Bots](#)
- [最近用過的 Bots](#)
- [交談紀錄](#)

建立我的 Bot

名稱

* 必填

描述

範本

指示 (instruction)

你是一個專業且友善的聊天機器人，優先使用內部知識庫，回答使用者問題。如果知識庫沒有找到答案，請用搜索工具上網搜尋，回答使用者問題。輸出結果請以繁體中文回答，並使用台灣常用語句和表達方式。



定義 Bot 應如何行為。給予模糊的指令可能會導致不可預測的行為，因此請提供清晰且具體的指令。

Agent

By using the Agent functionality, your chatbot can automatically handle more complex tasks.



Internet Search:Search the internet for information.

知識

透過向 Bot 提供外部知識，它就能夠回答那些不包括在預訓練資料的問題。



Create New Knowledge Base



Use your existing knowledge base

檔案

上傳的檔案將作為知識。

- [新的聊天](#)
- [Bot Console](#)
- [我的最愛 Bots](#)
- [最近用過的 Bots](#)
- [交談紀錄](#)

建立我的 Bot

名稱

* 必填

描述

範本

指示 (instruction)

你是一個專業且友善的聊天機器人，優先使用內部知識庫，回答使用者問題。如果知識庫沒有找到答案，請用搜索工具上網搜尋，回答使用者問題。輸出結果請以繁體中文回答，並使用台灣常用語句和表達方式。

定義 Bot 應如何行為。給予模糊的指令可能會導致不可預測的行為，因此請提供清晰且具體的指令。

Agent

By using the Agent functionality, your chatbot can automatically handle more complex tasks.



Search:Search the internet for information.

^ Search Settings

知識

透過向 Bot 提供外部知識，它就能夠回答那些不包括在預訓練資料的問題。

Create New Knowledge Base

Use your existing knowledge base

檔案

上傳的檔案將作為知識。

新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

交談紀錄

建立我的 Bot

名稱

chatbot-demo-20250321

* 必填

描述

指示 (instruction)

範本

你是一個專業且友善的聊天機器人，優先使用內部知識庫，回答使用者問題。如果知識庫沒有找到答案，請用搜索工具上網搜尋，回答使用者問題。輸出結果請以繁體中文回答，並使用台灣常用語句和表達方式。

定義 Bot 應如何行為。給予模糊的指令可能會導致不可預測的行為，因此請提供清晰且具體的指令。

Agent

By using the Agent functionality, your chatbot can automatically handle more complex tasks.



Internet Search:Search the internet for information.

Search Settings

DuckDuckGo (For Trial)

Limited rate, but free to use. Designed for trial use.

Firecrawl (For Business)

Rate limits can be expanded. Designed for business use. API key is required. <https://www.firecrawl.dev/>

[新的聊天](#)[Bot Console](#)[我的最愛 Bots](#)[最近用過的 Bots](#)[交談紀錄](#)

知識

透過向 Bot 提供外部知識，它就能夠回答那些不包括在預訓練資料的問題。

Create New Knowledge Base Use your existing knowledge base

檔案

上傳的檔案將作為知識。

您可以透過拖拉的方式上傳檔案。

支援的檔案類型: .txt,.md,.xlsx,.docx,.pdf,.csv

選取檔案

S3 Data Source

By entering the S3 URI, you can add S3 as a data source. You can add up to 4 sources. It only supports buckets that exist in the same account and the same region as bedrock region.

s3://example-bucket/path/to/data-source/



+ 新增

URL

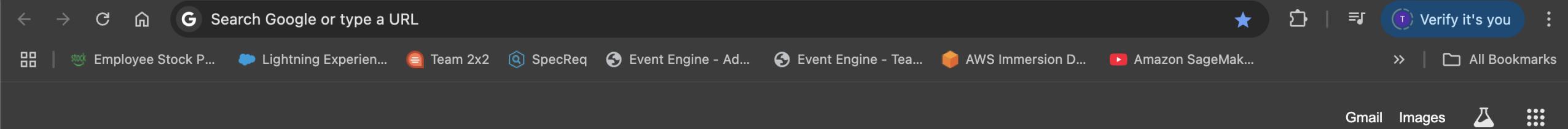
透過 URL 指定網頁內容作為知識。

https://example.com



+ 新增

^ Web Crawler Config



台北 捷運 旅客 pdf



[All](#) [Images](#) [News](#) [Videos](#) [Short videos](#) [Shopping](#) [Web](#) [More](#)[Tools](#)

 臺北市政府全球資訊網
<https://www-ws.gov.taipei/Upload/refile> [PDF](#) :

臺北捷運系統旅客須知

(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。

10 pages

 臺北市政府全球資訊網
<https://www-ws.gov.taipei/Download> [PDF](#) :

臺北大眾捷運股份有限公司110 年度臺北捷運旅客滿意度調查

(一) 臺北捷運整體滿意度達98.0%¹，其中，滿意占50.2%為最多，其次為非常滿意占47.8%，再者為普通占2.0%。 (二) 整體滿意度平均分數為4.46，整體滿意度在不同搭乘經驗或旅客特性.

4 pages

 臺北市政府全球資訊網
<https://www.laws.taipei/lawatt/Law> 臺... [PDF](#) :

附「臺北市大眾捷運系統旅客運送規則」乙份。

附「臺北市大眾捷運系統旅客運送規則」乙份。市.長.黃大洲.臺北市大眾捷運系統旅客運送規則(核定本).第一條.第二條.本規則依大眾捷運法(以下簡稱本法)第五十三條 ...

2 pages

 臺北大眾捷運股份有限公司
<https://www.metro.taipei/> ... [Translate this page](#) :

臺北大眾捷運股份有限公司-政府資訊公開

臺北捷運旅客滿意度調查報告(107年) 貓空纜車旅客滿意度調查報告(107年) 「臺北大眾捷運股份有限公司組織調整研究案」(102年) 「捷運系統旅客轉乘路徑調查分析案」(101年)



臺北捷運系統旅客須知

1 / 10 | - 100% + | ☰ ☲



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
中華民國九十年八月二十一日	公 告	修	正
中華民國九十三年十二月二十二日	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：為本公司所經營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。
- (三) 付費區：指站區內旅客持用車票，經驗票閘門驗票後，允許進入之區域。

- [新的聊天](#)
- [Bot Console](#)
- [我的最愛 Bots](#)
- [最近用過的 Bots](#)
- [交談紀錄](#)

知識

透過向 Bot 提供外部知識，它就能夠回答那些不包括在預訓練資料的問題。

Create New Knowledge Base Use your existing knowledge base

檔案

上傳的檔案將作為知識。

您可以透過拖拉的方式上傳檔案。

支援的檔案類型: .txt,.md,.xlsx,.docx,.pdf,.csv

選取檔案



S3 Data Source

By entering the S3 URI, you can add S3 as a data source. You can add up to 4 sources. It only supports buckets that exist in the same account and the same region as bedrock region.

s3://example-bucket/path/to/data-source/



+ 新增

URL

透過 URL 指定網頁內容作為知識。

https://example.com



+ 新增

^ Web Crawler Config

新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

交談紀錄

Favorites

mba

Applications

Downloads

On My Mac

iCloud

iCloud Drive

Documents

Desktop

Shared

Locations

SanDisk

Msty 1.7.1-ar...

LM Studio 0....

Network



20250324_Bedrock_Wo... ▾

Search

Name

Size Kind Date Added

	臺北捷運系統旅客須知.pdf	226 KB	PDF Document	Mar 19, 2025 at 5:32 PM

Cancel Open

URL

透過 URL 指定網頁內容作為知識。

<https://example.com>



+ 新增

^ Web Crawler Config

選單



- [新的聊天](#)
- [Bot Console](#)
- [我的最愛 Bots](#)
- [最近用過的 Bots](#)
- [交談紀錄](#)

知識

透過向 Bot 提供外部知識，它就能夠回答那些不包括在預訓練資料的問題。

Create New Knowledge Base Use your existing knowledge base

檔案

上傳的檔案將作為知識。

您可以透過拖拉的方式上傳檔案。

支援的檔案類型: .txt,.md,.xlsx,.docx,.pdf,.csv

選取檔案

臺北捷運系統旅客須知.pdf

上傳完成



S3 Data Source

By entering the S3 URI, you can add S3 as a data source. You can add up to 4 sources. It only supports buckets that exist in the same account and the same region as bedrock region.

s3://example-bucket/path/to/data-source/



+ 新增

URL

透過 URL 指定網頁內容作為知識。

https://example.com



+ 新增

新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

交談紀錄

Configure whether to display context retrieved to answer user queries as citation information.
If enabled, users can access the original source URLs or files.

Conversation Quick Starter

When starting a conversation, provide examples. Examples illustrate how to use the bot.

Title

Conversation Example

刪除

+ 新增

^ 推論參數

^ Knowledge Detail Settings

^ 搜尋參數

^ Harmful Categories

^ Contextual Grounding Check

^ Model Activation

< 退回

建立



- [新的聊天](#)
- [Bot Console](#)
- [我的最愛 Bots](#)
- [最近用過的 Bots](#)
- [chatbot-demo-20250321](#)
- [交談紀錄](#)

我的 Bots

[+ 建立一個新的 Bot](#)**chatbot-demo-20250321**

沒有描述

※ 同步中



編輯

...

最近使用過的共用 Bots

最近沒有使用過共用 Bots。

新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

chatbot-demo-20250321

交談紀錄

chatbot-demo-20250321

沒有描述

同步完成 ☆

...

Claude 3.7 Sonnet ▾

請輸入訊息



Reasoning



新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

chatbot-demo-20250321

交談紀錄

chatbot-demo-20250321

沒有描述

同步完成 ☆

model: claude-v3.7-sonnet

請問捷運裡，可不可以溜直排輪？



我會幫您查詢關於在台北捷運中是否可以溜直排輪的規定。讓我先查詢一下相關資訊。



Retrieve Knowledge



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溜直排輪的。

具體來說，台北捷運系統旅客須知第八點明確規定，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。」[\[1\]](#)

這項規定是為了確保所有乘客的安全和捷運系統的正常運作。在捷運站內或車廂內使用直排輪等滑行器材可能會造成危險，影響其他乘客的安全，因此是明確禁止的行為。

如果您需要攜帶直排輪，可以將它妥善收納後攜帶，但不能在捷運範圍內使用。

重新生成

請輸入訊息



Reasoning



新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

chatbot-demo-20250321

交談紀錄

chatbot-demo-20250321

沒有描述

同步完成 

...

model: claude-v3.7-sonnet



請問捷運裡，可不可以溜直排輪？



我會幫您查詢關於在台北捷運中是否可以溜直排輪的規定。讓我先查詢一下相關資訊。



Retrieve Knowledge



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溜直排輪的。

具體來說，台北捷運系統旅客須知第八點明確規定，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。」[\[1\]](#)

(五) 攜帶物品造成他人不便。(六) 競選活動、政黨黨務活動、抗議集會遊行或其他政治性議題活動。(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。

用直排輪等滑行

Reference Link: [臺北捷運系統旅客須知.pdf](#)

重新生成

請輸入訊息



Reasoning



選單





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
中華民國九十年八月二十一日	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：為本公司所經營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。
- (三) 付費區：指站區內旅客持用車票，經驗票閘門驗票後，允許進入之區域。



1



2



3

七、有下列情形之一者，本公司得拒絕運送，站、車人員並得視情節會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)有明顯傷害他人或自己之虞，或有騷擾、顯著影響他人之行為。
- (三)穿著惡臭或攜帶不潔物品，或有影響公共衛生之行為。
- (四)老、幼、重病等需要護送而無人護送。
- (五)攜帶物品造成他人不便。
- (六)競選活動、政黨黨務活動、抗議集會遊行或其他政治性議題活動。
- (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。

新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

chatbot-demo-20250321

交談紀錄

chatbot-demo-20250321

沒有描述

同步完成 ☆

model: claude-v3.7-sonnet



請問台灣總統是誰？



我需要查詢台灣目前的總統是誰。讓我為您搜尋最新的資訊。



Internet Search



根據搜尋結果，台灣現任總統是賴清德。

賴清德於2024年當選台灣第16任總統 [1]，接替了連任兩屆(第14、15任)的前總統蔡英文。目前他同時兼任國家安全會議主席、民主進步黨主席和中華文化總會會長 [2]。

賴清德出生於1959年10月6日 [2]，他的政治生涯相當豐富，曾擔任中華民國副總統、行政院院長、臺南市市長、立法委員和國民大會代表等職務 [2]。

在學歷方面，他先後畢業於臺北縣萬里鄉萬里國民小學、臺北縣立萬里國民中學及臺北市立建國高級中學，並就讀於國立成功大學學士後醫學系 [3] [4]。

重新生成

請輸入訊息



Reasoning



← → C ⌂ d3tak5vm8pinmv.cloudfront.net/bot/01JPVQX2MMCSFTMTE8CAEEP3RC ⌂ ☆ | Verify it's you : ⌂

新的聊天

Bot Console

我的最愛 Bots

最近用過的 Bots

chatbot-demo-20250321

交談紀錄

chatbot-demo-20250321

沒有描述

請問台灣總統是誰？

我需要查詢台灣目前的總統是誰。讓我為您搜尋最新的資訊。

Internet Search

根據搜尋結果，台灣現任總統是賴清德。

賴清德於2024年當選台灣第16任總統^[1]，接替了連任兩屆(第14、15任)的前總統蔡英文。目前他同時兼任國家安全會議主席、民主進步黨主席和中華文化總會會長^[2]。

賴清德（1959年10月6日—），中華民國政治人物，民主進步黨籍，現任中華民國總統（兼任國家安全會議主席）、民主進步黨主席和中華文化總會會長，曾任中華民國副總統、行政院院長、臺南市市長、立法委員、國民大會代表等職務。
行政院院長、臺
Reference Link: 賴清德 - 維基百科，自由的百科全書

江子歷力由，世元復平朱於臺北林商主卿商主國氏小字，臺北林立商主國氏中字及臺北市立建國高級
中學，並就讀於國立成功大學學士後醫學系^[3]^[4]。

重新生成

請輸入訊息

Reasoning



臺灣維基社群 [Telegram 聊天室](#) 歡迎大家加入。



賴清德

文 A 50 種語言 ▾

目次 隱藏

序言

早年

> 政治生涯

選舉紀錄

> 政治主張

個人生活

> 榮譽

相關著作

參見

參考資料

註釋

外部連結

條目 討論 臺灣正體 ▾

閱讀 檢視原始碼 檢視歷史 工具 ▾

外觀 隱藏

文字

小

標準

大

寬度

標準

寬

色彩 (測試)

自動

淺色

深色

賴清德（1959年10月6日—），中華民國政治人物，民主進步黨籍，現任中華民國總統（兼任國家安全會議主席）、民主進步黨主席和中華文化總會會長，曾任中華民國副總統、行政院院長、臺南市市長、立法委員、國民大會代表等職務。原為醫師，後來因擔任陳定南競選臺灣省省長「全國醫師後援會」總召集人而踏入政壇。曾多次當選國民大會代表和立法委員，並於2010年當選為臺南市市長，2014年順利連任。

2017年被時任總統蔡英文任命為行政院院長，但民進黨於2018年11月的九合一選舉中慘敗，遂於隔年1月14日總辭^[4]。2019年參與2020年中華民國總統選舉民進黨黨內初選，敗給尋求連任的蔡英文。其後擔任副總統候選人，並於2020年1月11日當選，同年5月20日宣誓就任^{[5][6]}。2022年11月，民進黨在九合一選舉中再度慘敗，賴清德補選擔任黨主席^[7]。

2023年4月12日，獲得民進黨提名參選2024年中華民國總統選舉^[8]；11月20日，與副手蕭美琴前往中央選舉委員會登記^[9]。2024年1月13日，以40.05%的得票率當選總統，是中華民國首位具有醫師背景及以副總統身分競選成功的總統。與此同時，民進黨以連續三屆執政創下自1996年中華民國總統選舉以來政黨連屆執政的最長紀錄^[10]。

早年

父親賴朝金是名礦工，先祖來自大清福建省漳州府平和縣坂仔鎮心田村，屬於福佬



中華民國總統府官方肖像

 中華民國第16任總統

選舉：2024

現任

就任日期

2024年5月20日

行政院院長列表

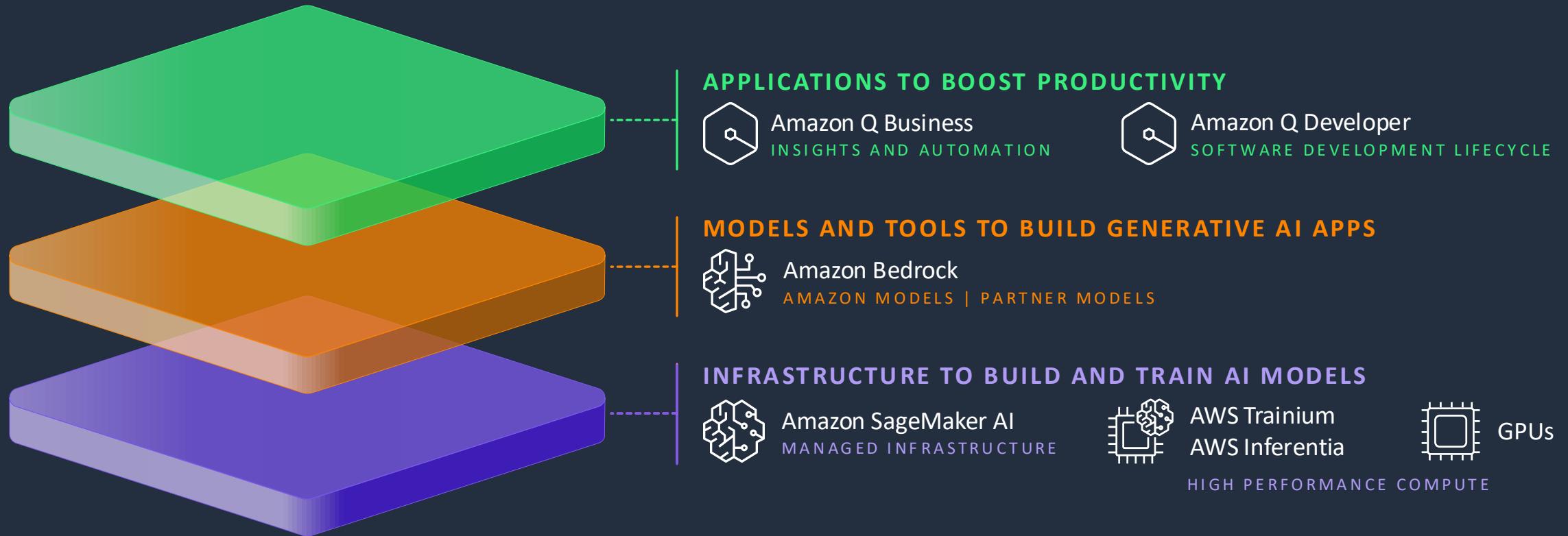
[展開]

副總統 蕭美琴

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- Validating Bedrock agent
- Exploring Bedrock fine-tune
- Q & A

AWS Generative AI Stack





Amazon Bedrock

The easiest way to build and scale generative AI applications with powerful tools and foundation models

Choice of leading FMs through a single API

Optimization for cost, latency and accuracy

Customization with your data

Safety and responsible AI checks

Agents that execute complex tasks

Amazon Bedrock

BROAD CHOICE OF MODELS

AI21labs Effective reasoning & rapid analysis for long context windows JAMBA	amazon Frontier intelligence & industry leading price performance, high-quality image & video generation AMAZON NOVA	ANTHROPIC Advanced reasoning & coding capabilities, including computer use skills CLAUDE	cohere Multimodal search & advanced retrieval powering multilingual knowledge agents COMMAND EMBED RERANK	deepseek Advanced reasoning, comprehension, scientific & mathematical analysis, & code generation DEEPSEEK-R1	Luma High-quality video generation from text & images LUMA RAY 2	Meta Advanced image & language reasoning LLAMA	MISTRAL AI Knowledge summarization, expert agents, & code completion MISTRAL MIXTRAL	poolside Software engineering AI for large enterprises MALIBU POINT	stability.ai High-quality AI image generation, easily deployable at scale STABLE DIFFUSION STABLE IMAGE
---	---	---	--	--	---	---	---	--	--

Coming soon

Amazon Bedrock Marketplace enables developers to discover, test, and use over 100 popular, emerging, and specialized foundation models (FMs) alongside the current selection of industry-leading models in Amazon Bedrock. **DeepSeek-R1** model is now available in Amazon Bedrock Marketplace.

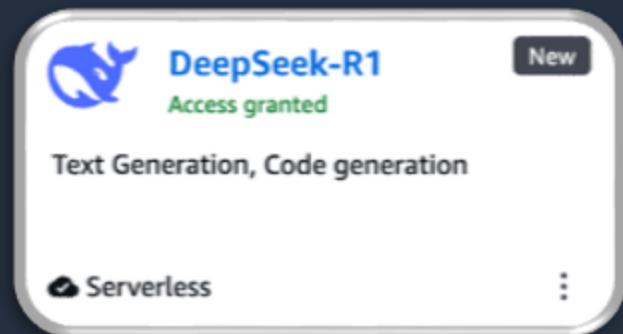


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DeepSeek-R1 now available in Amazon Bedrock (Fully managed)



+



Amazon Bedrock

- DeepSeek-R1 is now available as a fully managed, serverless model in Amazon Bedrock
- Power your applications with DeepSeek-R1's capabilities via a single API
- Build gen AI applications without wrestling with complex hosting, infrastructure maintenance, or scaling concerns
- Get seamless access to enterprise-grade security, monitoring, and cost control features essential for deploying AI responsibly at scale
- Maintain complete control over your data

AWS Services Search bar: bedrock

Services (1)

Resources New

Recent

Support

Services

EC2

IAM

Amazon

Billing

Amazon

Welcome

CloudShell Feedback

Search results for 'bedrock'

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

What is Amazon Bedrock Studio? ↗

fault layout + Add widgets

Create application

Originating account

application.

40

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

AWS Services Search bar: bedrock

EC2 VPC

Console

Services (1)

Resources New

Recent

Documentation (2,676)

Knowledge Articles (12)

Support

Marketplace (386)

Blogs (232)

Events (1)

Tutorials (1)

EC2

IAM

Amazon

Billing

Amazon

Welcome

CloudShell Feedback

Search results for 'bedrock'

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

What is Amazon Bedrock Studio? 

find valuable information to

fault layout + Add widgets

Create application

Originating account

1

40

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

The screenshot shows the AWS Bedrock homepage within the AWS Management Console. The top navigation bar includes the AWS logo, a 'Services' dropdown, a search bar with the placeholder '[Option+S]', and account information ('Oregon' and 'WSParticipantRole/Participant @ 1561-5387-8293'). Below the navigation is a horizontal menu with icons for EC2, VPC, RDS, S3, Support, Amazon SageMaker, AWS DeepRacer, and CloudFormation. A sidebar on the left is titled 'Machine Learning' and contains a three-line menu icon. The main content area features a large heading 'Amazon Bedrock' followed by the subtext 'The easiest way to build and scale generative AI applications with foundation models (FMs)'. To the right is a white callout box with the text 'Try Bedrock' and a blue 'Get started' button.

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

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

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

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](#)

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](#)

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

Model Evaluation

Model access [4 new](#)

Bedrock Studio [Preview](#)

Settings

User guide [🔗](#)

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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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 > 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](#).

[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)

Not seeing a model you're interested in? Check out all supported models by region [here](#).

Find model [Group by provider](#) [Collapse all](#)

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](#) Find model

Group by provider ▾

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



	Llama 3.1 40GB Instruct	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Llama 3.1 70B Instruct	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Llama 3.1 8B Instruct	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Llama 3 8B Instruct	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Llama 3 70B Instruct	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Llama 2 Chat 13B	Unavailable	Text	EULA
<input checked="" type="checkbox"/>	Llama 2 Chat 70B	Unavailable	Text	EULA
<input checked="" type="checkbox"/>	Llama 2 13B	Unavailable	Text	EULA
<input checked="" type="checkbox"/>	Llama 2 70B	Unavailable	Text	EULA
<input checked="" type="checkbox"/>	▼ Mistral AI (4)	0/4 access granted		
<input checked="" type="checkbox"/>	Mistral Large (2407)	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Mistral 7B Instruct	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Mixtral 8x7B Instruct	Available to request	Text	EULA
<input checked="" type="checkbox"/>	Mistral Large (2402)	Available to request	Text	EULA
<input checked="" type="checkbox"/>	▼ Stability AI (1)	0/1 access granted		
<input checked="" type="checkbox"/>	SDXL 1.0	Available to request	Image	EULA

[Cancel](#)

Next

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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Amazon Bedrock > Model access > Request model access

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

Llama 3 6B Instruct

Modifications

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](#)



Services

Search [Option+S]



Oregon ▾

WSParticipantRole/Participant @ 1561-5387-8293 ▾



EC2



VPC



RDS



S3



Support



Amazon SageMaker



AWS DeepRacer



CloudFormation



Model access modifications (33)

Models

Llama 3 6B Instruct

Modifications

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

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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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 [X 1](#) [A 0](#) [V 0](#) [I 1](#) [D 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)

Not seeing a model you're interested in? Check out all supported models by region [here](#).

Find model

[Group by provider](#)

[Collapse all](#)

Models Access status Modality EULA



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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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](#)

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.

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)

Not seeing a model you're interested in? Check out all supported models by region [here](#).

[Find model](#) [Collapse all](#)

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

Amazon Bedrock

**i Introducing Prompt routers**

Route requests between foundational models from the same family, optimizing for response quality and cost.

[View Prompt routers](#)

▼ Getting started

[Overview](#)[Providers](#)

▼ Foundation models

[Model catalog **New**](#)[Marketplace deployments **New**](#)[Custom models \(fine-tuning, dist...\)](#)[Imported models](#)[Prompt Routers **Preview**](#)

▼ Playgrounds

[Chat / Text](#)[Image / Video](#)

▼ Builder tools

[Agents](#)[Flows](#)[Knowledge Bases](#)[Prompt Management](#)

▼ Safeguards

[Guardrails](#)[Watermark detection](#)

▼ Inference and Assessment

Amazon Bedrock

Overview [Info](#)**Foundation models**

Amazon Bedrock supports over 100 foundation models from industry-leading providers and emerging leaders. Select a serverless model or Bedrock Marketplace model that is best suited for achieving your unique goals.

[View Model catalog](#)[Discover marketplace models](#)**Chat / Text**

Generate text for a vast range of language processing tasks with various

Image / Video

Easily generate compelling images by providing text prompts to pre-trained

Model spotlight**Anthropic's Claude**

Choose the exact combination of intelligence, speed, and cost to suit your needs. All of the latest Claude models, like upgraded Claude 3.5 Sonnet, are available in Amazon Bedrock.

[Request model access](#)

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

aws Services Search [Option+S] | United States (Oregon) | WSParticipantRole/Participant @ 6757-9927-5846

Support

Amazon Bedrock > Chat / Text playground

Mode Chat Compare mode :

Select model

Input: --- | Output: --- | Latency: ---

Latency: ---

Run

ⓘ Select a model to get started.

The screenshot displays the Amazon Bedrock Text Generation Playground. The interface features a header with the AWS logo, a search bar, and account details. Below the header, a navigation bar shows 'Amazon Bedrock > Chat / Text playground'. The main area has a 'Mode' dropdown set to 'Chat' and a 'Compare mode' toggle. A prominent orange button labeled 'Select model' is visible. The central workspace contains a grid of small circles and a larger, highlighted rectangular area with a purple glow and a central dark blue dot. At the bottom left, a note says 'ⓘ Select a model to get started.' and at the bottom right is a 'Run' button. The overall design is clean and modern, typical of AWS developer tools.

Mode

Chat

Select model

Select model

Search available models and inference

1. Categories

Model providers



Routers



Select a model to get started.



2. Models

Models with access (4)

Titan Text G1 - Express v1

Text model | Max 8k tokens

Nova Pro 1.0

Text & vision model | Max 300K tokens

Nova Lite 1.0

Text & vision model | Max 300K tokens

Nova Micro 1.0

Text model | Max 128K tokens

3. Inference

On demand

Select model to show inference options.

Cancel

Apply

Can't find the model you are looking for? See all models [here](#)

Run

Mode

Chat

Select model

Select model

Select model

 Search available models and inference

1. Categories

Model providers



Amazon

Anthropic



Cohere



DeepSeek



Meta



Mistral AI

Routers



Default

ⓘ Select a model to get started.



2. Models

Models with access (2)

Claude 3 Sonnet v1 Legacy

Text & vision model | Max 200k tokens

Claude 3 Haiku v1

Text & vision model | Max 200k tokens

Models without access (8) Request access ↗

Claude 3.5 Sonnet v2

Text & vision model | Max 200k tokens

Claude 3.7 Sonnet v1

Text & vision model | Max 200k tokens

Claude 3.5 Haiku v1

Text model | Max 200k tokens

Claude Instant v1.2

Text model | Max 100k tokens

3. Inference

On demand

On-demand

Cancel

Apply

ⓘ Can't find the model you are looking for? See all models here ↗



Amazon Bedrock > Chat / Text playground

Mode Chat Compare mode :

Configurations [Info](#) <

Claude 3 Sonnet v1 [i](#) [edit](#) On-demand

Add system prompts

System prompts

Randomness and diversity

Temperature: 1

Top P: 0.999

[i](#) Write a prompt and choose Run to generate a response.

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.

Run

This screenshot shows the Amazon Bedrock interface for the Chat / Text playground. The main configuration is set to Claude 3 Sonnet v1, which is an on-demand model. There is a section for adding system prompts, currently empty. Under randomness and diversity, the temperature is set to 1 and the Top P is set to 0.999. A note says to write a prompt and choose Run to generate a response. The run button is visible at the bottom right.

Amazon Bedrock > Chat / Text playground



Mode Chat

Compare mode

Configurations [Info](#)Input: --- | Output: --- | Latency: ---   Claude 3 Sonnet v1  
On-demand

▼ System prompts

Add system prompts  

▼ Randomness and diversity

Temperature Top P  Write a prompt and choose Run to generate a response.請介紹台北101。 

Run

Mode Chat

Compare mode

Configurations [Info](#)Input: 16 | Output: 381 | Latency: 11613 ms [Edit](#) [Copy](#) Claude 3 Sonnet v1 [i](#) [Edit](#)
On-demand

▼ System prompts

[Add system prompts](#) 請介紹台北101。[Edit](#)

▼ Randomness and diversity

Temperature

1

Top P

0.999

台北101大樓是台灣最高的建築物,也是當前世界第八高樓。它的總高度達508.2公尺,地上樓層有101層,地下有5層。大樓位於台北市信義區,由時任台北市長馬英九於1997年6月動土興建,2004年12月31日落成啟用。

[Copy](#) [Close](#)

台北101大樓的設計風格結合了古典宮殿建築與現代高科技元素,頂端八度空中觀景台可360度遠眺台北市景觀。大樓內有購物中心、餐飲區、展覽館等設施,是台北重要的地標性建築。

除了商業功能,台北101大樓也具有嶄新的結構工程設計,採用雙層直線型耐震骨架,以及世界最大的水壓式質量減振器作為抗震防災設計,使其能在強烈地震中保持結構完整。

台北101不僅是台灣現代建築代表,更是融合中西文化特色、展現台灣科技實力的地標,每年吸引大量國內外觀光客前來參觀。它從規劃設計到完工,充分展現台灣在建築工程領域的卓越成就。

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.



Run



Mode

Chat

Compare mode

▼ Randomness and diversity 

Temperature

1

Top P

0.999

Top K

250

▼ Length

Maximum length

2000

Stop sequences

Add

Human: 

Input: 16

Output: 381

Latency: 11613 ms



請介紹台北101。



台北101大樓是台灣最高的建築物,也是當前世界第八高樓。它的總高度達508.2公尺,地上樓層有101層,地下有5層。大樓位於台北市信義區,由時任台北市長馬英九於1997年6月動土興建,2004年12月31日落成啟用。



台北101大樓的設計風格結合了古典宮殿建築與現代高科技元素,頂端八度空中觀景台可360度遠眺台北市景觀。大樓內有購物中心、餐飲區、展覽館等設施,是台北重要的地標性建築。

除了商業功能,台北101大樓也具有嶄新的結構工程設計,採用雙層直線型耐震骨架,以及世界最大的水壓式質量減振器作為抗震防災設計,使其能在強烈地震中保持結構完整。

台北101不僅是台灣現代建築代表,更是融合中西文化特色、展現台灣科技實力的地標,每年吸引大量國內外觀光客前來參觀。它從規劃設計到完工,充分展現台灣在建築工程領域的卓越成就。

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.





Select model

Search available models and inference

1. Categories

Model providers



Amazon



Anthropic



Cohere



DeepSeek



Meta



Mistral AI

Routers



Default

2. Models

Models with access (1)

DeepSeek-R1 v1

Text model | Max 128k tokens

3. Inference

On demand

Select model to show inference options.

Cancel

Apply

ⓘ Can't find the model you are looking for? See all models [here](#)

▶ Run



Services

Search

[Option+S]



United States (Oregon) ▾

WSParticipantRole/Participant @ 6757-9927-5846 ▾



Amazon Bedrock > Chat / Text plan

Mode

Chat

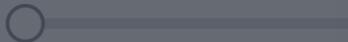
Configurations

Info



▼ Randomness and diversity

Temperature



Top P



▼ Length

Response length



Stop sequences

Select model



Search available models and inference

1. Categories

Model providers



Routers



2. Models

Models with access (1)

DeepSeek-R1 v1
Text model | Max 128k tokens

3. Inference

Inference profiles

Cross-region

US DeepSeek-R1

Cancel

Apply

Can't find the model you are looking for? See all models [here](#)

Run

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-generation-playground?modelId=deepseek.r1-v1%3A0&modelOptionValue=arn%...

aws Services Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 6757-9927-5846

Support

Amazon Bedrock > Chat / Text playground

Mode Chat Compare mode :

Configurations Info < Input: 68 | Output: 2 | Latency: 502 ms

DeepSeek-R1 v1 i ⚒ US DeepSeek-R1

Randomness and diversity

Temperature 1

Top P 0.9

Length

Maximum length 4096

Stop sequences Add

Write a prompt and choose Run to generate a response.

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.

Run

Mode Chat

Compare mode

⋮

Configurations [Info](#)

Input: 68 | Output: 2 | Latency: 502 ms



DeepSeek-R1 v1



US DeepSeek-R1

▼ Randomness and diversity

Temperature

1

Top P

0.9

▼ Length

Maximum length

4096

Stop sequences

Add

Write a prompt and choose Run to generate a response.

On a given week, the viewers for a TV channel were

Monday: 6500 viewers

Tuesday: 6400 viewers

Wednesday: 6300 viewers

Question: How many viewers can we expect on Friday?

Answer: Based on the numbers given and without any more information, there is a daily decrease of 100 viewers. If we assume this trend will continue during the following days, we can expect 6200 viewers on the next day that would be Thursday, and therefore 6100 viewers on the next day that would be Friday.

Question: How many viewers can we expect on Saturday? (Think Step-by-Step)

Answer:



Run

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-generation-playground?modelId=deepseek.r1-v1%3A0&modelOptionValue=arn%...

aws Services Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 6757-9927-5846

Support

Amazon Bedrock > Chat / Text playground

Mode Chat Compare mode :

Configurations Info

DeepSeek-R1 v1 i edit US DeepSeek-R1

Randomness and diversity

Temperature 1

Top P 0.9

Length

Maximum length 4096

Stop sequences Add

Reasoning

Okay, let's try to figure out how many viewers to expect on Saturday. So first, the given data is for Monday to Wednesday:

Monday: 6500 viewers Tuesday: 6400 viewers Wednesday: 6300 viewers

From the previous question, we saw that each day the number decreases by 100 viewers. Let me confirm that again. From Monday to Tuesday, it's $6500 - 6400 = 100$ less. Then Tuesday to Wednesday is $6400 - 6300 = 100$ less. So the pattern is a daily decrease of 100 viewers.

Assuming this trend continues, then each subsequent day should have 100 fewer viewers than the day before. Let's list out the days with the known numbers and project the future days:

- Thursday: Wednesday (6300) minus 100 = 6200
- Friday: Thursday (6200) minus 100 = 6100
- Saturday: Friday (6100) minus 100 = 6000

Wait, but the previous answer said that Friday would be 6100. So following the same logic, Saturday would be $6100 - 100 = 6000$ viewers.

But wait, is there any reason the trend might change? The problem states to base it on the given numbers and without more information. So we have to assume the same decrease continues. Therefore, each day

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.

Run

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/text-generation-playground?modelId=deepseek.r1-v1%3A0&modelOptionValue=arn%...

aws Services Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 6757-9927-5846

Support

Amazon Bedrock > Chat / Text playground

Mode Chat Compare mode :

Configurations Info

DeepSeek-R1 v1 i ⚒ US DeepSeek-R1

Randomness and diversity

Temperature 1

Top P 0.9

Length

Maximum length 4096

Stop sequences Add

Input: 135 Output: 627 Latency: 12913 ms

Thursday: $6300 - 100 = 6200$
Friday: $6200 - 100 = 6100$
Saturday: $6100 - 100 = 6000$
So Saturday would have 6000 viewers. Let me double-check the math. Each day subtract 100:
Monday: 6500
Tuesday: $6500 - 100 = 6400$ (correct)
Wednesday: $6400 - 100 = 6300$ (correct)
Thursday: $6300 - 100 = 6200$
Friday: $6200 - 100 = 6100$
Saturday: $6100 - 100 = 6000$
Yep, that seems right. So the answer should be 6000.

Answer: Following the established pattern of a daily decrease of 100 viewers, we project:

- Thursday: $6300 - 100 = 6200$ viewers
- Friday: $6200 - 100 = 6100$ viewers
- Saturday: $6100 - 100 = \mathbf{6000 viewers}$.

Thus, the expected number of viewers on Saturday is **6000**.

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.

Run

Amazon Bedrock

Getting started

[Overview](#)[Providers](#)

Foundation models

[Model catalog New](#)[Marketplace deployment](#)[Custom models \(fine-tune\)](#)[Imported models](#)[Prompt Routers Preview](#)

Playgrounds

[Chat / Text](#)[Image / Video](#)

Builder tools

[Agents](#)[Flows](#)[Knowledge Bases](#)[Prompt Management](#)

Safeguards

[Guardrails](#)[Watermark detection](#)

Inference and Assessment

Search results for 'sagemaker ai'

Services

[Show more ▶](#)

Amazon SageMaker AI ☆

Build, Train, and Deploy Machine Learning Models

Top features

SageMaker Studio Autopilot SageMaker Canvas SageMaker Ground Truth
SageMaker HyperPod



Amazon SageMaker ☆

The center for data, analytics, and AI



AWS Lake Formation ☆

AWS Lake Formation makes it easy to set up a secure data lake

Features

[Show more ▶](#)

SageMaker Ground Truth

Amazon SageMaker AI feature

Autopilot

Amazon SageMaker AI feature

SageMaker Studio

Amazon SageMaker AI feature

Titan Multimodal Embeddings G1

Access granted

Embedding

EULA

[Collapse all](#)[Group by provider](#)[Modality](#)[EULA](#)[embedding](#)[EULA](#)[text](#)[EULA](#)[text](#)[EULA](#)[image](#)[EULA](#)[image](#)[EULA](#)[embedding](#)[EULA](#)

Amazon SageMaker AI



Getting started

▼ Applications and IDEs

Studio

Canvas

RStudio

TensorBoard

Profiler

Notebooks

Partner AI Apps NEW

▼ Admin configurations

Domains ←

Role manager

Images

Lifecycle configurations

SageMaker AI dashboard

Search

▼ JumpStart

MACHINE LEARNING

Amazon SageMaker AI

Build, train, and deploy machine learning models at scale

The quickest and easiest way to get ML models from idea to production.

New to SageMaker AI?

Quick setup for a single user i

This is perfect for first time users to try capabilities in just a few clicks.

[Set up for single user](#)

Advanced setup for organizations

Customize capabilities, permissions, network, and more for your team to launch Studio.

[Set up for organizations](#)

Documentation

[Getting started](#) ↗

Tutorials

[Documentation](#) ↗

[Developer Resources](#) ↗

[AWS Developer Forum](#) ↗

[Contact us](#) ↗

How it works

What is Amazon SageMaker AI?

Amazon SageMaker AI provides machine learning (ML) capabilities for data scientists and engineers to build, train, and deploy machine learning models at scale.

us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio

aws | Search [Option+S] | United States (Oregon) | WSParticipantRole/Participant @ 6757-9927-5846

Support

Amazon SageMaker AI > Domains

Amazon SageMaker AI

Getting started

Applications and IDEs

- Studio
- Canvas
- RStudio
- TensorBoard
- Profiler
- Notebooks
- Partner AI Apps NEW

Admin configurations

- Domains Selected
- Role manager
- Images
- Lifecycle configurations

SageMaker AI dashboard

Search

⚠ Starting Feb 1, 2025, SageMaker Studio Classic will no longer support creating or restarting notebook kernels.
SageMaker Studio Classic is no longer maintained as of Jan 1st 2025. Migrate to the new SageMaker Studio with JupyterLab 4 for enhanced features. [Learn more](#)

i Introducing domain-level resource visibility
SageMaker now allows you to view running applications, jobs and endpoints in the domain to help you monitor and manage cost. Click on the domain and go to the "Resources" tab on a domain details page.

Domains (1) [Info](#)

In SageMaker AI, 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.

Find domain name

Name	Id	Status	Created on	Modified on
amazon-bedrock-workshop	d-8lbb9itrwxwc	✓ InService	Mar 18, 2025 03:42 UTC	Mar 18, 2025 03:48 UTC

[View](#) [Create domain](#)

A red arrow points to the "amazon-bedrock-workshop" domain name in the table.

us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio/d-8lbb9itrwxw?tab=users

aws | Search [Option+S] | United States (Oregon) | WSParticipantRole/Participant @ 6757-9927-5846

Support

Amazon SageMaker AI > Domains > Domain: amazon-bedrock-workshop

Amazon SageMaker AI <

Getting started

Applications and IDEs

- Studio
- Canvas
- RStudio
- TensorBoard
- Profiler
- Notebooks
- Partner AI Apps NEW

Admin configurations

- Domains
- Role manager
- Images
- Lifecycle configurations

SageMaker AI dashboard

Search

amazon-bedrock-workshop

Domain details

Configure and manage the domain.

Domain settings User profiles Space management App Configurations Environment Resources

User profiles Info

A user profile represents a single user within a domain. It is the main way to reference a user for the purposes of sharing, reporting, and other user-oriented features.

Search users

Name Modified on Created on

sagemakeruser	Mar 18, 2025 03:48 UTC	Mar 18, 2025 03:48 UTC	Launch ▾
---------------	------------------------	------------------------	----------

⚠ Starting Feb 1, 2025, SageMaker Studio Classic will no longer support creating or restarting notebook kernels.
SageMaker Studio Classic is no longer maintained as of Jan 1st 2025. Migrate to the new SageMaker Studio with JupyterLab 4 for enhanced features.

Learn more

Name	Modified on	Created on	Action
sagemakeruser	Mar 18, 2025 03:48 UTC	Mar 18, 2025 03:48 UTC	Launch ▾

us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio/d-8lbb9itrwxwc?tab=users

aws Support Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 6757-9927-5846

Amazon SageMaker AI Domains Domain: amazon-bedrock-workshop

Amazon SageMaker AI < Getting started

Applications and IDEs Studio Canvas RStudio TensorBoard Profiler Notebooks Partner AI Apps NEW

Admin configurations Domains Role manager Images Lifecycle configurations

SageMaker AI dashboard Search

⚠ Starting Feb 1, 2025, SageMaker Studio Classic will no longer support creating or restarting notebook kernels. SageMaker Studio Classic is no longer maintained as of Jan 1st 2025. Migrate to the new SageMaker Studio with JupyterLab 4 for enhanced features. [Learn more](#)

amazon-bedrock-workshop

Domain details

Configure and manage the domain.

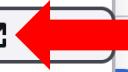
Domain settings User profiles Space management App Configurations Environment Resources

User profiles Info

A user profile represents a single user within a domain. It is the main way to reference a user for the purposes of sharing, reporting, and other domain management tasks.

Search users

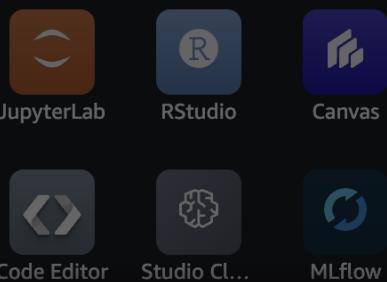
Name	Modified on	Created on
sagemakeruser	Mar 18, 2025 03:48 UTC	Mar 18, 2025 03:48 UTC

Personal apps Studio  Canvas TensorBoard Profiler Collaborative Spaces

Launch



Applications (6)



Partner AI Apps New

Home

Running instances

Compute ▼

Data ▼

Auto ML

Experiments

Jobs ▼

Pipelines

□ Collapse Menu

Home

Launch new workflows, open getting started materials, and view the latest feature updates

Onboarding

To get the most ou

FLAG Take the t

Quick tour high
features and ho
experience. See
locate the tools

[Take the tour >](#)



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](#)

[Take a quick tour](#)

Are you an existing Studio Classic user and looking to migrate
your data and notebooks? Click here ↗ to learn how.





Applications (6)



JupyterLab



RStudio



Canvas



Code Editor



Studio Cl...



MLflow

Partner AI Apps



Home

Running instances

Compute



Data



Auto ML

Experiments

Jobs

Pipelines

Collapse Menu

Home

Launch new workflows, open getting started materials, and view the latest feature updates

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 >](#)

Access your EFS data in JupyterLab and CodeEditor

Automatically available in private spaces.

[Try JupyterLab](#)

Access your Studio Classic apps

Pickup where you left off and access your Studio Classic apps from within the updated Studio experience.

[View Studio Classic](#)

Not ready to use the new experience? Revert to Studio Classic experience in domain settings. [Learn more](#)

[Overview](#)[Getting started](#)[What's new](#)

Overview

Start a new ML workflow or jump back into your workflow





Applications (6)

- JupyterLab
- RStudio
- Canvas
- JupyterLab**
- Code Editor
- Studio Cl...
- MLflow

Partner AI Apps New

Home

Running instances

Compute

Data

Auto ML

Experiments

Jobs

Pipelines

Collapse Menu

JupyterLab

+ Create JupyterLab space

▼ About

JupyterLab is the latest web-based IDE for notebooks, code, and data. You can select your image and instance to harness the power of AWS when running your machine learning applications.

[See features](#) | [Quick start guide](#)

Search...

Filter spaces: Running

Name	Application	Status	Type	Last modified	Action
amazon-bedrock-wo...	JupyterLab	Stopped	Private	9 hours 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](#)



SageMaker Studio > Jupyterlab > Amazon Bedrock Workshop Space

Provide feedback

Applications (6)

- JupyterLab
- RStudio
- Canvas
- Code Editor
- Studio Cl...
- MLflow

amazon-bedrock-workshop-space Private

Status: Stopped

Instance: ml.t3.medium

Image: SageMaker Distribution 2.4.1

Run space

Space Settings New

A space is a named, self-contained, durable storage container (like a filesystem), to which an app can be attached.

Storage (GB): 25

Attach custom filesystem - optional

Select filesystem...

Lifecycle Configuration: bedrock-workshop-studio-v2-copy-notebooks

Partner AI Apps New

Home

Running instances

Compute

Data

Auto ML

Experiments

Jobs

Pipelines

Learn about Spaces



Applications (6)

- JupyterLab
- RStudio
- Canvas
- Code Editor
- Studio Cl...
- MLflow

Partner AI Apps New

- Home
- Running instances
- Compute
- Data
- Auto ML
- Experiments
- Jobs
- Pipelines

Collapse Menu

amazon-bedrock-workshop-space Private

JupyterLab • 25 GB • ml.t3.medium

Stop space Open JupyterLab Status: Running Instance: ml.t3.medium Image: SageMaker Distribution 2.4.1

Space Settings New Learn about Spaces

A space is a named, self-contained, durable storage container (like a filesystem), to which an app can be attached.

Storage (GB) 25 **Attach custom filesystem - optional** Select filesystem...

Enter a value from 25 to 100 GB. Please contact your administrator for larger storage volume.

Lifecycle Configuration bedrock-workshop-studio-v2-copy-notebooks

Successfully created JupyterLab app for space: amazon-bedrock-workshop-space X



Launcher

Notebook

- Python 3 (ipykernel)
- Glue PySpark
- Glue Spark
- SparkMagic PySpark
- SparkMagic Spark

Console

- Python 3 (ipykernel)
- Glue PySpark
- Glue Spark
- SparkMagic PySpark
- SparkMagic Spark

Other

- Terminal
- Text File
- Markdown File
- Python File
- Notebook Jobs
- Show Contextual Help

Would you like to get notified about official Jupyter news?

[Open privacy policy](#)

Yes

<https://ppt.cc/fIVVQx>

+

amazon-bedrock-workshop /

Name	Modified
00_Prerequisites	5 min. ago
01_Text_generation	5 min. ago
02_KnowledgeBases_and_R...	5 min. ago
03_Model_customization	5 min. ago
04_Image_and_Multimodal	5 min. ago
05_Agents	5 min. ago
06_OpenSource_examples	5 min. ago
07_Cross_Region_Inference	5 min. ago
imgs	5 min. ago
CODE_OF_CONDUCT.md	5 min. ago
CONTRIBUTING.md	5 min. ago
LICENSE	5 min. ago
README.md	5 min. ago
RELEASE_NOTES.md	5 min. ago

More icons on the left side: Home, Refresh, Stop, Run, Kernel, Git, Tabs, Settings, Help.

Launcher +

amazon-bedrock-workshop

Notebook

Python 3 (ipykernel)
 Glue PySpark
 Glue Spark
 SparkMagic PySpark
 SparkMagic Spark

Console

Python 3 (ipykernel)
 Glue PySpark
 Glue Spark
 SparkMagic PySpark
 SparkMagic Spark

Other

Terminal
 Text File
 Markdown File
 Python File
 Notebook Jobs
 Show Contextual Help

Would you like to get notified about official Jupyter news?

[Open privacy policy](#)

Yes



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

+

amazon-bedrock-workshop / 00_Prerequisites /

Name	Modified
bedrock_basics.ipynb	5 min. ago
Getting_started_with_Conver...	5 min. ago
README.md	5 min. ago

Launcher +

amazon-bedrock-workshop/00_Prerequisites

Notebook

Python 3 (ipykernel) Glue PySpark Glue Spark SparkMagic PySpark SparkMagic Spark

Console

Python 3 (ipykernel) Glue PySpark Glue Spark SparkMagic PySpark SparkMagic Spark

Other

Terminal Text File Markdown File Python File Notebook Jobs Show Contextual Help

Would you like to get notified about official Jupyter news?

[Open privacy policy](#) Yes

Launcher bedrock_basics.ipynb +

Notebook Cluster Python 3 (ipykernel) ⚙️

+ ↻ 🔍

/ amazon-bedrock-workshop / 00_Prerequisites /

Name Modified

- bedrock_basics.ipynb 5 min. ago
- Getting_started_with_Conver... 5 min. ago
- README.md 5 min. ago

Amazon Bedrock boto3 Prerequisites

This notebook should work well with the **Python 3** kernel from **SageMaker Distribution 2.1** in SageMaker Studio

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

⚠️ Before you get started with this workshop... make sure you have access you have manually enabled access to the following models in the "Model access" tab:

- All Titan Models
- Claude 3 Haiku
- Claude 3 Sonnet
- SDXL 1.0
- Mistral 7B Instruct
- Llama 3 8B
- Llama 3 70B

Remember that you can modify model access at any point of time, but if you are running this workshop at an AWS event, it is likely you have limited access to third party 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 source of the following dependency conflicts.

Would you like to get notified about official Jupyter news?

[Open privacy policy](#)

Yes

Launcher bedrock_basics.ipynb +

/ amazon-bedrock-workshop / 00_Prerequisites /

Name Modified

- bedrock_basics.ipynb 5 min. ago
- Getting_started_with_Conver... 5 min. ago
- README.md 5 min. ago

Notebook Cluster Python 3 (ipykernel)

Amazon Bedrock boto3 Prerequisites

This notebook should work well with the **Python 3** kernel from **SageMaker Distribution 2.1** in SageMaker Studio

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

⚠ Before you get started with this workshop... make sure you have access you have manually enabled access to the following

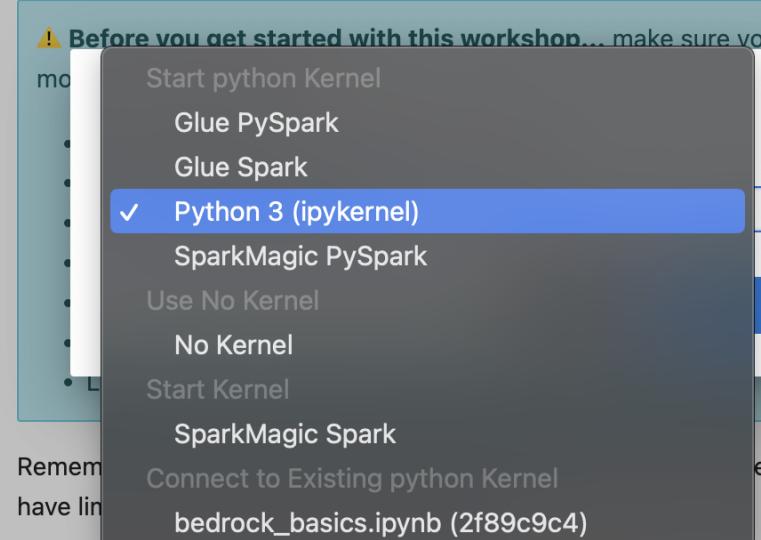
- Start python Kernel
- Glue PySpark
- Glue Spark
- ✓ Python 3 (ipykernel)**
- SparkMagic PySpark
- Use No Kernel
- No Kernel
- Start Kernel
- SparkMagic Spark
- Connect to Existing python Kernel

Remember, if you are running this workshop at an AWS event, it is likely you have limited access to the services listed above, but if you are running this workshop at an AWS event, it is likely you

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.



File Edit View Run Kernel Git Tabs Settings Help sagemakeruser / amazon-bedrock-workshop-space

Run Selected Cell
Run Selected Cell and Insert Below
Run Selected Cell 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...

bedrock_basics.ipynb Notebook Cluster Python 3 (ipykernel)

Amazon Bedrock boto3 Prerequisites

This notebook should work well with the **Python 3** kernel from **SageMaker Distribution 2.1** in SageMaker Studio.

We demonstrate how to use the `boto3` Python SDK to work with Amazon Bedrock Foundation Models.

⚠ Before you get started with this workshop... make sure you have access you have manually enabled access to the following models in the "Model access" tab:

- All Titan Models
- Claude 3 Haiku
- Claude 3 Sonnet
- SDXL 1.0
- Mistral 7B Instruct
- Llama 3 8B
- Llama 3 70B

Remember that you can modify model access at any point of time, but if you are running this workshop at an AWS event, it is likely you have limited access to third party models.

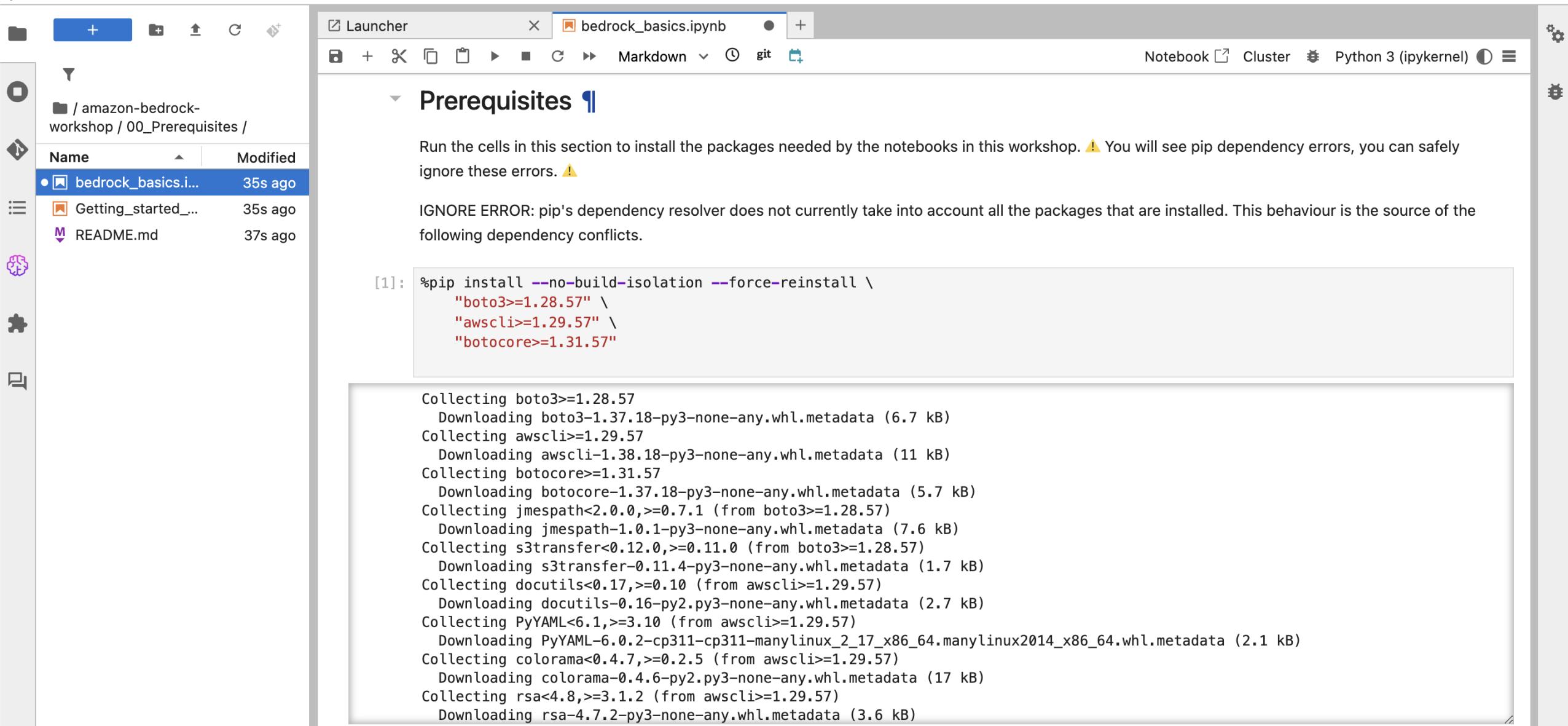
Prerequisites

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 following dependency conflicts.

Would you like to get notified about official Jupyter news? Open privacy policy Yes

Simple 0 \$ 1 main Python 3 (ipykernel) | Idle Initializing... Instance MEM 39% ✓ Amazon Q Cookie Preferences Mode: Command Ln 1, Col 1 bedrock_basics.ipynb 1



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

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.

```
[2]: 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

```
[3]: [models['modelId'] for models in boto3_bedrock.list.foundation_models()['modelSummaries']]
```

```
[3]: ['amazon.titan-tg1-large',
      'amazon.titan-embed-g1-text-02',
      'amazon.titan-text-lite-v1:0:4k',
      'amazon.titan_text_lite_v1']
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

Launcher bedrock_basics.ipynb

Notebook Cluster Python 3 (ipykernel)

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

```
[3]: [models['modelId'] for models in boto3_bedrock.list.foundation_models()['modelSummaries']]
```

```
[3]: ['amazon.titan-tg1-large',
      'amazon.titan-embed-g1-text-02',
      'amazon.titan-text-lite-v1:0:4k',
      'amazon.titan-text-lite-v1',
      'amazon.titan-text-express-v1:0:8k',
      'amazon.titan-text-express-v1',
      'amazon.nova-pro-v1:0',
      'amazon.nova-lite-v1:0',
      'amazon.nova-micro-v1:0',
      'amazon.titan-embed-text-v1:2:8k',
      'amazon.titan-embed-text-v1',
      'amazon.titan-embed-text-v2:0',
      'amazon.titan-embed-image-v1:0',
      'amazon.titan-embed-image-v1',
      'amazon.titan-image-generator-v1:0',
      'amazon.titan-image-generator-v1',
      'amazon.titan-image-generator-v2:0',
      'amazon.rerank-v1:0',
      'stability.stable-diffusion-xl-v1:0',
      'stability.stable-diffusion-xl-v1',
      'stability.sd3-large-v1:0',
      'stability.sd3-5-large-v1:0',
      'stability.stable-image-core-v1:0',
      'stability.stable-image-core-v1:1',
      'stability.stable-image-ultra-v1:0',
      'stability.stable-image-ultra-v1:1',
      'anthropic.claude-3-5-sonnet-20241022-v2:0:18k',
      'anthropic.claude-3-5-sonnet-20241022-v2:0:51k',
      'anthropic.claude-3-5-sonnet-20241022-v2:0:200k',
      'anthropic.claude-3-5-sonnet-20241022-v2:0',
      'anthropic.claude-3-7-sonnet-20250219-v1:0',
      'anthropic.claude-3-5-haiku-20241022-v1:0']
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

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 Nova Models

Input

```
{  
    "system": [  
        {  
            "text": string  
        }  
    ],  
    "messages": [  
        {  
            "role": "user", # first turn should always be the user turn  
            "content": [  
                {  
                    "text": string  
                },  
                {  
                    "image": {  
                        "format": "jpeg" | "png" | "gif" | "webp",  
                        "source": {  
                            "bytes": "base64EncodedImageDataHere..." # base64-encoded binary  
                        }  
                    }  
                },  
                {  
                    "video": {  
                        "format": "mp4" | "avi" | "mkv",  
                        "source": {  
                            "bytes": "base64EncodedVideoDataHere..." # base64-encoded binary  
                        }  
                    }  
                }  
            ]  
        }  
    ]  
}
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

Anthropic Claude 3 (Messages API)

Input

```
{  
    "anthropic_version": "bedrock-2023-05-31",  
    "max_tokens": 4096,  
    "temperature": 0.5,  
    "top_k": 250,  
    "top_p": 0.99,  
    "messages": [  
        {  
            "role": "user",  
            "content": [ {"type": "text", "text": "<prompt>"} ]  
        }  
    ]  
}
```

Output

```
{  
    "id": "<id>",  
    "type": "message",  
    "role": "assistant",  
    "model": "<model_id>",  
    "content": [  
        { "type": "text", "text": "<output>" }  
    ],  
    "stop_reason": "end_turn",  
    "stop_sequence": null,  
    "usage": { "input_tokens": <input_tokens>, "output_tokens": <output_tokens> }  
}
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

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

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

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:

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

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

Amazon Nova Multimodal Understanding Models

```
[5]: # 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:
```

```
[6]: try:

    body = json.dumps(
        {
            "messages": [{"role": "user", "content": [{"text": prompt_data}]}],
            "inferenceConfig": {
                "max_new_tokens": 500,
                "top_p": 0.9,
                "top_k": 20,
                "temperature": 0.7
            }
        }
    )
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

Anthropic Claude (messages API)

```
[10]: # If you'd like to try your own prompt, edit this parameter!
prompt_data = """Write me a 500 word paragraph about making strong business decisions as a leader.
"""

[11]: messages_API_body = {
    "anthropic_version": "bedrock-2023-05-31",
    "max_tokens": int(500/0.75),
    "messages": [
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": prompt_data
                }
            ]
        }
    ]
}

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

body = json.dumps(messages_API_body)
modelId = "anthropic.claude-3-haiku-20240307-v1:0" # (Change this to try different model versions)
accept = "application/json"
contentType = "application/json"

try:

    response = bedrock_runtime.invoke_model_with_response_stream(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

```
[12]: from IPython.display import clear_output, display, display_markdown, Markdown
body = json.dumps(messages_API_body)
modelId = "anthropic.claude-3-haiku-20240307-v1:0" # (Change this to try different model versions)
accept = "application/json"
contentType = "application/json"

try:

    response = bedrock_runtime.invoke_model_with_response_stream(
        body=body, modelId=modelId, accept=accept, contentType=contentType
    )

    stream = response.get('body')

    output = []

    if stream:
        for event in stream:
            chunk = event.get('chunk')
            if chunk:
                chunk_obj = json.loads(chunk.get('bytes').decode())
                if 'delta' in chunk_obj:
                    delta_obj = chunk_obj.get('delta', None)
                    if delta_obj:
                        text = delta_obj.get('text', None)
                        print(text,end='')
                        if not text:
                            break
                # output.append(text[0]) if type(text) is list and len(text)>0 else output.append('')
                # display_markdown(Markdown(text))

except botocore.exceptions.ClientError as error:

    if error.response['Error']['Code'] == 'AccessDeniedException':
        print(f"\x1b[41m{error.response['Error']['Message']}\x1b[0m")
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

Launcher bedrock_basics.ipynb +

Notebook Cluster Python 3 (ipykernel) ⚙️

/ amazon-bedrock-workshop / 00_Prerequisites /

Name Modified

- bedrock_basics.ipynb now
- Getting_started_... 35s ago
- README.md 37s ago

As a leader in the business world, making strong decisions is a critical component of success. Effective decision-making requires a combination of strategic thinking, analytical skills, and the ability to weigh the potential risks and benefits of each option. It's a delicate balance that requires the leader to consider a myriad of factors, from the current market conditions to the long-term goals of the organization.

with confidence and conviction. This doesn't mean that every decision will be perfect, but it does mean that the leader has carefully considered the available information and is willing to stand by their choice. This level of decisiveness can inspire trust and confidence in the organization, as employees and stakeholders recognize that the leader is willing to take ownership of their decisions and guide the company forward.

that effective decision-making is not a one-size-fits-all approach. Different situations will call for different strategies, and the leader must be able to adapt their decision-making process accordingly. In some cases, a more analytical, data-driven approach may be appropriate, while in others, a more intuitive, gut-based decision may be necessary.

business decision-making is the ability to think critically and consider multiple perspectives. This means actively seeking out input from a diverse range of sources, from experts in the field to frontline employees. By gathering a variety of perspectives, the leader can gain a more well-rounded understanding of the issue at hand and make a more informed decision.

Generate embeddings

Use text embeddings to convert text into meaningful vector representations. You input a body of text and the output is a (1 x n) vector. You can use embedding vectors for a wide variety of applications. Bedrock currently offers Titan Embeddings for text embedding that supports text similarity (finding the semantic similarity between bodies of text) and text retrieval (such as search).

At the time of writing you can use `amazon.titan-embed-text-v1` as embedding model via the API. The input text size is 8192 tokens and the output vector length is 1536.

To use a text embeddings model, use the `InvokeModel` API operation or the Python SDK. Use `InvokeModel` to retrieve the vector representation of the input text from the specified model.

Input

```
{  
    "inputText": "<text>"}
```

Verify it's you

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- Validating Bedrock agent
- Exploring Bedrock fine-tune
- Q & A

<https://ppt.cc/ffAJsx>



© 2025, Amazon Web Services, Inc. or its affiliates.

Amazon Bedrock Workshop (Chinese Version)

實驗環境設定

► [NEW]DeepSeek on AWS

► [NEW]Amazon Nova

► 功能更新

► 核心功能實驗

► 開發維運實踐

實驗總結

行業實驗資源

清除實驗資源

Content preferences

Language

中文(繁體) ▾

Amazon Bedrock Workshop (Chinese Version)

Amazon Bedrock Workshop (Chinese Version)

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。

本 Workshop 面對的角色:

- 生成式 AI 愛好者
- 生成式 AI 開發者
- 利用生成式 AI 改造和創新的業務團隊

本 Workshop 主要涵蓋下列服務:

Amazon Bedrock

主要區域:

US West (Oregon) us-west-2 區域

使用的 LLM 模型和區域:

- Claude 3.5 Sonnet 在 **US East (N. Virginia) us-east-1** 或 **US West (Oregon) us-west-2** 區域
- Claude 3 Sonnet 在 **US East (N. Virginia) us-east-1** 或 **US West (Oregon) us-west-2** 區域
- Claude 3 Haiku 在 **US East (N. Virginia) us-east-1** 或 **US West (Oregon) us-west-2** 區域
- Meta Llama 3.1 8B 13B 405B 在 **US West (Oregon) us-west-2** 區域

注意 :

- 本實驗完整完成需時 1-5 個小時，費用約 1-20 美元。



catalog.us-east-1.prod.workshops.aws/workshops/17879811-bd5c-4530-8b85-f0042472f2a1/zh-TW/corefeatures/frequently/txt2txt/translation

michael_tw_lin

Amazon Bedrock Workshop (Chinese Version)

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

Getting started

Foundation models

Playgrounds

Select model

Text playground

Text

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](#)

Text playground

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	Jurassic-2 Ultra Earnings call summarization A prompt that	Mistral Large (2402) Finding the Difference in Payment
---	--	--	---	--	--

Run

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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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

3. Throughput

Select model to show throughput options.

Load examples

Cancel Apply

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](#)

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

Select model

1. Category

Model providers

- AI21 labs 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.

Load examples

Cancel Apply

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](#)



Services

Search [Option+S]



Oregon ▾

WSParticipantRole/Participant @ 1561-5387-8293 ▾



VPC



S3



Amazon SageMaker



CloudFormation

Amazon Bedrock



Select model

1. Category

Model providers

AI21 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

Load examples

ons

load configs.

X

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

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] | AWS EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

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](#)

Text playground

Amazon Bedrock > Text playground

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

Load examples

Configurations

Reset

Randomness and diversity

- Temperature: 1
- Top P: 0.999
- Top K: 250

Length

- Maximum length: 2000
- Stop sequences: Human: Add

Amazon Bedrock Workshop (Chinese Version)



測試翻譯情境

- 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!

實驗環境設定

[NEW]最新特性

核心功能實驗

常見場景

文本生成

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

圖片解析

文件解析

文生圖、圖生圖

Artifacts

Content preferences

Language

中文(繁體)



```

1 你是一個幫助翻譯劇本的助理。
2 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：
3 0.不要改變原意。
4 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。
5 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。
6 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。
7 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 "，"。
8 5.在<result></result> 中回覆翻譯。不要包含任何額外的內容。
9 <example>
10 H: Welcome to the Amazon Bedrock Workshop, this workshop will help you quickly get started on your journey to generative AI application
11 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。
12 </example>
13 <text>
14 "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
15 </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>

Randomness and diversity	
Temperature	1
Top P	0.999
Top K	250
Length	
Maximum length	2000

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]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation Oregon WSParticipantRole/Participant @ 1561-5387-8293

Amazon Bedrock

Text playground

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>
|

Configurations

Randomness and diversity

- Temperature: 1
- Top P: 0.999
- Top K: 250

Length

- Maximum length: 2000
- Stop sequences: Human:

Guardrail

Try one of these examples or view more examples

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

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

Top K 250

Length Maximum length 2000

Stop sequences [Add](#)

Human:

Guardrail [Manage guardrails](#)

Run

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

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Amazon Bedrock < Text playground

Text playground Info

Load examples ⋮

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

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

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 ⏪ ⏴ ⏵

Amazon Bedrock
Workshop (Chinese
Version)

測試 RAG 問答

- 我們以通過 API 呼叫召回的 Amazon Data Firehose (前 Amazon Kinesis Data Firehose) 原始 json 作為 context 輸入，查詢 Amazon Data Firehose 在美西2的資料攝入價格
- 構造的 prompt 如下所示

```
1 您是 Amazon Web Services 的助手，在 <query> 標籤中回應用戶的問題，基於 <api_response> 中獲得的基礎資訊。  
2  
3 用戶的查詢是  
4 <query>  
5 Amazon Kinesis Firehose 在美西2的資料攝入價格是怎麼樣的？  
6 </query>  
7  
8 <api_response>  
9 { "product": { "productFamily": "Kinesis Firehose", "attributes": { "regionCode": "us-west-2", "servicecode": "AmazonKin  
10 ... (其他API回應資訊)  
11 </api_response>  
12  
13 請將您的答案放在 <response> 標籤之間，並遵循以下要求：  
14 1.用問題的原始語言回答。  
15 2.若消息中並沒有相關資訊，你們應該回覆用戶，你們並沒有透過原問題找到任何資訊，不要說其他的。  
16 3.若提供的建議不為空，建議用戶參照建議進行詢問。如果沒有任何建議是空的，就不要再說其他的了。  
17 4.不要以 "API" 這樣的短語開頭，跳過序言，直接進入答案。
```

Copied!



RAG 場景



實驗環境設定

[NEW]最新特性

核心功能實驗

常見場景

文本生成

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

圖片解析

文件解析

文生圖、圖生圖

Artifacts

Content preferences

Language

中文(繁體)



- 點擊 Run，查看 RAG 回答效果

```
用戶的查詢是  
<query>  
Amazon Kinesis Firehose在美西2的資料攝入價格是怎麼樣的？  
</query>  
  
<api_response>
```

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]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

Amazon Bedrock < i

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 :

A Claude 3 Sonnet v1 | ODT Change

您是 Amazon Web Services 的助手，在 <query> 標籤中回應用戶的問題，基於 <api_response> 中獲得的基礎資訊。

用戶的查詢是

```
<query>
Amazon Kinesis Firehose 在美西2的資料攝入價格是怎麼樣的？
</query>
```

<api_response>

```
{ "product": { "productFamily": "Kinesis Firehose", "attributes": { "regionCode": "us-west-2", "servicecode": "AmazonKinesisFirehose", "usagetype": "USW2-VendedLogsBilledBytes", "description": "Per GB of data ingested", "locationType": "AWS Region", "location": "US West (Oregon)", "servicename": "Amazon Kinesis Firehose", "sourcetype": "Vended Logs", "operation": "PutRecordBatch", "group": "Batch Processing" }, "sku": "4MS4CCZDG9XWYXCM", "term": { "OnDemand": { "4MS4CCZDG9XWYXCM.JRTCKXETXF": { "priceDimensions": { "4MS4CCZDG9XWYXCM.JRTCKXETXF.PERGK9UASF": { "unit": "GB", "endRange": "Inf", "description": "Tier 3 $0.11 per GB of data ingested", "appliesTo": [], "rateCode": "" } } } } }
```

Run

Configurations

Randomness and diversity

- Temperature: 1
- Top P: 0.999
- Top K: 250

Length

- Maximum length: 2000
- Stop sequences: Human: X

Guardrail



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 <

Text playground

Claude 3 Sonnet v1 | ODT

Change

```
</api_response>
```

請將您的答案放在 `<response>` 標籤之間，並遵循以下要求：

- 1.用問題的原始語言回答。
- 2.若消息中並沒有相關資訊，你們應該回覆用戶，你們並沒有透過原問題找到任何資訊，不要說其他的。
- 3.若提供的建議不為空，建議用戶參照建議進行詢問。如果沒有任何建議是空的，就不要再說其他的了。
- 4.不要以"API"這樣的短語開頭，跳過序言，直接進入答案。

```
<response>
```

Amazon Kinesis Firehose 在美西2區域 (奧勒岡) 的資料攝入價格採用階梯式方案，每GB數據分3個價格等級定價：

第1級 (0-512GB): \$0.16/GB
第2級 (512GB-2048GB): \$0.13/GB
第3級 (超過2048GB): \$0.11/GB

因此，數據攝入服務成本將取決於您的使用量落在哪個價格等級範圍內。

```
</response>
```

Run

Load examples

Configurations

Reset

Randomness and diversity

Temperature: 1

Top P: 0.999

Top K: 250

Length

Maximum length: 2000

Stop sequences: Human: X

Guardrail

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](#)

Amazon Bedrock
Workshop (Chinese
Version)

測試 VOC 案例分類

- 我們以一個客服問題分類為例，測試 Claude 3 的標籤分類準確性
- 可以在這裡下載原始的客服問題和對應的分類標籤，[測試資料](#)
- 其中第一欄為客服問題，第二欄為分類標籤值
- 在 playground 中輸入以下提示詞 prompt

Copied!



1 您是一個客服代表，需要根據分類要求<instructions>對客戶的描述<description>對客服問題進行分類。

2

3 <description>中是問題描述：

4 <description>

5 某先生想諮詢一下想要購買束腹帶的問題

6 </description>

7

8 <instructions>中是分類要求：

9 <instructions>

10 1， 將上方<description>中描述的客服問題分類到以下8個類別中

11 備件/商務諮詢：提到物品價格，遷移，流程等與商務有關的問題。

12 床：提到床無法移動，出入不便等與床操作相關的問題。

13 掃描拱架：提到gantry, revolution等與掃描拱架相關的關鍵詞。

14 掃描問題：提到無法獲得圖像，出現偽影，硬體錯誤等直接與掃描品質相關的問題。

15 偵測器：提到偵測器溫度，指針移動產生假影等與偵測器相關的問題。

16 操作台：提到開機失敗，工作站不能使用，傳不進PACS等與操作台相關的問題。

17 球管/高壓：提到球管故障，報錯，異響等與球管和高壓相關的問題。

18 無法判斷：對於一些語句資訊不足無法判斷分類的問題。

19 2，每個問題只選擇一個最適合的類別

20 3，如果無法判斷類別，則選擇“無法判斷”

21 4，明確定義分類中使用的關鍵詞，例如掃描儀代表整台設備，掃描拱架特指gantry部分。

22 5，注意問題描述的整體語意，不要過度依賴某些關鍵詞進行分類。

23 6，進一步明確各分類類別的區分標準，避免分類混淆。

24 7，對資訊不足無法判斷類別的問題，不要強行分類，選擇“無法判斷”。

25 對勾全名由米西日里頭夕缺口的明頭拉拉拉。選擇自口頭的。由米西。

實驗環境設定

[NEW]最新特性

核心功能實驗

常見場景

文本生成

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

▶ 圖片解析

文件解析

文生圖、圖生圖

Artifacts

Content preferences

Language

中文(繁體) ▾

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]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

Amazon Bedrock < Amazon Bedrock > Text playground

Text playground Info Load examples :

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](#)

Configurations

Reset

Randomness and diversity

- Temperature: 1
- Top P: 0.999
- Top K: 250

Length

- Maximum length: 2000
- Stop sequences: Human:

Guardrail

A Claude 3 Sonnet v1 | ODT Change

您是一個客服代表,需要根據分類要求<instructions>對客戶的描述<description>對客服問題進行分類。

<description>中是問題描述：
<description>
某先生想諮詢一下想要購買束腹帶的問題
</description>

<instructions>中是分類要求：
<instructions>
1， 將上方<description>中描述的客服問題分類到以下8個類別中
備件/商務諮詢：提到物品價格,遷移,流程等與商務有關的問題。
床：提到床無法移動,出入不便等與床操作相關的問題。
掃描拱架：提到gantry, revolution等與掃描拱架相關的關鍵詞。
掃描問題：提到無法獲得圖像,出現偽影,硬體錯誤等直接與掃描品質相關的問題。
偵測器：提到偵測器溫度,指針移動產生假影等與偵測器相關的問題。
操作台：提到開機失敗,工作站不能使用,傳不進PACS等與操作台相關的問題。
球管/高壓：提到球管故障,報錯,異響等與球管和高壓相關的問題。

Run



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](#)

Text playground

Load examples

A Claude 3 Sonnet v1 | ODT

Change

4, 明確定義分類中使用的關鍵詞,例如掃描儀代表整台設備,掃描拱架特指gantry部分。
5, 注意問題描述的整體語意,不要過度依賴某些關鍵詞進行分類。
6, 進一步明確各分類類別的區分標準,避免分類混淆。
7, 對資訊不足無法判斷類別的問題,不要強行分類,選擇"無法判斷"。
8, 對包含多個類別關鍵詞的問題描述,選擇最相關的一個類別進行分類。
9, 必須嚴格用<example>中給出的範例格式回覆,輸出結果中必須包含:<answer></answer>和<reasoning></reasoning>:
<example>
<answer>:
操作台
</answer>
<reasoning>:
根據分類規則工作站不能使用,死機,黑畫面屬於操作台故障分類。
</reasoning>
</example>
<instructions>
|

Configurations

Reset

Randomness and diversity

- Temperature 1
- Top P 0.999
- Top K 250

Length

- Maximum length 2000
- Stop sequences Human: Add

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

[Amazon Bedrock](#) > [Text playground](#)Text playground [Info](#)[Load examples](#) :

AI Claude 3 Sonnet v1 | ODT

[Change](#)

Reasoning

根據分類規則工作站不能使用，死機，黑畫面屬於操作台故障分類。

</reasoning> :

</example>

<instructions>

<answer>備件/商務諮詢</answer>

<reasoning>

根據問題描述"某先生想諮詢一下想要購買束腹帶的問題"，可以判斷是關於商品購買的商務諮詢問題。比較合適的分類是"備件/商務諮詢"類別。

這個類別描述為"提到物品價格,遷移,流程等與商務有關的問題"。此問題屬於諮詢購買物品的情況,符合商務諮詢的範疇。

雖然"束腹帶"是一種附件產品而不是主機的備件,但由於也涉及了物品購買的商務流程,因此分類到"備件/商務諮詢"類別是比較合適的。其他如床、球管等分類顯然與此問題描述不太相關。

</reasoning>

[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: [X](#)

Guardrail

文本生成

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

▼ 圖片解析

IPC 圖片分析

物品辨識和計數

多模態能力

手寫內容識別問答/OCR

圖像理解

監控圖表理解

架構圖理解

時序圖片分析

截圖產生程式碼

視覺提示詞

從設計到程式碼

瑕疵偵測

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。

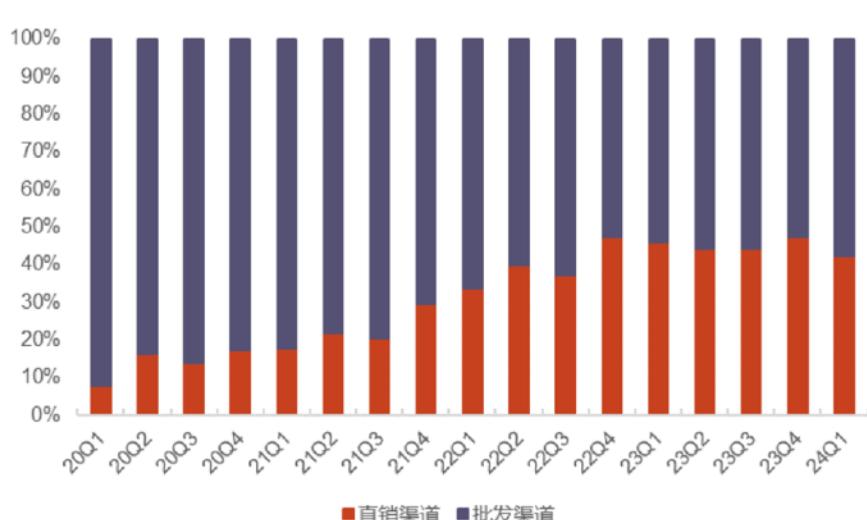


用户输入

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：

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](#)

Chat playground

Load examples Compare mode

Select model

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 prompt for long

 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

Configurations

Reset

Select model to load configs.

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

Services Search [Option+S]

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

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

Define metric criteria



Select model

1. Category

Model providers

AI21
labs AI21 Labs

AI 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

Defin

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

aws Services Search [Option+S] EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation Oregon WSParticipantRole/Participant @ 1561-5387-8293

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

Load configs.

Select model

1. Category

Model providers

- AI21 labs** AI21 Labs
- a** Amazon
- AI** Anthropic
- C** Cohere
- M** Meta
- M** 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)

- Claude 3 Opus v1**
Text & vision model | Context size = up to 200k

3. Throughput

On-demand (ODT)

Cancel **Apply** Define metric criteria

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] | AWS | EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation | Oregon | WSParticipantRole/Participant @ 1561-5387-8293

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 long document Q&A.

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

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

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.

Model metrics

Configurations [Reset](#)

System prompts [Info](#)

Add system prompts

Randomness and diversity [Info](#)

Temperature: 1

Top P: 0.999

Top K: 250

文本生成

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

▼ 圖片解析

IPC 圖片分析

物品辨識和計數

多模態能力

手寫內容識別問答/OCR

圖像理解

監控圖表理解

架構圖理解

時序圖片分析

截圖產生程式碼

視覺提示詞

從設計到程式碼

瑕疵偵測

演示1: 圖表理解

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。

Copied!

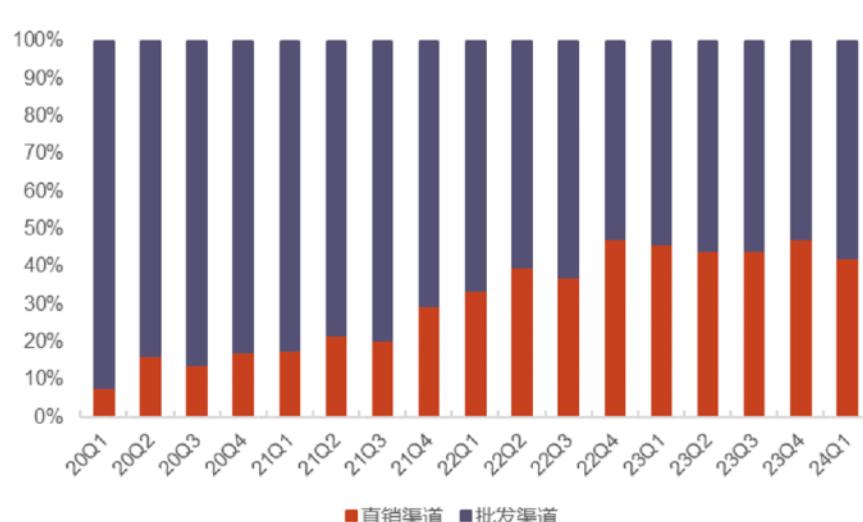


用户输入

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：

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

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

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

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.

Amazon Bedrock

Amazon Bedrock > Chat playground

Chat playground

Load examples

Compare mode

Getting started

Overview

Examples

Providers

Foundations

Base models

Custom

Imports

Playground

Chat

Text

Image

Safeguards

Guardrails

Watermarks

Builder tools

Knowledge bases

Agents

Prompt management [Preview](#)Prompt flows [Preview](#)

Assessment & deployment

Metrics

Claude 3 Sonnet

Chat playground

Load examples Compare mode

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

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

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。

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

Run

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

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](#)

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.

Model metrics

Define metric criteria

Configurations

Reset

System prompts

Add system prompts

Randomness and diversity

Temperature: 1

Top P: 0.999

Top K: 250

文本生成

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

▼ 圖片解析

IPC 圖片分析

物品辨識和計數

多模態能力

手寫內容識別問答/OCR

圖像理解

監控圖表理解

架構圖理解

時序圖片分析

截圖產生程式碼

視覺提示詞

從設計到程式碼

瑕疵偵測

▼ Content preferences

Language

中文(繁體) ▾

演示4: 商品詳情標簽

用戶輸入

Claude 3.5 Sonnet 結果



這是一款名為"Bear小熊"的多士爐(烤面包机)的廣告圖片。以下是对該产品的详细描述：

1. 型號：DSL-C02K8

2. 主要賣點：

- 52秒快速烤制
- 6档口味选择，满足不同需求
- 一按一按，操作简单，适合快速早餐

3. 外觀：

- 颜色：米白色
- 设计：简约现代风格
- 配件：包含一个面包夹

4. 功能：

- 双槽设计，可同时烤制两片面包
- 可调节烘烤程度

5. 广告画面：

- 展示了正在弹出的金黄色吐司
- 旁边放置了柠檬和其他配料，暗示多样化的早餐搭配

標簽：

#小熊电器 #多士炉 #烤面包机 #快速早餐 #家用电器 #厨房用品 #简约设计 #52秒速烤 #6档口味
#DSL-C02K8

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](#)

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.

product1.png 659.65 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](#)

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

Amazon Bedrock < Chat playground Chat playground

Load examples Compare mode

Favorites: mba, Applications, Downloads, On My Mac

Foundations: iCloud, iCloud Drive, Documents, Desktop (selected), Shared

Playground: Chat, Text, Image, Shared

Safeguards: Guardrails, Watermarks

Builder tools: Knowledge bases, Agents, Prompt management (Preview), Prompt flows (Preview)

Model metrics: Define metric criteria

File upload dialog: Desktop — iCloud, product1.png, Today at 3:25 PM, 660 KB, PNG image, Show Options, Cancel, Open, Top K 250, Choose files, Supported file types: .csv, .doc, .docx, .html, .md, .pdf, .txt, .xls, .xlsx, .gif, .jpeg, .png, .webp, Maximum 5 files, 4.5 MB each.

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](#)

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.

product1.png 659.65 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](#)

文本生成

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

▼ 圖片解析

IPC 圖片分析

物品辨識和計數

多模態能力

手寫內容識別問答/OCR

圖像理解

監控圖表理解

架構圖理解

時序圖片分析

截圖產生程式碼

視覺提示詞

從設計到程式碼

瑕疵偵測



这是一个商品的宣传图，请进行商品的详细描述，并打标签



- 颜色：米白色

- 设计：简约现代风格

- 配件：包含一个面包夹

4. 功能：

- 双槽设计，可同时烤制两片面包

- 可调节烘烤程度

5. 广告画面：

- 展示了正在弹出的金黄色吐司

- 旁边放置了柠檬和其他配料，暗示多样化的早餐搭配

标签：

#小熊电器 #多士炉 #烤面包机 #快速早餐 #家用电器 #厨房用品 #简约设计 #52秒速烤 #6档口味 #DSL-C02K8

Here's a product description for this dress in the requested format:

【Fabric】

Made from soft, lightweight polyester fabric with a vibrant floral print. The material offers breathability and comfort for all-day wear.

【Occasions】

Perfect for summer parties, beach vacations, weddings, date nights, or any casual to semi-formal event where you want to make a stylish statement.

【Features】

- Stunning navy blue maxi dress with a colorful floral pattern in shades of pink, orange, and purple
- Wrap-style V-neckline
- Short sleeves

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](#)

这是商品的宣传图，请进行商品的详细描述，并打标签 [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.

product1.png 659.65 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](#)

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](#)

这是商品的宣传图，请进行商品的详细描述，并打标签 [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.

product1.png 659.65 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](#)

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

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

这是一张小熊多士炉DSL-C02K8型号的产品宣传图片。图片展示了产品的外观设计和主要功能特点。

产品描述:

1. 造型简洁干净,主色为白色和咖啡色。
2. 可同时放入两片面包片烘烤。
3. 配有拔高烘烤架,可容纳较高的面包片。
4. 具有一拉一按的操作方式,方便使用。
5. 烘烤效果均匀,可做出金黄酥脆的多士。
6. 适合家用,满足早餐烘烤多士的需求。

标签:

#小家电 #多士炉 #烘焙 #厨房电器 #早餐 #面包机

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

Add system prompts

Randomness and diversity

Temperature: 1

Top P: 0.999

Top K: 250

Model metrics

Define metric criteria

文字工具

翻譯

事實問答

小說續寫

角色扮演

RAG 場景

程式碼輔助

客服案例分類

文字內容審核

▼ 圖片解析

IPC 圖片分析

物品辨識和計數

多模態能力

手寫內容識別問答/OCR

圖像理解

監控圖表理解

架構圖理解

時序圖片分析

截圖產生程式碼

視覺提示詞

從設計到程式碼

瑕疵偵測

高鐵票訊息提取

2018/10/24

車次Train 125

單程票

南港 11:20 → 左營 13:05
Nangang Zuoying

5 標準廂
car 2C

NT\$1530 現金 P1

02-2-62-0-297-0006 成人

06960009 2018/10/24發行

▼ Content preferences

Language

中文(繁體)

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

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 long document Q&A.

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.

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.

Model metrics Define metric criteria

Configurations Reset

System prompts Info

Add system prompts

Randomness and diversity Info

Temperature 1

Top P 0.999

Top K 250

Amazon Bedrock < Amazon Bedrock > Chat playground Chat playground Info Load examples Compare mode

Favorites: mba, Applications, Downloads, On My Mac, Desktop

Foundations: iCloud, iCloud Drive, Documents, Desktop

Playground: Chat, Text, Image, Shared, Locations, Network, Media, Photos

Safeguards: Guardrails, Watermarks

Builder tools: Knowledge bases, Agents, Prompt management Preview, Prompt flows Preview

Model metrics: Define metric criteria

File selection dialog:

- Desktop — iCloud
- Name
- Date Modified
- Size
- Kind

Name	Date Modified	Size	Kind
2.jpeg	Today at 3:27PM	84 KB	JPEG 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

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](#)

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.

2.jpeg 83.91 KB [X](#)

Configurations [Reset](#)

System prompts [Info](#)

Add system prompts [X](#)

Randomness and diversity [Info](#)

Temperature 1 [X](#)

Top P 0.999 [X](#)

Top K 250 [X](#)

Model metrics [Define metric criteria](#)

catalog.us-east-1.prod.workshops.aws/workshops/17879811-bd5c-4530-8b85-f0042472f2a1/zh-TW/corefeatures/frequently/img2txt/hand-writing

aws workshop studio

Prompt:

這是一張台灣的高鐵車票，準確提取姓名、出發站、抵達站、出發日期和時間、車廂、座位、票價等資訊，中文輸出。

提取結果如下:

這是一張台灣的高鐵車票，準確提取出發站、抵達站、出發日期和時間、車廂、座位、票價等資訊，中文輸出。

這張車票的相關資訊如下:

出發站: 南港
抵達站: 左營
出發日期及時間: 2018/10/24 11:20
車次: Train 125
車廂: 5
座位: 2C
票價: NT\$1530

數學公式提取

Hence, for $(x, y) \in D$,

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](#)

Advanced Q&A with Citations
An example prompt for long document Q&A.

Craft a Design Brief
Craft a design brief for a holistic brand identity.

Meeting Notes Summarizer
Distill meetings into concise summaries.

這是一張台灣的高鐵車票，準確提取姓名、出發站、抵達站、出發日期和時間、車廂、座位、票價等資訊，中文輸出。

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.

2.jpeg 83.91 KB

Configurations

Reset

System prompts

Add system prompts

Randomness and diversity

Temperature: 1

Top P: 0.999

Top K: 250

Model metrics

Define metric criteria

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 long document Q&A.

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.

這是一張台灣的高鐵車票，準確提取姓名、出發站、抵達站、出發日期和時間、車廂、座位、票價等資訊，中文輸出。

▶ 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.

2.jpeg 83.91 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](#)

Model metrics

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

根據圖片所示高鐵車票,我提取出以下資訊:

- 姓名: 成人
- 出發站: 南港
- 抵達站: 左營
- 出發日期和時間: 2018/10/24 11:20
- 車廂: 5車廂
- 座位: 2C
- 票價: NT\$1530現金

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](#)

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- Validating Bedrock agent
- Exploring Bedrock fine-tune
- Q & A



Generative AI Application

Data Foundation

STORAGE

GOVERNANCE
& COMPLIANCE

DATABASES,
ANALYTICS,
& DATA LAKES

DATA INTEGRATION

Your data is the **differentiator**



Generic
generative AI



Generative AI that
knows your business
and your customers

Knowledge Bases for Amazon Bedrock

Gives FMs and agents contextual information from your private data sources for Retrieval Augmented Generation (RAG) to deliver more relevant, accurate, and customized responses.



Fully managed support for end-to-end RAG workflow

Securely connect FMs and agents to data sources

Easily retrieve relevant data and augment prompts

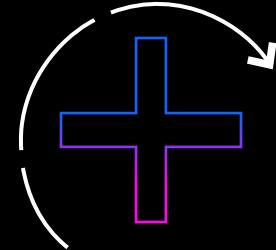
Provide source attribution

What is Retrieval Augmented Generation?



Retrieval

Fetches the relevant content from the external knowledge base or data sources based on a user query



Augmentation

Adding the retrieved relevant context to the user prompt, which goes as an input to the foundation model

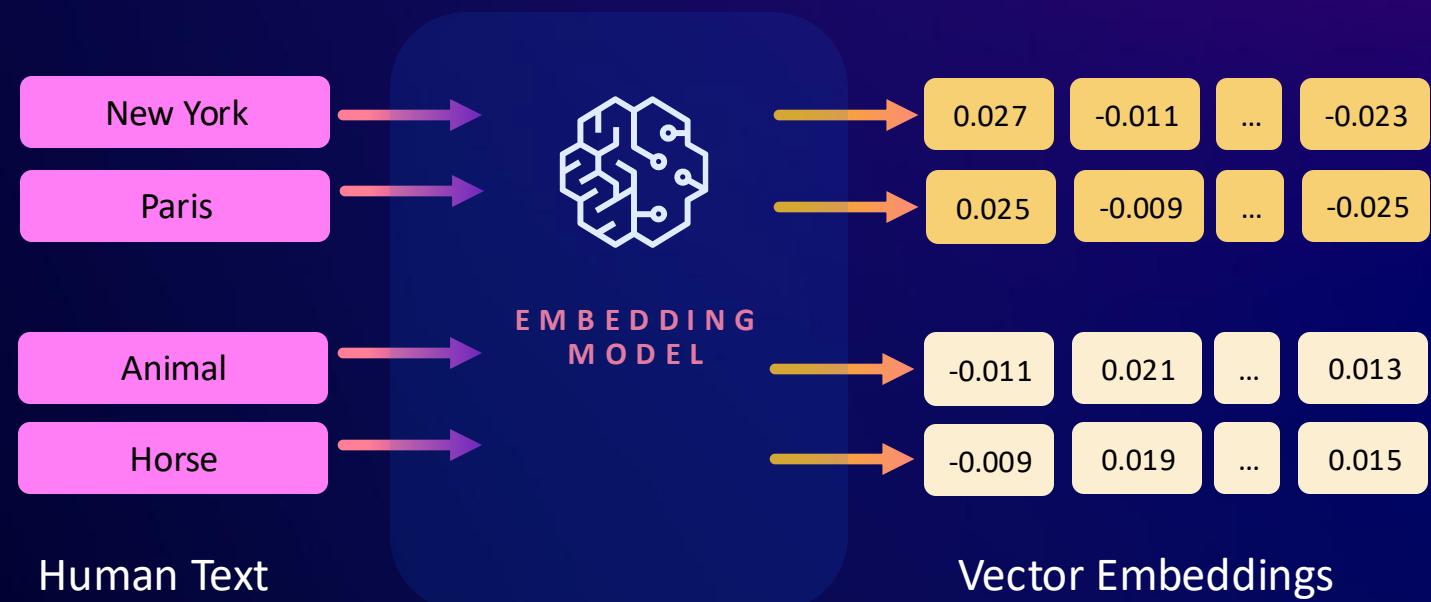


Generation

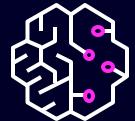
Response from the foundation model based on the augmented prompt.

What are embeddings?

- Numerical representation of text (vectors) that captures semantics and relationships between words.
- Embedding models capture features and nuances of the text.
- Rich embeddings can be used to compare text similarity.
- Multilingual Text Embeddings can identify meaning in different languages.



Titan text embeddings model



Amazon Titan Text Embeddings

v2.0

Translates text inputs (words, phrases) into numerical representations (embeddings). Comparing embeddings produces more relevant and contextual responses than word matching.

Max Tokens: **8,000**

Output Vectors: **1,536**

Language: **Multilingual** (25 languages)

Model ID: *amazon.titan-embed-g1-text-02*



Highlights

- Titan Text Embeddings offers fast, cost effective, high-performance, accurate embeddings in 25 languages.
- Optimized for text retrieval tasks, semantic similarity and clustering.
- Applications of this model includes semantic search and personalization.

Data Ingestion Workflow

KNOWLEDGE BASES FOR AMAZON BEDROCK

Fully managed
data ingestion
workflow



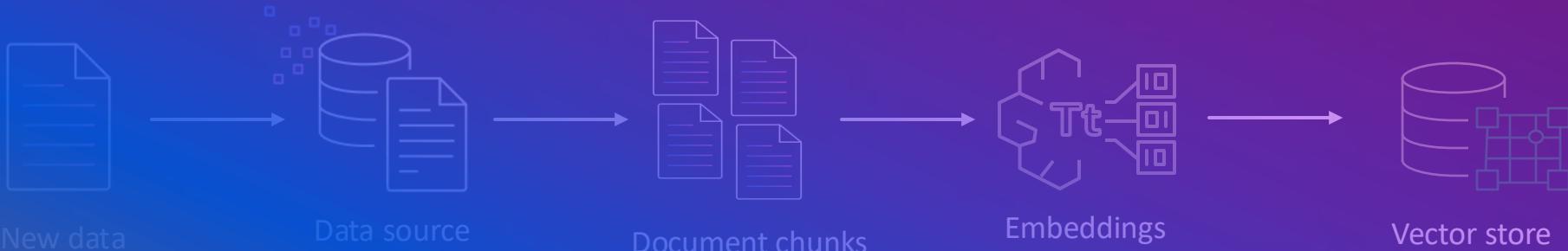
- Choose your data source (Amazon S3)
- Support for incremental updates
- Multiple data file formats supported
- Choose your chunking strategy
 - Fixed chunks
 - No chunking
 - Default (200 tokens)
- Choose your embedding model
 - Amazon Titan
- Choose your vector store
 - Open search serverless
 - Pinecone
 - Redis



Fully managed data ingestion

KNOWLEDGE BASES FOR AMAZON BEDROCK

Fully managed
data ingestion
workflow



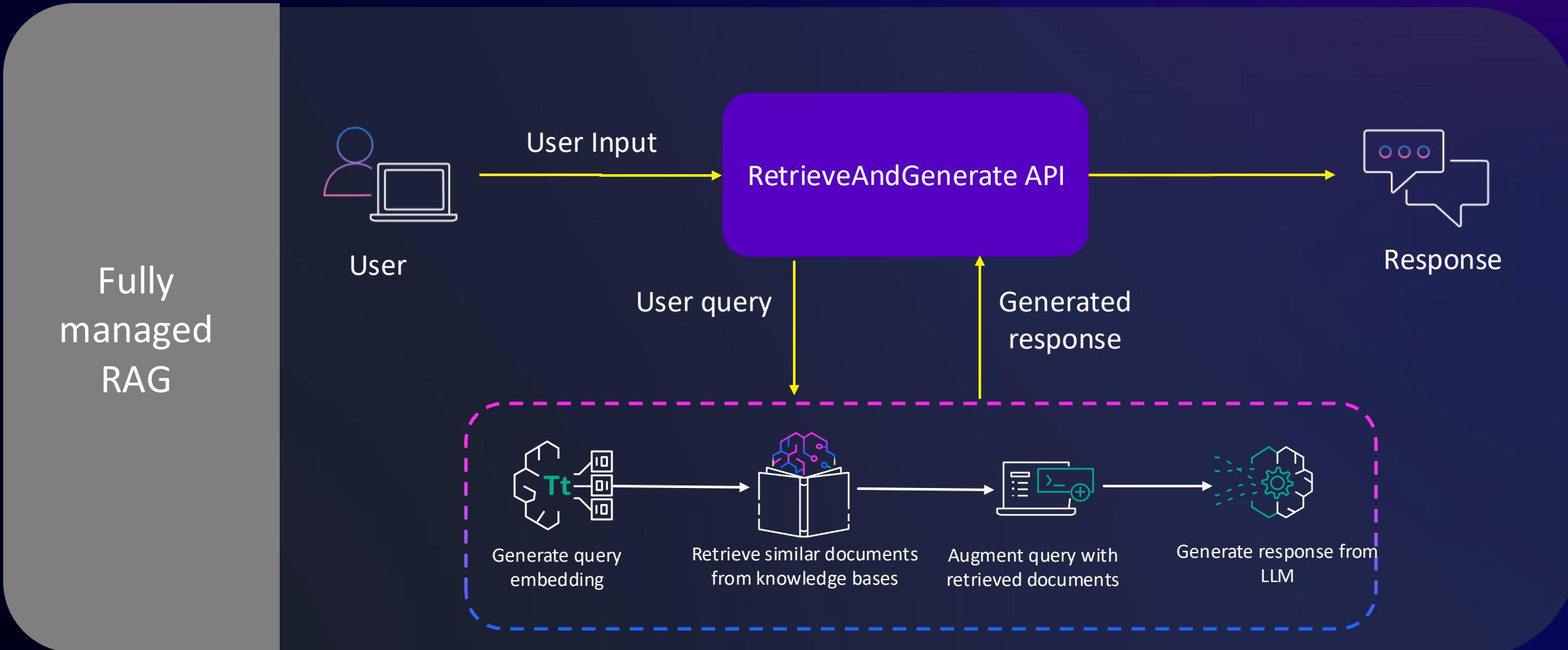
Automated and fully managed data ingestion using
Knowledge Bases for Amazon Bedrock

- Support for incremental updates
- Multiple data file formats supported
- Choose your data source (Amazon S3, AWS Lambda, AWS Step Functions, AWS Lambda@Edge)
- Choose your chunking strategy (Fixed chunks, No chunking, Default (200 tokens))
- Choose your embedding model (Amazon Titan)
- Choose your vector store (Open search serverless, Pinecone, Redis)



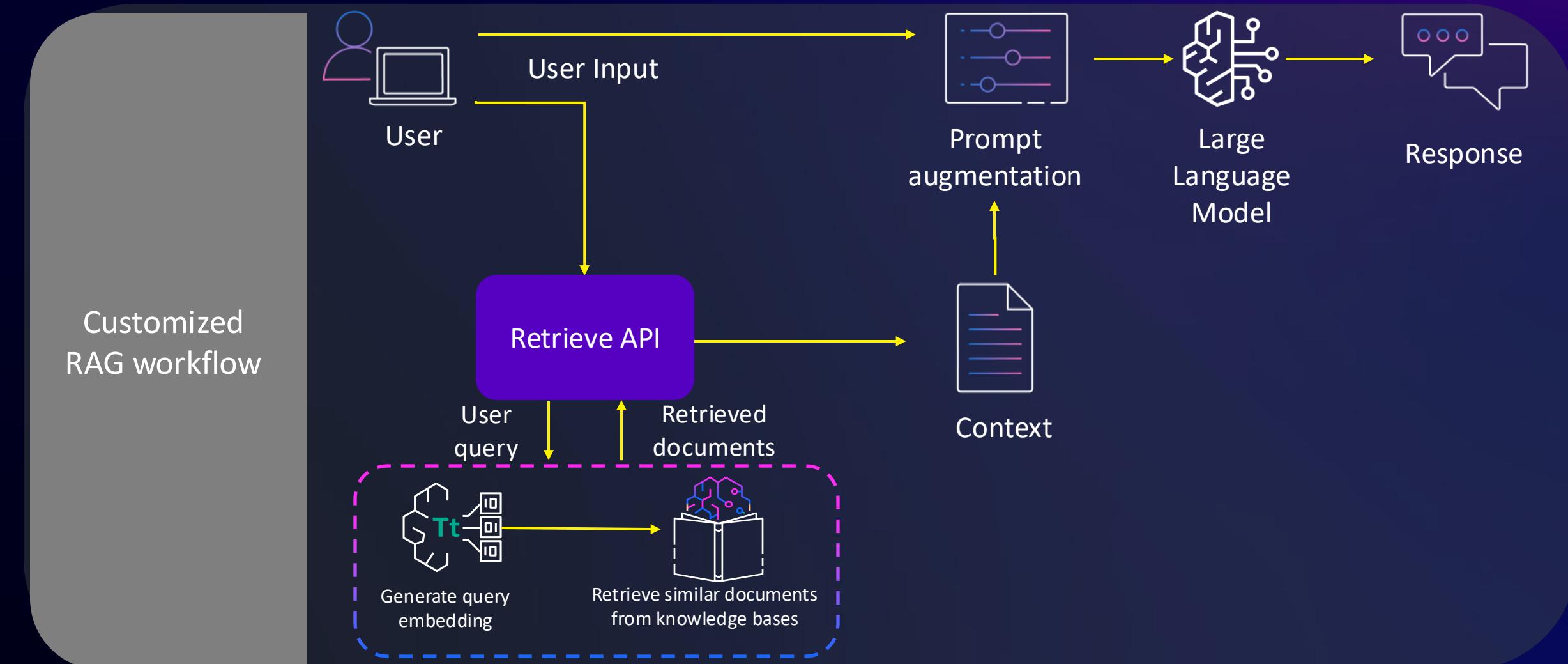
RetrieveAndGenerate API

KNOWLEDGE BASES FOR AMAZON BEDROCK



Customize RAG workflows using Retrieve API

KNOWLEDGE BASES FOR AMAZON BEDROCK



[尚未上傳至 Amazon Bedrock]

瑕疵偵測

影片精彩片段

文件解析

文生圖、圖生圖

Artifacts

比較模式

▼ 安全防護

圖片浮水印

模型評估

安全圍籬 Guardrail

▼ 建構工具

▼ 知識庫 (Knowledge Base)

知識庫 (Knowledge Base)

與文件聊天

▶ 智能代理

提示詞管理

提示詞工作流

▶ 定製模型

▶ 開發維運實踐

實驗總結

行業實驗資源

▼ Content preferences

Language

中文(繁體)

知識庫 (Knowledge Base)

使用 Knowledge Bases for Amazon Bedrock 快速構建知識庫

一、上傳文件到 S3 bucket

1. 下載要查詢的文件 [\[\]](#)，這個連結是 Amazon Bedrock 的英文操作手冊，也可以使用您自己的文件。

支援的文件格式如下：

- Plain text (.txt)
- Markdown (.md)
- HyperText Markup Language (.html)
- Microsoft Word document (.doc/.docx)
- Comma-separated values (.csv)
- Microsoft Excel spreadsheet (.xls/.xlsx)
- Portable Document (.pdf)

注意：單個文件不可以大於 50MB

2. 通過 <https://console.aws.amazon.com/s3/> [\[\]](#) 打開 Amazon S3 控制台，然後點擊 建立儲存貯體(Create bucket) 按鈕。
3. 紿定儲存貯體名稱，例如：`kbbucketxxxx`，`xxxx` 請改為隨機的4位數字，點擊最下面的 建立儲存貯體(Create bucket) 按鈕。如果已存在，請修改名稱，繼續建立
4. 點擊建立好的儲存貯體，點擊建立資料夾，建立一個資料夾，點擊上傳(Upload)按鈕，點擊新增檔案(Add File)按鈕，上傳前面步驟已下載的文件或您自己的文件，上傳成功後進入對應資料夾中，點擊 複製 S3 URI 按鈕。您也可以直接上傳資料夾(Add Folder)，成功後複製文件夾的 S3 URI。

二、知識庫

[未完成] 上傳至 Amazon S3

瑕疵偵測

影片精彩片段

文件解析

文生圖、圖生圖

Artifacts

比較模式

▼ 安全防護

圖片浮水印

模型評估

安全圍籬 Guardrail

▼ 建構工具

▼ 知識庫 (Knowledge Base)

知識庫 (Knowledge Base)

與文件聊天

▶ 智能代理

提示詞管理

提示詞工作流

▶ 定製模型

▶ 開發維運實踐

實驗總結

行業實驗資源

▼ Content preferences

Language

中文(繁體) ▾

知識庫 (Knowledge Base)

使用 Knowledge Bases for Amazon Bedrock 快速構建知識庫

一、上傳文件到 S3 bucket

1. 下載要查詢的文件 [\[\]](#)，這個連結是 Amazon Bedrock 的英文操作手冊，也可以使用您自己的文件。

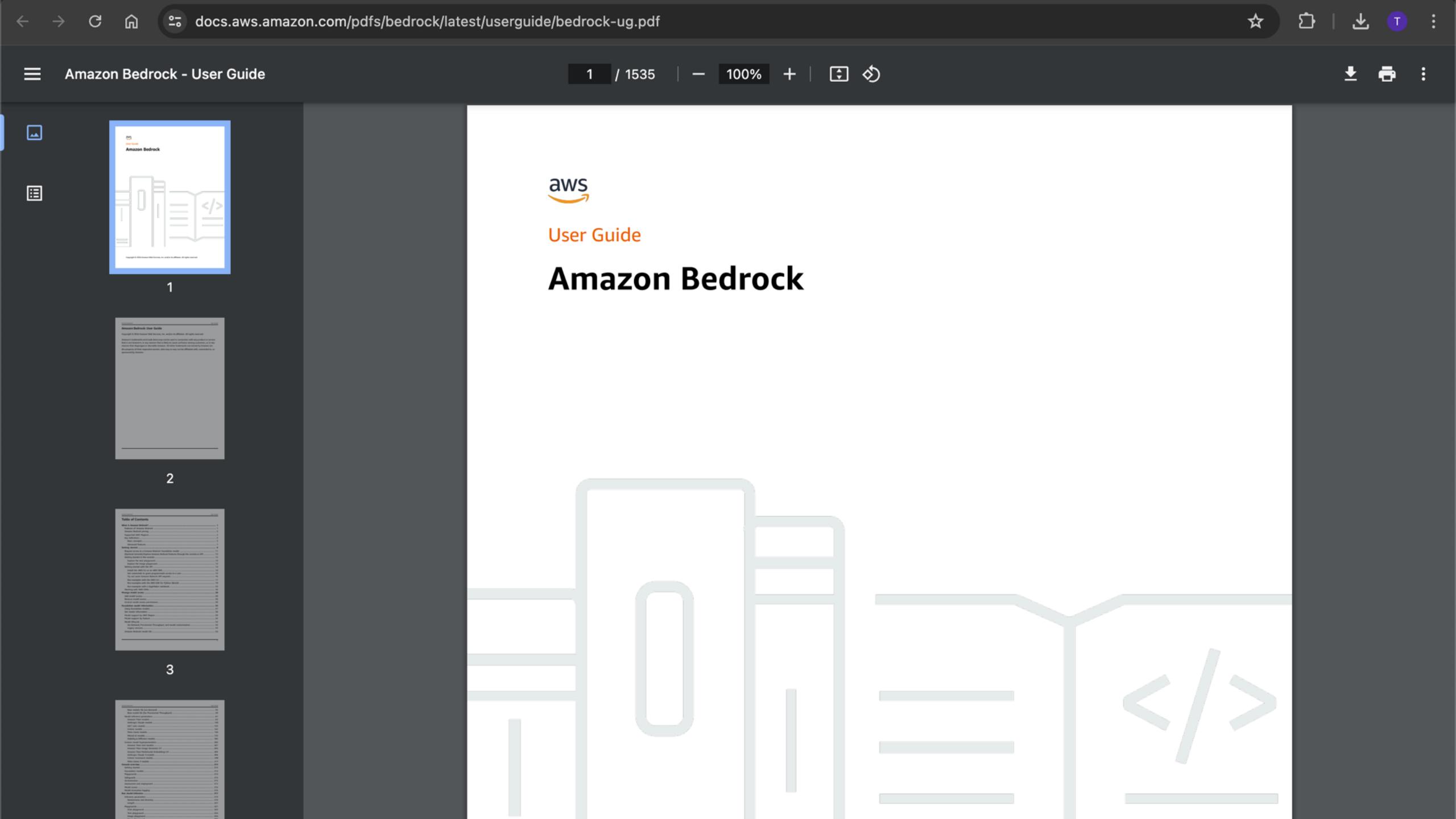
支援的文件格式如下：

- Plain text (.txt)
- Markdown (.md)
- HyperText Markup Language (.html)
- Microsoft Word document (.doc/.docx)
- Comma-separated values (.csv)
- Microsoft Excel spreadsheet (.xls/.xlsx)
- Portable Document (.pdf)

注意：單個文件不可以大於 50MB

2. 通過 <https://console.aws.amazon.com/s3/> [\[\]](#) 打開 Amazon S3 控制台，然後點擊 建立儲存貯體(Create bucket) 按鈕。
3. 紿定儲存貯體名稱，例如：`kbbucketxxxx`，`xxxx` 請改為隨機的4位數字，點擊最下面的 建立儲存貯體(Create bucket) 按鈕。如果已存在，請修改名稱，繼續建立
4. 點擊建立好的儲存貯體，點擊建立資料夾，建立一個資料夾，點擊上傳(Upload)按鈕，點擊新增檔案(Add File)按鈕，上傳前面步驟已下載的文件或您自己的文件，上傳成功後進入對應資料夾中，點擊 複製 S3 URI 按鈕。您也可以直接上傳資料夾(Add Folder)，成功後複製文件夾的 S3 URI。

二、知識庫



1



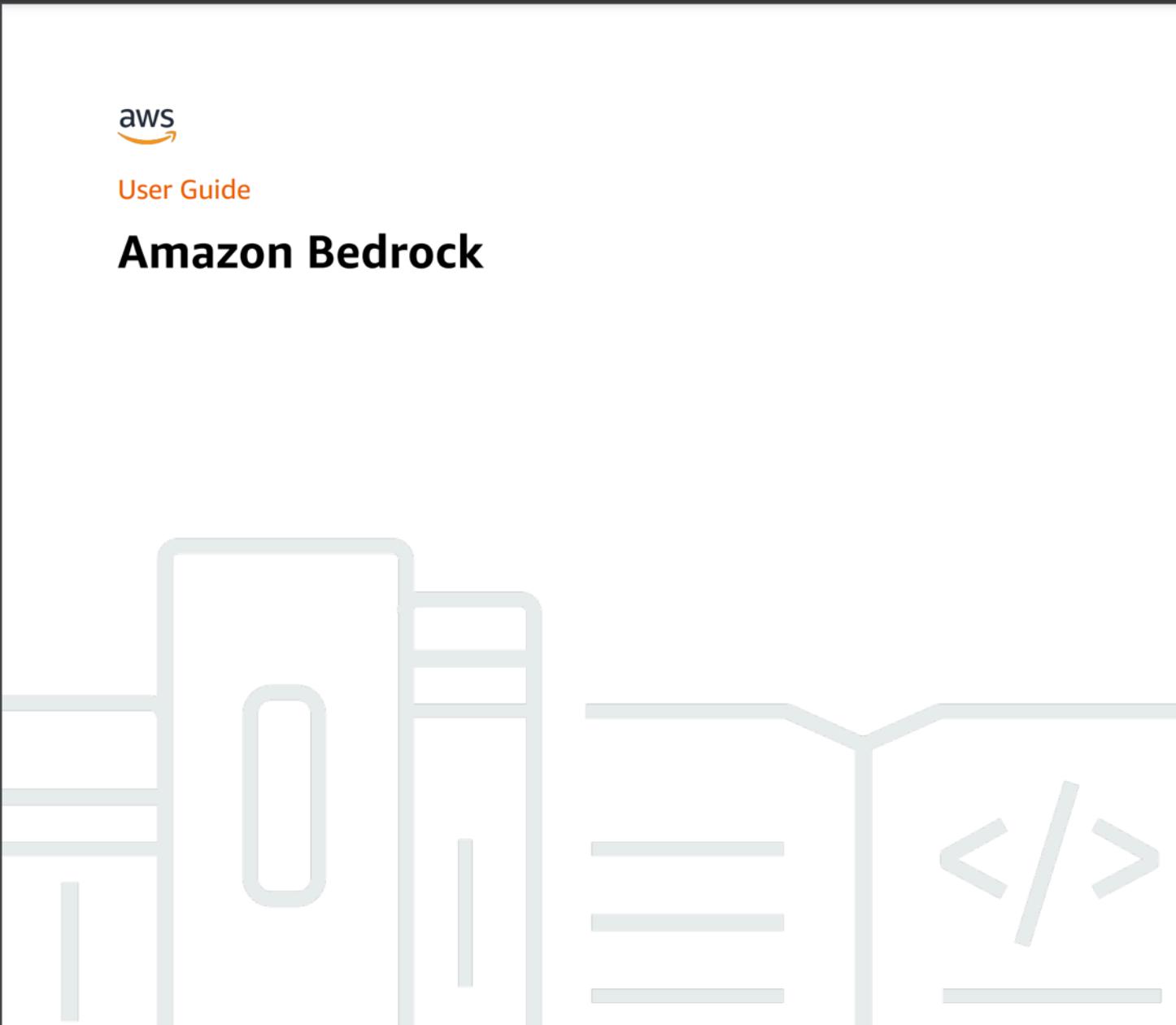
2



3



4



AWS Services Search: s3

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

Metrics

Search results for 's3'

Services (8)

Features (39)

Resources New

Documentation (26,830)

Knowledge Articles (288)

Marketplace (1,851)

Blogs (1,415)

Events (26)

Tutorials (12)

Services

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

See all 8 results ▶

Features

Imports from S3

DynamoDB feature

Feature spotlight

S3 feature

S3 Access Grants

S3 feature

Claude's Sonnet

See all 39 results ▶

examples

Compare mode

Configurations

System prompts

Add system prompts

Randomness and diversity

Temperature: 1

Top P: 0.999

Top K: 250

Define metric criteria

CloudShell Feedback

© 2024, Amazon Web Services, Inc. or its affiliates.

Privacy Terms Cookie preferences

Storage

Amazon S3

Store and retrieve any amount of data from anywhere

Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance.

How it works



Introduction to Amazon S3  Copy link

aws 

aws.amazon.com/s3

Create a bucket

Every object in S3 is stored in a bucket. To upload files and folders to S3, you'll need to create a bucket where the objects will be stored.

[Create bucket](#) 

Pricing

With S3, there are no minimum fees. You only pay for what you use. Prices are based on the location of your S3 bucket.

Estimate your monthly bill using the [AWS Simple Monthly Calculator](#) 

[View pricing details](#) 

Resources

us-west-2.console.aws.amazon.com/s3/bucket/create?region=us-west-2

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

Amazon S3 > Buckets > Create bucket

Create bucket Info

Buckets are containers for data stored in S3.

General configuration

AWS Region
US West (Oregon) us-west-2

Bucket type Info

General purpose
Recommended for most use cases and access patterns.
General purpose buckets are the original S3 bucket type.
They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

Directory - New
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.

Bucket name Info

Bucket name must be unique within the global namespace and follow the bucket naming rules. [See rules for bucket naming](#)

Copy settings from existing bucket - *optional*
Only the bucket settings in the following configuration are copied.

[Choose bucket](#)

Format: s3://bucket/prefix

us-west-2.console.aws.amazon.com/s3/bucket/create?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 > Create bucket

Create bucket Info

Buckets are containers for data stored in S3.

General configuration

AWS Region
US West (Oregon) us-west-2

Bucket type Info

General purpose
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

Directory - New
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.

Bucket name Info
kbbucket-20240727

Bucket name must be unique within the global namespace and follow the bucket naming rules. [See rules for bucket naming](#)

Copy settings from existing bucket - *optional*
Only the bucket settings in the following configuration are copied.

[Choose bucket](#)

Format: s3://bucket/prefix

us-west-2.console.aws.amazon.com/s3/bucket/create?region=us-west-2&bucketType=general

aws Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Add tag

Default encryption Info

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption type Info

Server-side encryption with Amazon S3 managed keys (SSE-S3)

Server-side encryption with AWS Key Management Service keys (SSE-KMS)

Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)

Secure your objects with two separate layers of encryption. For details on pricing, see DSSE-KMS pricing on the Storage tab of the [Amazon S3 pricing page](#).

Bucket Key

Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. [Learn more](#)

Disable

Enable

▶ Advanced settings

After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.

Cancel Create bucket



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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

Successfully created bucket "kbbucket-20240727" View details X

To upload files and folders, or to configure additional bucket settings, choose View details.

Amazon S3 > Buckets

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

Buckets are containers for data stored in S3.

Find buckets by name < 1 > ⚙️

C Copy ARN Empty Delete Create bucket

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)

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

Services Search [Option+S] | Amazon S3 > Buckets > kbbucket-20240727

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

Amazon S3 > Buckets > kbbucket-20240727

kbbucket-20240727 Info

Objects Properties Permissions Metrics Management Access Points

Objects (0) Info C 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
No objects					
You don't have any objects in this bucket.					
Upload					

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)

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

Services Search [Option+S] Services Search [Option+S]

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

Cancel Open

Destination Info

Destination s3://kbbucket-20240727

Desktop — iCloud

Name	Date Modified	Size	Kind
bedrock-ug.pdf	Today at 3:36 PM	11.8 MB	PDF Document

Search

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 (1 Total, 11.2 MB)

All files and folders in this table will be uploaded.

Find by name < 1 >

<input type="checkbox"/>	Name	Folder	Type
<input type="checkbox"/>	bedrock-ug.pdf	-	application/pdf

Destination Info

Destination
<s3://kbbucket-20240727>

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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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

Files and folders (1 Total, 11.2 MB)

All files and folders in this table will be uploaded.

Find by name

<input type="checkbox"/>	Name	Folder	Type
<input type="checkbox"/>	bedrock-ug.pdf	-	application/pdf

Remove Add files Add folder

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



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

Upload succeeded View details below.

Upload: status

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

Summary

Destination	Succeeded	Failed
s3://kbbucket-20240727	1 file, 11.2 MB (100.00%)	0 files, 0 B (0%)

Files and folders Configuration

Files and folders (1 Total, 11.2 MB)

Find by name < 1 >

Name	Folder	Type	Size	Status	Error
bedrock-up	-	application/	11.2 MB	Succeeded	-

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

Services Search [Option+S] | Amazon S3 > Buckets > kbbucket-20240727

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

Amazon S3 > Buckets > kbbucket-20240727

kbbucket-20240727 Info

Objects Properties Permissions Metrics Management Access Points

Objects (1) Info C 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 > ⚙️

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	bedrock-ug.pdf	pdf	July 27, 2024, 15:39:10 (UTC+08:00)	11.2 MB	Standard

AWS Services Search: bedrock Oregon WSParticipantRole/Participant @ 1561-5387-8293

EC2 VPC

Amazon S3 kbbuck

Services (1)

Resources New

Documentation (2,676)

Knowledge Articles (12)

Marketplace (386)

Blogs (232)

Events (1)

Tutorials (1)

Find objects

Name

File

Search results for 'bedrock'

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

What is Amazon Bedrock Studio? ↗

Create folder Upload

explicitly grant them permissions. [Learn](#)

< 1 > 🔍

Storage class ▾

Standard

The screenshot shows the AWS Bedrock homepage within the AWS Management Console. The top navigation bar includes the AWS logo, a 'Services' dropdown, a search bar with the placeholder '[Option+S]', and a user profile section showing 'Oregon' and a participant role. Below the navigation is a secondary header with links for EC2, VPC, RDS, S3, Support, Amazon SageMaker, AWS DeepRacer, and CloudFormation. A 'Machine Learning' sidebar is visible on the left. The main content area features a large heading 'Amazon Bedrock' followed by a subtext: 'The easiest way to build and scale generative AI applications with foundation models (FMs)'. To the right is a call-to-action box with a 'Try Bedrock' heading and a 'Get started' button.

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

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

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

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](#)

us-west-2.console.aws.amazon.com/bedrock/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

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
- Provisioned Throughput

Model access [4 new](#)

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

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

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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 (0)

Edit Delete Test knowledge base Create knowledge base

Find knowledge base

Name	Status	Description	Source files	Creation time	Last sync w...	Last sync
No knowledge base						
No knowledge base to display						

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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 > Knowledge bases > Create knowledge base

Step 1 Provide knowledge base details

Step 2 Configure data source

Step 3 Select embeddings model and configure vector store

Step 4 Review and create

Provide knowledge base details

Knowledge base details

Knowledge base name

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 50 characters.

Knowledge base description - optional

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 200 characters.

IAM permissions

Certain permissions are necessary to access other services or perform actions in order to create this resource. For more information, see [service role](#) for Amazon Bedrock

Runtime role

Create and use a new service role

Use an existing service role

Service role name

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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](#)

Choose data source

Select the data source that you want to configure in the next step.

 **Amazon S3**

Object storage service that stores data as objects within buckets.

 **Web Crawler - Preview**

Web page crawler that extracts content from public web pages you are authorized to crawl.

Third party data sources

 **Confluence - Preview**

Collaborative work-management tool designed for project planning, software development and product management.

 **Salesforce - Preview**

Customer relationship management (CRM) tool for managing support, sales, and marketing data.

 **Sharepoint - Preview**

Collaborative web-based service for working on documents, web pages, web sites, lists, and more.

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.

No tags associated with the resource.

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

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

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
Provisioned Throughput

Model access [4 new](#)
Bedrock Studio [Preview](#)
Settings
User guide
Bedrock Service Terms

Object storage service that stores data as objects within buckets.

Web Crawler - [Preview](#)
Web page crawler that extracts content from public web pages you are authorized to crawl.

Third party data sources

Confluence - Preview
Collaborative work-management tool designed for project planning, software development and product management.

Salesforce - Preview
Customer relationship management (CRM) tool for managing support, sales, and marketing data.

Sharepoint - Preview
Collaborative web-based service for working on documents, web pages, web sites, lists, and more.

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.
No tags associated with the resource.

[Add new tag](#)
You can add up to 50 more tags.

[Cancel](#)

[Next](#)

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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 > Create knowledge base

Step 1: Provide knowledge base details

Step 2: Configure data source

Step 3: Select embeddings model and configure vector store

Step 4: Review and create

Configure data source

Configure for the chosen data source

Amazon S3 [Info](#)

Provide details to connect Amazon Bedrock to your S3 data source.

Data source: knowledge-base-quick-start-cdw4o-data-source

[Delete](#)

Data source name
knowledge-base-quick-start-cdw4o-data-source
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Data source location
 This AWS account
 Other AWS account

S3 URI
To increase the accuracy and relevance of your responses, add a .metadata.json file containing metadata for your data source to your S3 bucket. [Info](#)

Choose a s3 location [View](#) [Browse S3](#)

Add customer-managed KMS key for S3 data - *optional*
If you encrypted your S3 data, provide the KMS key here so that Bedrock can decrypt it.

Chunking and parsing configurations [Info](#)
Choose between default or advanced customization.

Default
 Custom

Amazon Bedrock <

Amazon Bedrock > Knowledge bases > Create knowledge base



- Step 1
Provide knowledge base details
- Step 2
Configure data source

Configure data source

Configure for the chosen data source



Info

Choose an archive in S3

S3 buckets

Buckets (1/1)

Find S3 bucket



< 1 >



Name

Creation date

 kbbucket-20240727

2024-07-27T07:38:20.000Z

Cancel

Choose



Choose a s3 location

View

Browse S3

 Add customer-managed KMS key for S3 data - optional

If you encrypted your S3 data, provide the KMS key here so that Bedrock can decrypt it.

Chunking and parsing configurations Info

Choose between default or advanced customization.

 Default

Uses default parsing and chunking strategy.

 Custom

Customize the parsing and chunking strategy, including using advanced parsing.

- Getting started
- Overview
 - Examples
 - Providers

Four

- Base
- Custom
- Import

Play

- Chat
- Text
- Image

Safe

- Guard

Watermark detection

Builder tools

Knowledge bases

Agents

Prompt management PreviewPrompt flows Preview

Assessment & deployment

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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 > Knowledge bases > Create knowledge base

Step 1: Provide knowledge base details

Step 2: **Configure data source**

Step 3: Select embeddings model and configure vector store

Step 4: Review and create

Configure data source

Configure for the chosen data source

Amazon S3 [Info](#)

Provide details to connect Amazon Bedrock to your S3 data source.

▼ Data source: knowledge-base-quick-start-cdw4o-data-source [Delete](#)

Data source name
knowledge-base-quick-start-cdw4o-data-source
Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Data source location
 This AWS account
 Other AWS account

S3 URI
To increase the accuracy and relevance of your responses, add a .metadata.json file containing metadata for your data source to your S3 bucket. [Info](#)

s3://kbbucket-20240727 [X](#) [View](#) [Browse S3](#)

Add customer-managed KMS key for S3 data - *optional*
If you encrypted your S3 data, provide the KMS key here so that Bedrock can decrypt it.

Chunking and parsing configurations [Info](#)
Choose between default or advanced customization.

Default
 Custom

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

aws Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

Amazon Bedrock <

configure vector store

Step 4 Review and create

▼ 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

▼ Data source: knowledge-base-quick-start-cdw4o-data-source

Delete

Data source name

knowledge-base-quick-start-cdw4o-data-source

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Data source location

This AWS account

Other AWS account

S3 URI

To increase the accuracy and relevance of your responses, add a .metadata.json file containing metadata for your data source to your S3 bucket. [Info](#)

s3://kbbucket-20240727

[View](#) [Browse S3](#)

Add customer-managed KMS key for S3 data - *optional*

If you encrypted your S3 data, provide the KMS key here so that Bedrock can decrypt it.

Chunking and parsing configurations [Info](#)

Choose between default or advanced customization.

Default

Uses default parsing and chunking strategy.

Custom

Customize the parsing and chunking strategy, including using advanced parsing.

► Advanced settings - *optional*

Add data source

You can add 4 more data source(s).

Cancel Previous Next

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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 > Create knowledge base

Step 1: Provide knowledge base details

Step 2: Configure data source

Step 3: Select embeddings model and configure vector store

Step 4: Review and create

Select embeddings model and configure vector store

Choose an embeddings model to convert the data that you will provide in the next step, and provide details for a vector data store in which Bedrock can store, manage, and update your embeddings. The embeddings model and vector store cannot be changed after creation of knowledge base.

Embeddings model

Select an embeddings model to convert your data into an embedding. Pricing depends on the model. [Learn more](#)

-  [Titan Text Embeddings v2](#) 
By Amazon
-  [Titan Embeddings G1 - Text v1.2](#) 
By Amazon
-  [Embed English v3](#) 
By Cohere
-  [Embed Multilingual v3](#) 
By Cohere

Vector dimensions

Select the vector dimension size for your embeddings model to balance accuracy, cost, and latency. Higher dimensions improves overall accuracy and requires more vector storage. [Learn more](#)

1024

Vector database

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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 > Create knowledge base

Step 1: Provide knowledge base details

Step 2: Configure data source

Step 3: Select embeddings model and configure vector store

Step 4: Review and create

Select embeddings model and configure vector store

Choose an embeddings model to convert the data that you will provide in the next step, and provide details for a vector data store in which Bedrock can store, manage, and update your embeddings. The embeddings model and vector store cannot be changed after creation of knowledge base.

Embeddings model

Select an embeddings model to convert your data into an embedding. Pricing depends on the model. [Learn more](#)

Titan Text Embeddings v2 By Amazon 

Titan Embeddings G1 - Text v1.2 By Amazon

Embed English v3 By Cohere

Embed Multilingual v3 By Cohere

Vector dimensions

Select the vector dimension size for your embeddings model to balance accuracy, cost, and latency. Higher dimensions improves overall accuracy and requires more vector storage. [Learn more](#)

1024

Vector database

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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](#)
- Prompt flows [Preview](#)

Embed English v3 By Cohere

Embed Multilingual v3 By Cohere

Vector dimensions

Select the vector dimension size for your embeddings model to balance accuracy, cost, and latency. Higher dimensions improves overall accuracy and requires more vector storage. [Learn more](#)

1024

Vector database

Let Amazon create a vector store on your behalf or select a previously created store to allow Bedrock to store, update and manage embeddings. You will be billed directly from the vector store provider. [Learn more](#)

Select how you want to create your vector store.

Quick create a new vector store - *Recommended*
We will create an Amazon OpenSearch Serverless vector store on your behalf. This cost-efficient option is intended only for development and can't be migrated to production workload later. [Learn more](#)

Choose a vector store you have created
Select Amazon OpenSearch Serverless, Amazon Aurora, MongoDB Atlas, Pinecone or Redis Enterprise Cloud and provide field mappings.

Enable redundancy (active replicas) - *optional*
The default configuration has active replicas disabled, which is optimal for development workloads. Enable this option if you want to enable redundant active replicas, which may increase storage costs.

Add customer-managed KMS key for Amazon OpenSearch Serverless vector - *optional*
If you encrypted your OpenSearch data, provide the KMS key here so that Bedrock can decrypt it.

Cancel Previous Next

←

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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 > Knowledge bases > Create knowledge base

Step 1

- Provide knowledge base details
- Configure data source
- Select embeddings model and configure vector store

Step 4

[Review and create](#)

Review and create

Step 1: Provide details

[Edit](#)

Knowledge base details		
Knowledge base name	Knowledge base description	Service role
knowledge-base-quick-start-6zlo3	—	AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3

Tags (0)

Key	Value
No tags to display	

Step 2: Setup up data source

[Edit](#)

Data source: knowledge-base-quick-start-cdw4o-data-source

Data source name	Account ID	S3 URI

Red arrow pointing down to the "Edit" button in the Step 2 section.

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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](#)
- Prompt flows [Preview](#)

Assessment & deployment

Customer-managed KMS Key for S3

-

KMS key for transient data storage

-

Chunking strategy

Default

Parsing strategy

Default

Lambda function

-

S3 bucket for Lambda function

-

Data deletion policy

Delete

Step 3: Select embeddings model and configure vector store

[Edit](#)

Embeddings model

Model
Titan Text Embeddings v2

Vector dimensions
1024

Vector store

Quick create vector store - Recommended

We will create an Amazon OpenSearch Serverless vector store in your account on your behalf.

[Cancel](#) [Previous](#) [Create knowledge base](#)

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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](#)

Preparing vector database in Amazon OpenSearch Serverless. This process may take several minutes to complete.

Amazon Bedrock > Knowledge bases > Create knowledge base

Step 1

- Provide knowledge base details
- Configure data source
- Select embeddings model and configure vector store
- Review and create

Review and create

Step 1: Provide details

[Edit](#)

Knowledge base details

Knowledge base name	Knowledge base description	Service role
knowledge-base-quick-start-6zlo3	—	AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3

Tags (0)

Key	Value
No tags to display	

Step 2: Setup up data source

[Edit](#)

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/create-knowledge-base

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 Opensearch Serverless vector database is ready.

Notifications 0 0 1 0 2

Amazon Bedrock > Knowledge bases > Create knowledge base

Step 1 Provide knowledge base details

Step 2 Configure data source

Step 3 Select embeddings model and configure vector store

Step 4 Review and create

Review and create

Step 1: Provide details

Knowledge base details

Knowledge base name knowledge-base-quick-start-6zlo3 **Knowledge base description** — **Service role** AmazonBedrockExecutionRoleForKnowledgeBase_6zlo3

Tags (0)

Key	Value
No tags to display	

Step 2: Setup up data source

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

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](#)

Knowledge base 'knowledge-base-quick-start-6zlo3' is created successfully. Sync one or more data sources to index your content for searching. Syncing can take from a few minutes to a few hours. [Go to data sources](#)

Amazon Bedrock > Knowledge bases > knowledge-base-quick-start-6zlo3

knowledge-base-quick-start-6zlo3

Test Delete Edit

Knowledge base overview

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

Test knowledge base

Generate responses [Edit](#)

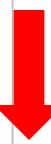
Select model

One or more data sources have not been synced. [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](#)



us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

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)

No tags

No tags to display

[Manage tags](#)

Add Edit Delete Sync

Data sources contain information returned when querying a Knowledge base.

Find data source

knowledgebase Available S3 15615387... [s3://k](#)

Embeddings model

Model	Vector dimensions
Titan Text Embeddings v2	1024

Test knowledge base

Generate responses [Edit](#) [Delete](#)

Select model

One or more data sources have not been synced. [Go to data sources](#)

Configure your retrieval and responses [X](#)
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](#)



us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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

No tags

No tags to display

Manage tags

Data sources contain information returned when querying a Knowledge base.

Find data source

knowledge... Syncing S3 15615387... s3://k

Embeddings model

Model	Vector dimensions
Titan Text Embeddings v2	1024

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



us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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

No tags

No tags to display

Manage tags

Find data source

Data so...	Status	Data sour...	Account ID	Source
knowledge...	Available	15615387...	s3://k...	

Embeddings model

Model	Vector dimensions
Titan Text Embeddings v2	1024

Test knowledge base

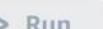
Generate responses

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 

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

Services Search [Option+S]

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Oregon WSParticipantRole/Participant @ 1561-5387-8293

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

Data so...	Status	Data sour...	Account ID	Source
knowledge...	Available	S3	15615387...	s3://k...

Embeddings model

Model	Vector dimensions
Titan Text Embeddings v2	1024

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 

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

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

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Select model

1. Category

Model providers

AI 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

3. Throughput

Provisioned throughput is not supported for knowledge bases.

Cancel Apply

Get 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

retrieval and search strategy base, select the on .

base by running a query to disable response generation information stored from your Generate responses above.

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

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

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

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

3. Throughput

On-demand (ODT)

Cancel Apply

Not seeing a model you are interested in? Check out all supported models [here](#)

base by running a query to disable response generation information stored from your Generate responses above.

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0

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

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

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

3. Throughput

On-demand (ODT)

Not seeing a model you are interested in? Check out all supported models [here](#)

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](#)

retrieval and search strategy base, select the on .

base by running a query to disable response generation information stored from your Generate responses above.

瑕疵偵測

影片精彩片段

文件解析

文生圖、圖生圖

Artifacts

比較模式

▼ 安全防護

圖片浮水印

模型評估

安全圍籬 Guardrail

▼ 建構工具

▼ 知識庫 (Knowledge Base)

[知識庫 \(Knowledge Base\)](#)

與文件聊天

▶ 智能代理

提示詞管理

提示詞工作流

▶ 定製模型

▶ 開發維運實踐

實驗總結

行業實驗資源

▼ Content preferences

Language

中文(繁體) ▾

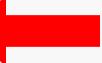
Cancel

Apply

Copied!



然後輸入要查詢的問題,等待輸出。

1 如何配置 Amazon Bedrock 知識庫? 

輸出如下圖:

 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 .

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](#) Prompt flows [Preview](#)

Assessment & deployment

knowledge-base-quick-start-6zlo3

Test Delete Edit

Knowledge base overview

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.

如何配置 Amazon Bedrock 知識庫? 

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

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](#) Prompt flows [Preview](#)

Assessment & deployment

knowledge-base-quick-start-6zlo3

Test Delete Edit

Knowledge base overview

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 [Edit](#) [Delete](#)

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.

如何配置 Amazon Bedrock 知識庫? [Run](#)

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/knowledge-bases/knowledge-base-quick-start-6zlo3/GRNV74BLMR/0?modelId=a... ☆

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

EC2 VPC RDS S3 Support Amazon SageMaker AWS DeepRacer CloudFormation

Amazon Bedrock

Knowledge base overview

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](#)

configurations icon .

如何配置 Amazon Bedrock 知識庫?

要配置 Amazon Bedrock 知識庫,您可以按照以下步驟: 1. 收集源文件並將它們存儲在支持的數據源中,例如 Amazon S3。 2. 在 Amazon Bedrock 控制台中創建一個新的知識庫,提供知識庫名稱、描述和 IAM 角色。 3. 選擇您的數據源並配置連接信息。 4. 選擇一個向量存儲來存儲數據的嵌入向量表示。您可以快速創建一個新的 Amazon OpenSearch Serverless 向量數據庫,或使用現有的向量存儲。 5. 選擇一個模型將您的數據轉換為嵌入向量。 6. 檢查知識庫配置並創建知識庫。 創建完成後,知識庫的狀態將變為"就緒"。[\[1\]](#)[\[2\]](#)[\[3\]](#)[\[4\]](#)[\[5\]](#) 一旦知識庫就緒,您就可以同步數據源以保持內容最新。在控制台中選擇您的知識庫,然後在數據源概覽部分選擇"同步"。[\[6\]](#)

[Show source details >](#)

Enter your message here ▶ Run

Amazon Bedrock

Test Knowledge base



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

[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

如何配置 Amazon Bedrock 知識庫?

要配置 Amazon Bedrock 知識庫,您可以按照以下步驟: 1. 收集源文件並將它們存儲在支持的數據源中,例如 Amazon S3。2. 在 Amazon Bedrock 控制台中創建一個新的知識庫,提供知識庫名稱、描述和 IAM 角色。3. 選擇您的數據源並配置連接信息。4. 選擇一個向量存儲來存儲數據的嵌入向量表示。您可以快速創建一個新的 Amazon OpenSearch Serverless 向量數據庫,或使用現有的向量存儲。5. 選擇一個模型將您的數據轉換為嵌入向量。6. 檢查知識庫配置並創建知識庫。創建完成後,知識庫的狀態將變為"就緒"。[\[1\]](#)[\[2\]](#)[\[3\]](#)[\[4\]](#)[\[5\]](#) 一旦知識庫就緒,您就可以同步數據源以保持內容最新。在控制台中選擇您的知識庫,然後在數據源概覽部分選擇"同步"。[\[6\]](#)

Source details (6)

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

Query configurations (1)**Source chunk 1**

Gather source documents to add to your knowledge base. 2. Store your source documents in a supported data source and configure the connection information to connect to and crawl your data. 3. (Optional if using Amazon S3 to store your source documents) Create a metadata file for each source document to allow for filtering of results during knowledge base query. 4. (Optional) Set up a vector index in a supported vector store to index your data. You can use the Amazon Bedrock console to create an Amazon OpenSearch Serverless vector database for you. 5. Create and configure your knowledge base. 6. Ingest your data by letting knowledge bases generate embeddings with an embeddings model and storing them in a supported vector store. 7. Set up your application or agent to query the knowledge base and return augmented responses. 492

<https://docs.aws.amazon.com/bedrock/latest/userguide/kb-how-it-works.html>

https://docs.aws.amazon.com/bedrock/latest/APIReference/API_agent-runtime_RetrieveAndGenerate.html

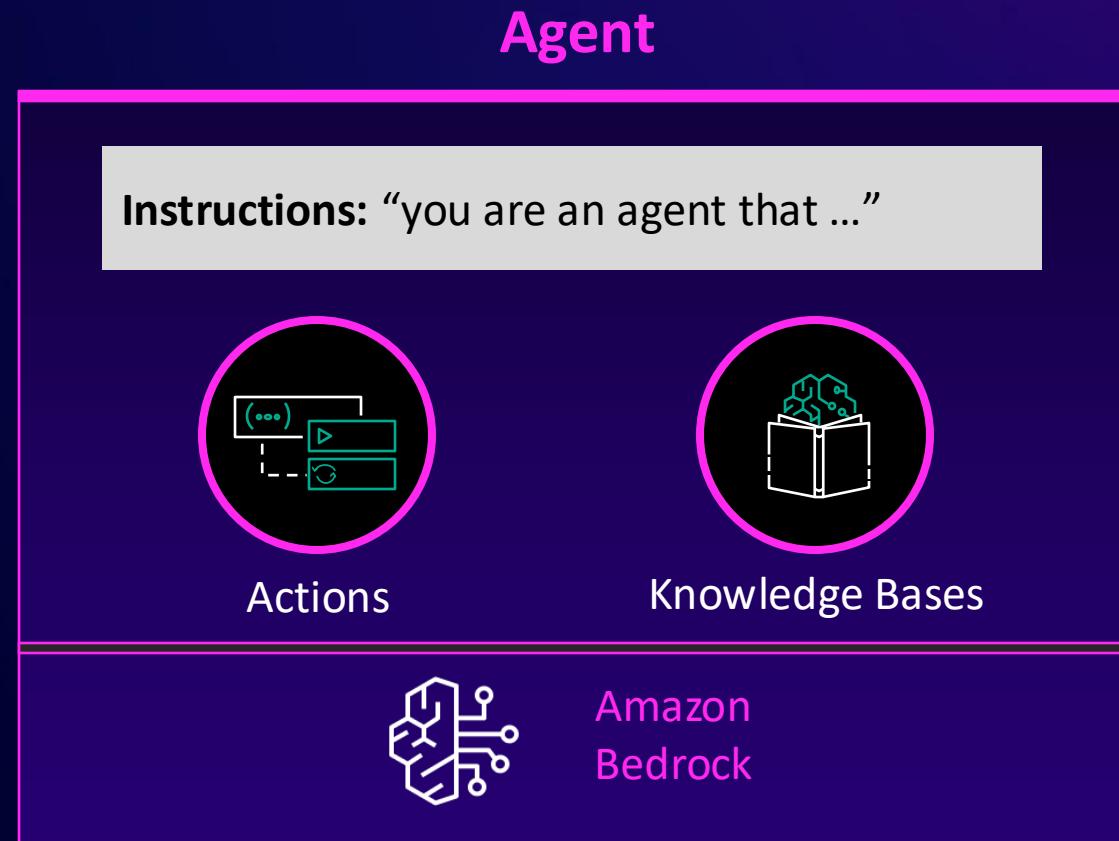
https://docs.aws.amazon.com/bedrock/latest/APIReference/API_agent-runtime_Retrieve.html

<https://docs.aws.amazon.com/bedrock/latest/userguide/data-source->

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- **Validating Bedrock agent**
- Exploring Bedrock fine-tune
- Q & A

Agent basics



Agents can combine Actions and Knowledge Bases



HR Policy Assistant v2

Instructions: “you are an HR agent, helping employees understand HR policies and manage vacation time”

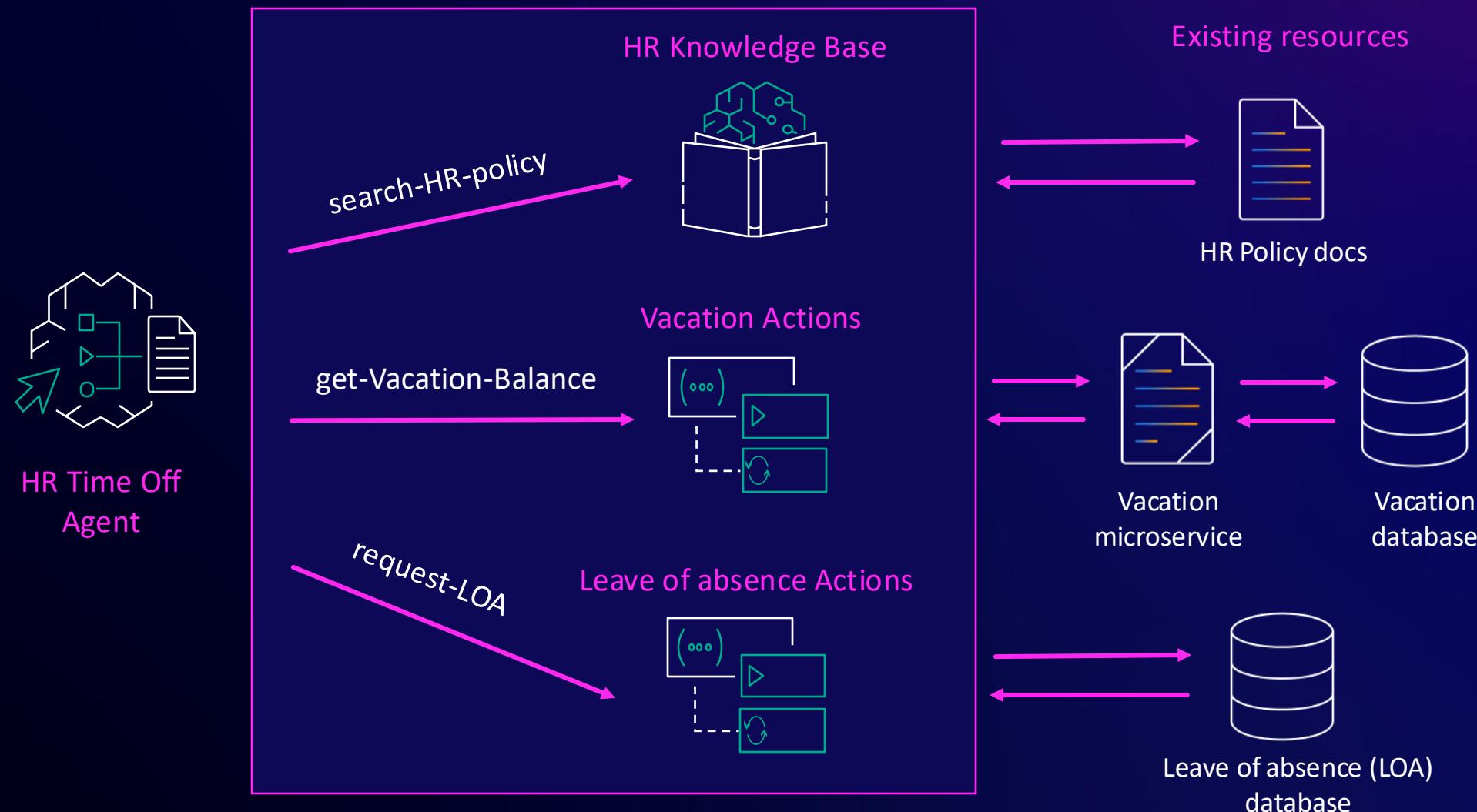
HR Knowledge Base

Vacation Policy
Contains the entire vacation policy for the company

HR Actions

Request Vacation
In: start date, end date
Out: approval status, remaining balance

Agents build on existing enterprise resources

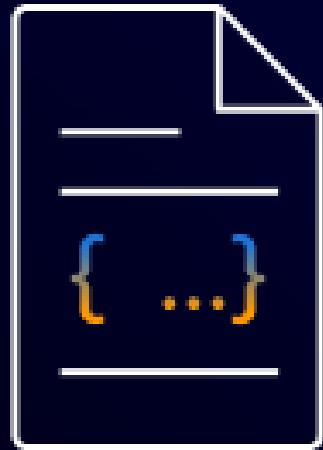


Action Group example

UtilityActionGroup

Description

API Schema



/action1
/action2

Lambda Function



```
def do_action1()  
def do_action2()
```

Action Group example

UtilityActionGroup

Description: “This action group provides a set of commonly used actions. Use these actions for things like sending emails and getting team member lists.”

API Schema

```
{ "openapi": "3.0.0",
  "info": { "title": "Utility Actions",
            "description": "... },
  "paths": {
    "/sendEmail": {
      "post": {
        "description": "This operation ...",
        "operationId": "sendEmail",
        "requestBody": { ... },
        "responses": {
          "200": {
            "text/plain": { ... }
          }
        }
      }
    }
  }
}
```

Lambda Function

```
def lambda_handler(event, context):
    if event['apiPath'] == '/sendEmail':
        result = sendEmail(event)
    elif event['apiPath'] == '/getTeam':
        result = getTeam(event)

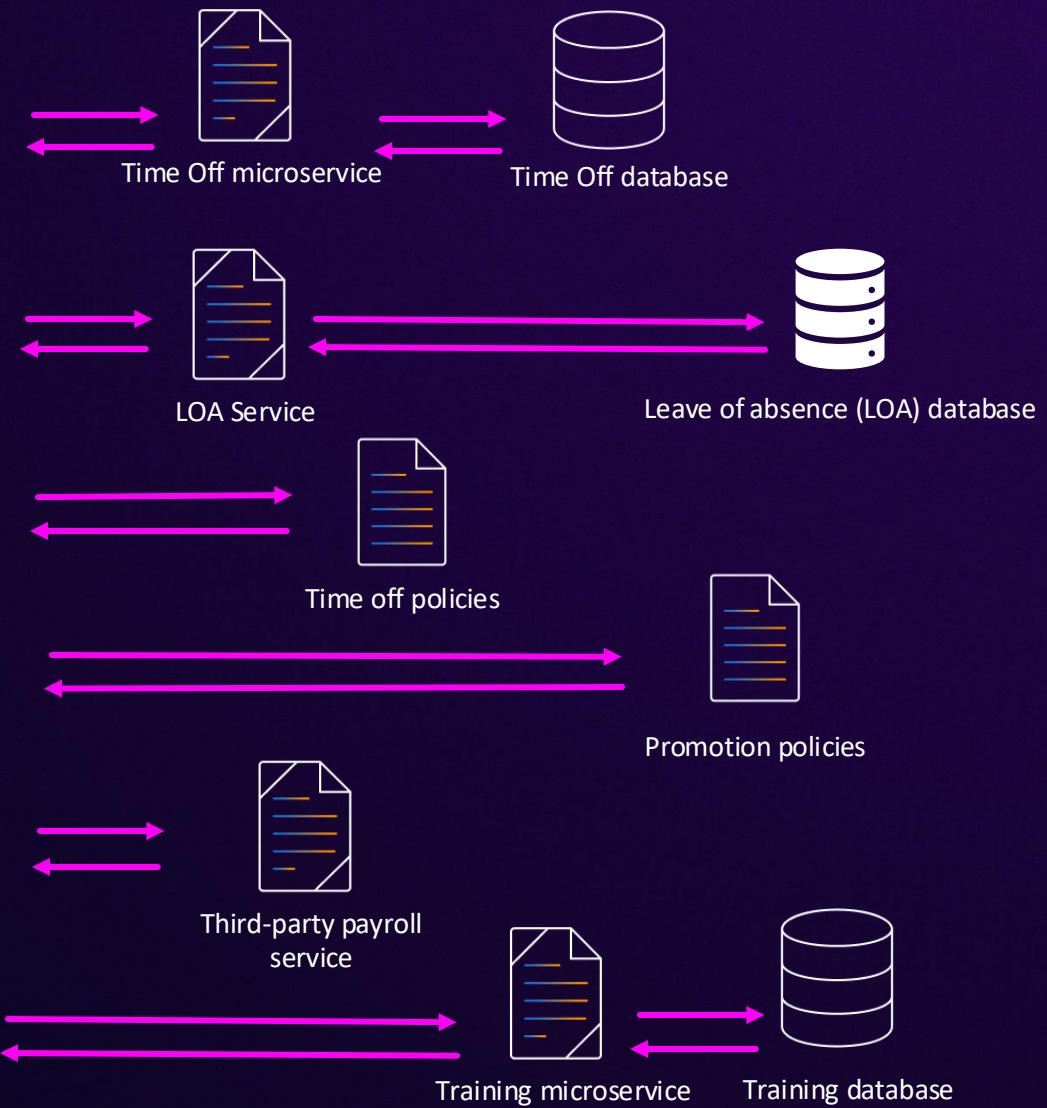
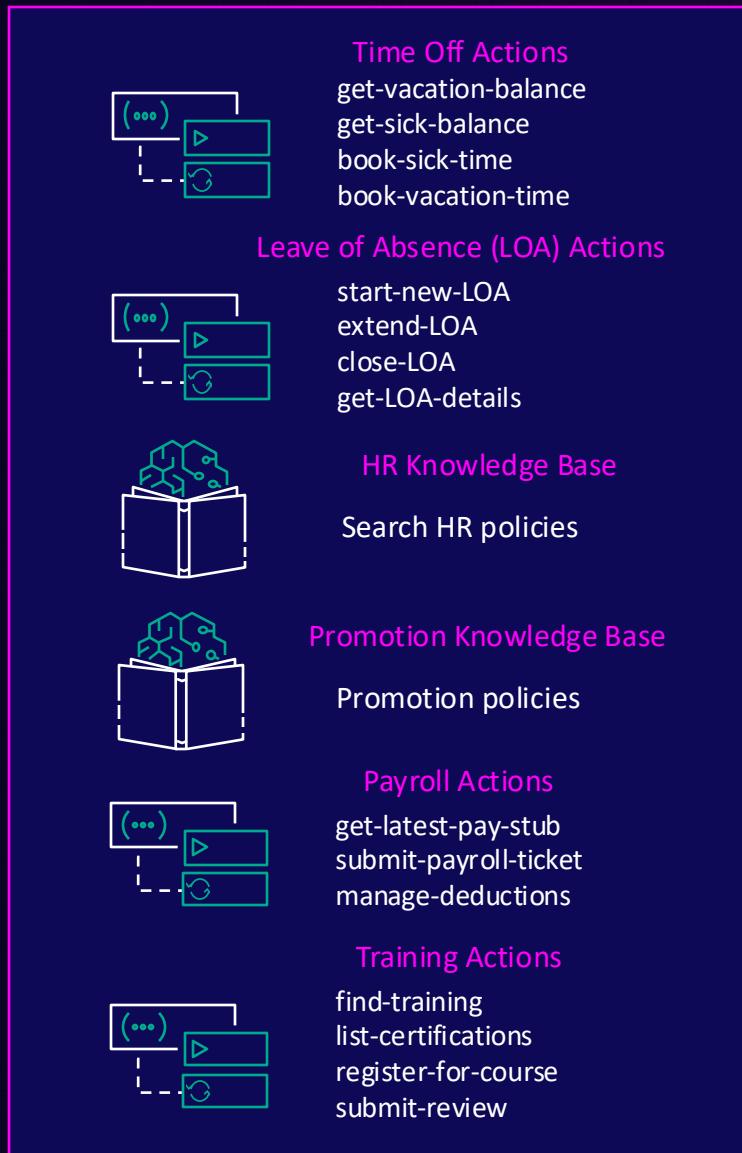
    response_body = {'application/json':
                    {'body': result}}
    action_response = { ...
                       'responseBody': response_body}

    return {'messageversion': '1.0',
           'response': action_response}
```

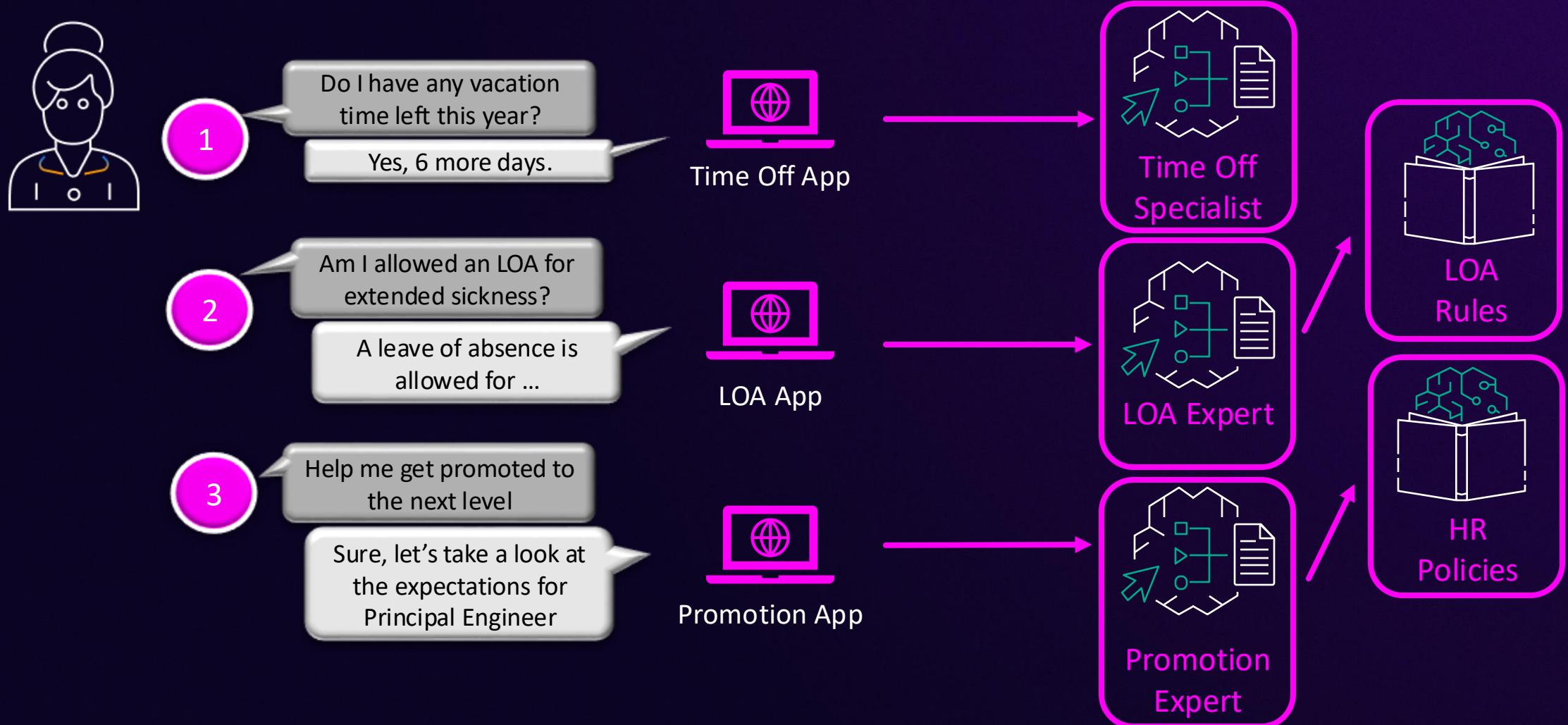


... but if you take a SINGLE agent too far

V4



Using multiple agents helps . . .



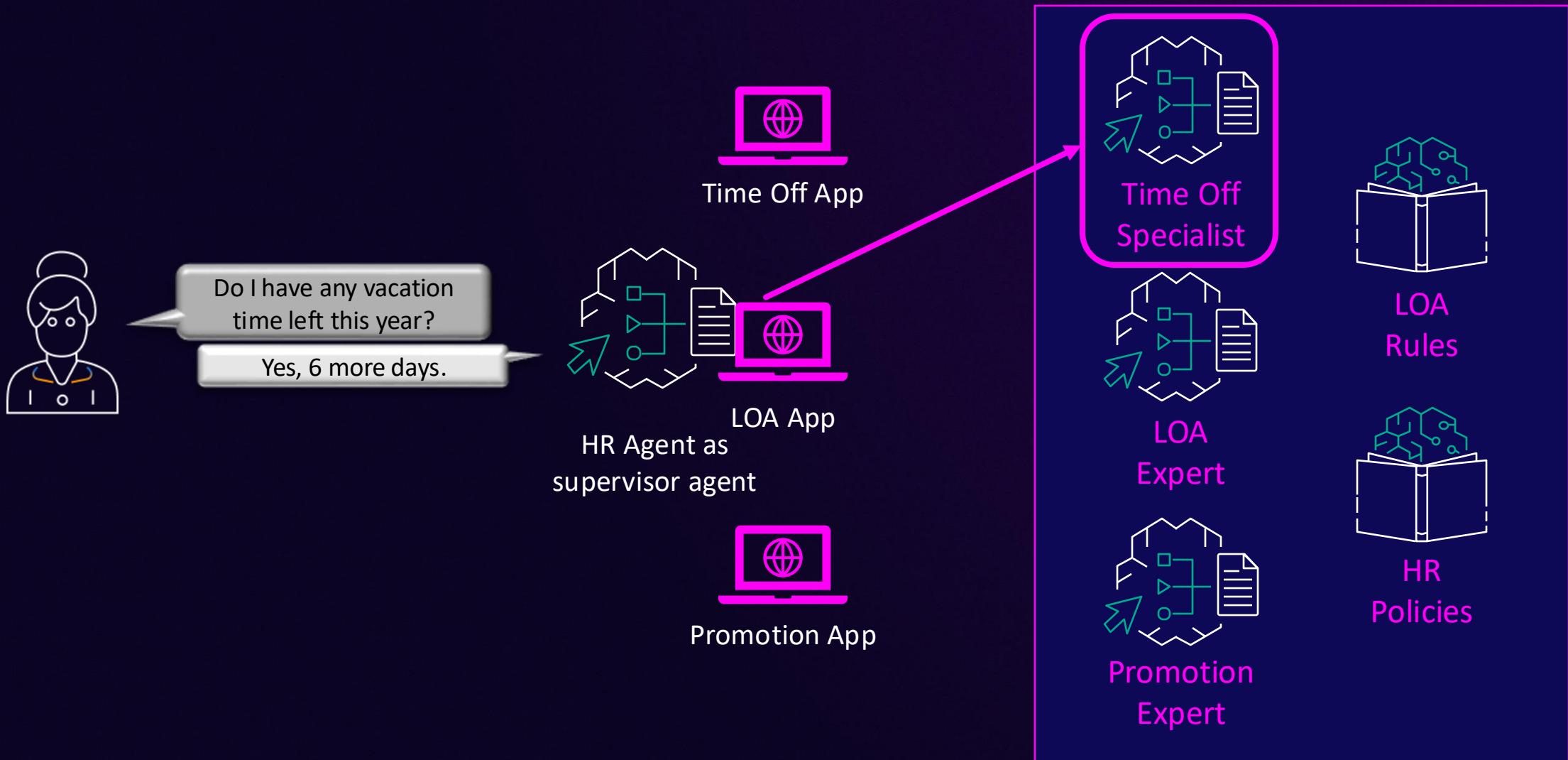
Amazon Bedrock Agents multi-agent collaboration

Scaling agentic experiences

Preview



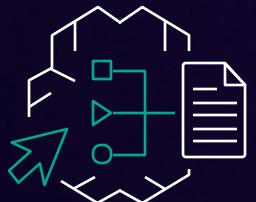
Unify customer experience



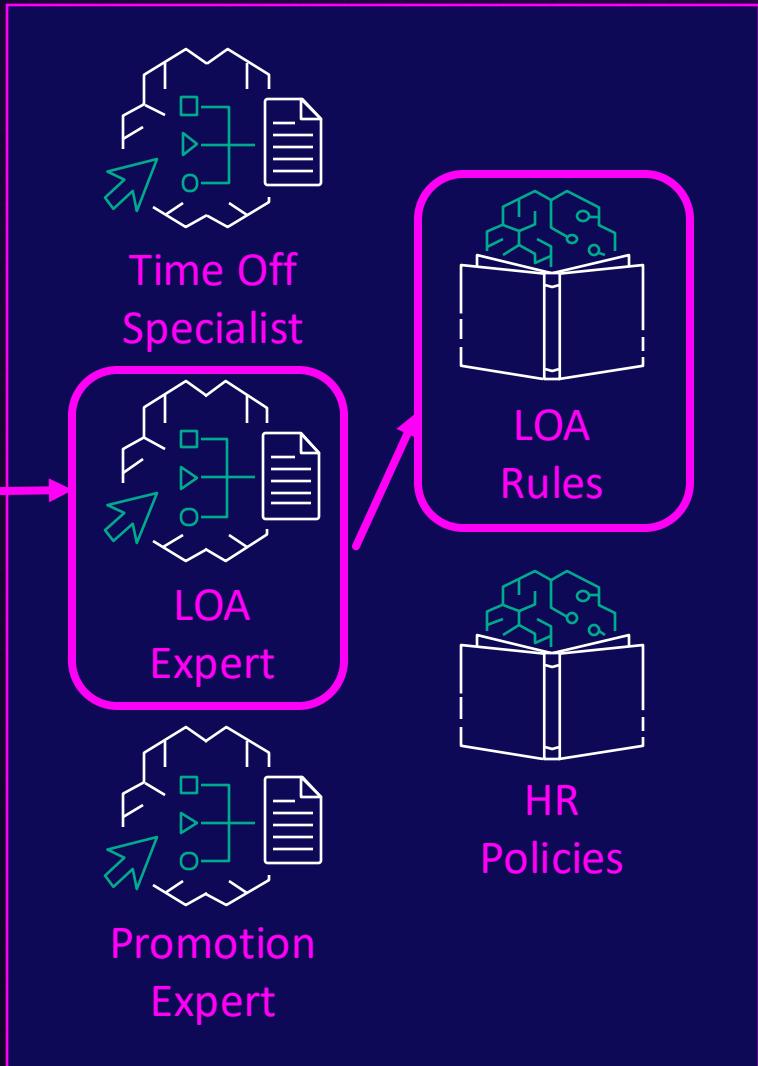
Unify customer experience



Am I allowed an LOA for extended sickness?
A leave of absence is allowed for ...



HR Agent as supervisor agent



Unify customer experience

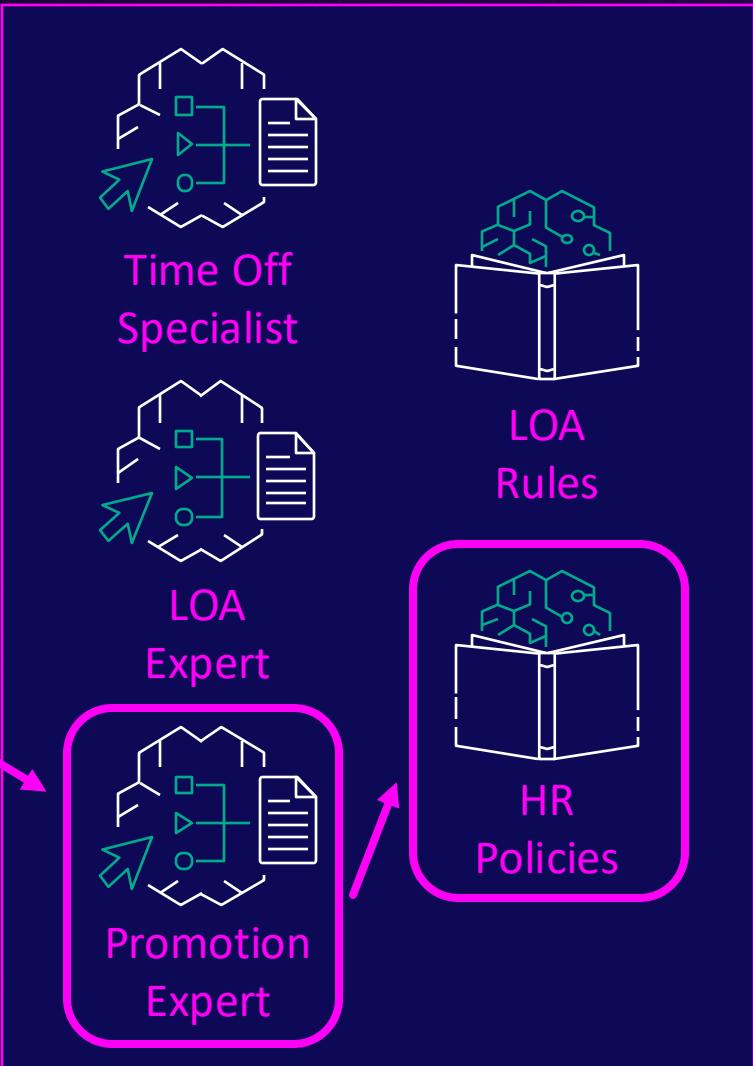
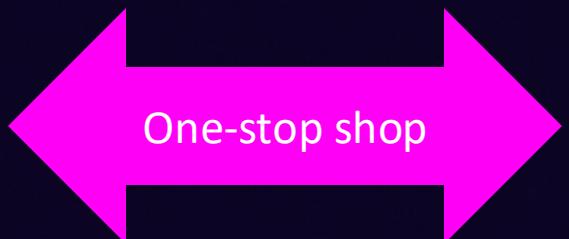
Supervisor provides one-stop shop for customers, with optimized routing and seamless conversation handling



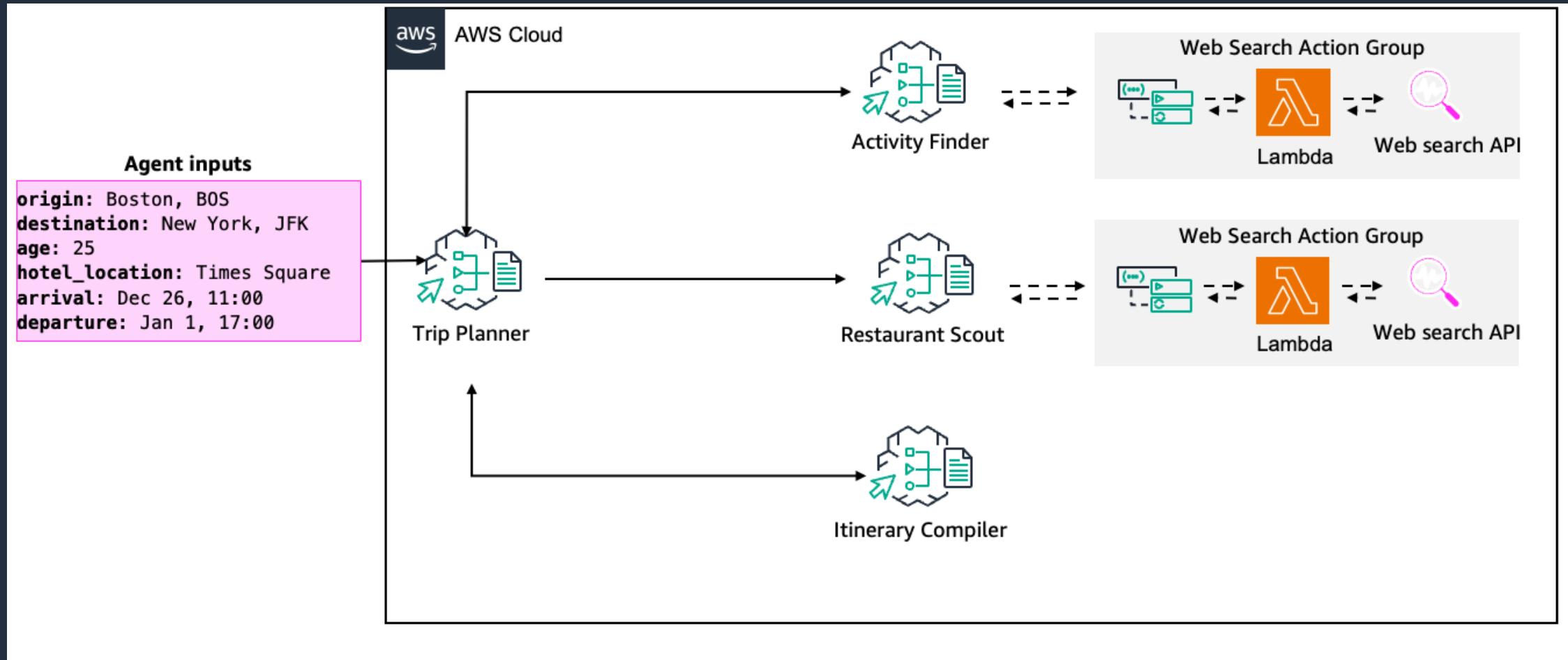
Help me get promoted to the next level
Sure, let's take a look at the expectations for Principal Engineer



HR Agent as supervisor agent



A Trip Planner: a Multi-Agent Illustration



https://github.com/awslabs/amazon-bedrock-agent-samples/tree/main/examples/multi_agent_collaboration/trip_planner_agent



© 2025, Amazon Web Services, Inc. or its affiliates.

**Final response:**

Here's your comprehensive day-by-day itinerary for your trip to New York City from Dec 26 (11:00) to Jan 1 (17:00):

Day 1 (Dec 26):

- 11:00 AM: Arrive and check-in to hotel near Times Square
- 1:00 PM: Lunch at Junior's Restaurant (1515 Broadway)
- 3:00 PM: Visit Macy's Holiday Display at Herald Sq...

Agent made a total of 14 LLM calls, using 38536 tokens (in: 33099, out: 5437), and took 135.2 total seconds

Here's your comprehensive day-by-day itinerary for your trip to New York City from Dec 26 (11:00) to Jan 1 (17:00):

Day 1 (Dec 26):

- 11:00 AM: Arrive and check-in to hotel near Times Square
- 1:00 PM: Lunch at Junior's Restaurant (1515 Broadway)
- 3:00 PM: Visit Macy's Holiday Display at Herald Square
- 6:00 PM: Dinner at Empire Steak House (151 W 49th St)
- 8:00 PM: Evening walk in Times Square

Day 2 (Dec 27):

- 9:00 AM: Breakfast at Best Bagel & Coffee (225 W 35th st)
- 11:00 AM: Winter Train Ride (various routes from NYC)
- 3:00 PM: Late lunch at Friedmans (132 W 31st St)
- 6:00 PM: Lektrik: A Festival of Lights
- 8:30 PM: Dinner at Bourbon Steak (1 W 46th St)

Day 3 (Dec 28):

- 9:00 AM: Breakfast at Times Square Diner (807 8th Ave)
- 11:00 AM: Observation Deck Visit (Empire State Building or Top of the Rock)
- 2:00 PM: Lunch at Junior's Restaurant
- 4:00 PM: Explore Central Park
- 7:00 PM: Dinner at Wolfgang's Steakhouse (250 W 41st St)
- 9:00 PM: Evening entertainment (Broadway show or comedy club)

Day 4 (Dec 29):

- 9:00 AM: Breakfast at Friedmans



- 4:00 PM: Explore Central Park
- 7:00 PM: Dinner at Wolfgang's Steakhouse (250 W 41st St)
- 9:00 PM: Evening entertainment (Broadway show or comedy club)

Day 4 (Dec 29):

- 9:00 AM: Breakfast at Friedmans
- 11:00 AM: Visit the Metropolitan Museum of Art
- 3:00 PM: Late lunch at Times Square Diner
- 5:00 PM: Santa's Secret Holiday Immersive Experience
- 8:00 PM: Dinner at Empire Steak House

Day 5 (Dec 30):

- 9:00 AM: Breakfast at Best Bagel & Coffee
- 11:00 AM: Visit the Statue of Liberty and Ellis Island
- 4:00 PM: Late lunch at Junior's Restaurant
- 6:00 PM: Kwanzaa Celebrations (various locations, including Apollo Theater)
- 9:00 PM: Dinner at Bourbon Steak

Day 6 (Dec 31 - New Year's Eve):

- 10:00 AM: Brunch at Friedmans
- 12:00 PM: Last-minute shopping or museum visit
- 3:00 PM: Return to hotel to prepare for New Year's Eve
- 6:00 PM: Early dinner at Wolfgang's Steakhouse
- 8:00 PM: Head to Times Square for New Year's Eve Celebration
- 12:00 AM: Watch the famous ball drop and celebrate the New Year

Day 7 (Jan 1):

- 11:00 AM: Late breakfast/brunch at Times Square Diner
- 1:00 PM: Relaxing walk in Central Park or visit a museum
- 4:00 PM: Pack and prepare for departure
- 5:00 PM: Depart for airport

Time taken: 135.2 seconds

(.venv) → amazon-bedrock-agent-samples git:(main) ✘

```
mba@mbas-MacBook-Air:~/Desktop/_Jupyter/GenAIWorkshop_20250324/amazon-bedrock-agent-samples
```

```
(.venv) → amazon-bedrock-agent-samples git:(main) ✘ python3 examples/multi_agent_collaboration/trip_planner_agent/main.py --recreate_agents "false"
boto3 version: 1.37.18
```

```
Creating supervisor agent...
```

```
Invoking supervisor agent...
```

```
time before call: 2025-03-22 13:39:16.471621
```

```
invokeAgent API request ID: 52743ee4-1c9a-4769-b2fc-a5fc7653169d
invokeAgent API session ID: trip_planner-1742621956-fe701bf2-06df-11f0-b48b-ee69f0faef78
agent id: Z3NQRCAVGE, agent alias id: QE7SH0BMOM
---- Step 1 ----
```

```
Took 8.4s, using 2034 tokens (in: 1754, out: 280) to complete prior action, observe, orchestrate.
I'll help coordinate these tasks sequentially. Let me break this down:
```

1. First, I need to get activities from the activity finder
2. Then, I'll get restaurant recommendations
3. Finally, I'll have the itinerary compiler create a comprehensive plan
4. I'll make sure to use the specified working memory table

```
Let me start with Task 1 by contacting the activity finder.
```

```
Using sub-agent collaborator: 'activity_finder [C7FUPYQCY2/XAAKZKRSUL]' passing input text:
```

```
Please find activities in New York City for a 25-year-old traveler. They will be staying near Times Square from Dec 26 (11:00) to Jan 1 (17:00). No specific interests provided. Please store your findings in the working memory table 'trip-planner-2278f8ac-557a-471f-b57f-ea0b5ca4c8ed' under the key 'a'
```

```
---- Step 1.1 [using sub-agent name:activity_finder, id:C7FUPYQCY2/XAAKZKRSUL] ----
```

---- Step 1.1 [using sub-agent name:activity_finder, id:C7FUPYQCY2/XAAKZRSUL] ----

Took 5.1s, using 1403 tokens (in: 1207, out: 196) to complete prior action, observe, orchestrate.

To find activities in New York City for a 25-year-old traveler staying near Times Square from Dec 26 to Jan 1, I'll need to use the web search function to gather information about suitable activities. After that, I'll store the findings in the working memory table as requested. Let's start with the web search.

Using tool: web_search with these inputs:

```
[{"name": "topic", "type": "string", "value": "general"}, {"name": "days", "type": "string", "value": "30"}, {"name": "search_query", "type": "string", "value": "Best activities in New York City near Times Square for young adults Dec 26 - Jan 1"}]
```

--tool outputs:

Here are the top search results for the query 'Best activities in New York City near Times Square for young adults Dec 26 - Jan 1':
: {"query": "Best activities in New York City near Times Square for young adults Dec 26 - Jan 1", "follow_up_questions": null, "answer": null, "images": [], "results": [{"url": "ht..."}]}

---- Step 1.2 [using sub-agent name:activity_finder, id:C7FUPYQCY2/XAAKZRSUL] ----

Took 18.2s, using 2976 tokens (in: 2303, out: 673) to complete prior action, observe, orchestrate.

Based on the search results, I've found some interesting activities for the 25-year-old traveler staying near Times Square from Dec 26 to Jan 1. I'll now compile this information and store it in the working memory table as requested.

Using tool: set_value_for_key with these inputs:

```
[{"name": "value", "type": "string", "value": "{\n    \"activities\": [\n        {\n            \"name\": \"Kwanzaa Celebrations\", \n            \"location\": \"Various locations, including Apollo Theater\", \n            \"description\": \"Participate in festive Kwanzaa happenings, including a Kwanzaa crawl to support Black-owned establishments and celebrations at the Apollo Theater.\", \n            \"rating\": \"Not provided\"\n        },\n        {\n            \"name\": \"Winter Train Ride\", \n            \"location\": \"Various routes from NYC\", \n            \"description\": \"Enjoy a cozy winter train ride, watching snow-covered hills roll by on a picturesque journey.\", \n            \"rating\": \"Not provided\"\n        },\n        {\n            \"name\": \"Macy\\'s Holiday Display\", \n            \"location\": \"Herald Square\", \n            \"description\": \"Marvel at the Macy\\'s holiday display, featuring an adorable tale about Tiptoe the reindeer.\", \n            \"rating\": \"Not provided\"\n        },\n        {\n            \"name\": \"Lektrik: A Festival of Lights\", \n            \"location\": \"Not specified\", \n            \"description\": \"Walk on illuminated trails lined with dozens of light sculptures.\", \n            \"rating\": \"Not provided\"\n        },\n        {\n            \"name\": \"Santa\\'s Secret Holiday Immersive Experience\", \n            \"location\": \"Not specified\", \n            \"description\": \"A naughty-themed holiday immersive experience for adults.\", \n            \"rating\": \"Not provided\"\n        },\n        {\n            \"name\": \"Observation Deck Visit\", \n            \"location\": \"Not specified (likely Empire State Building or Top of the Rock)\", \n            \"description\": \"Ascend to an iconic observation deck for panoramic night views, historical exhibits, and a King Kong-themed photo experience.\", \n            \"rating\": \"Not provided\"\n        },\n        {\n            \"name\": \"New Year\\'s Eve in Times Square\", \n            \"location\": \"Times Square\", \n            \"description\": \"Experience the world-famous New Year\\'s Eve ball drop and celebration.\", \n            \"rating\": \"Not provided\"\n        }\n    ]\n}, {\n    \"name\": 'table_name', \n    \"type\": 'string', \n    \"value\": 'tr...'}
```



---- Step 1.3 [using sub-agent name:activity_finder, id:C7FUPYQCY2/XAAKZKRSUL] ----

Took 16.2s, using 3811 tokens (in: 3599, out: 212) to complete prior action, observe, orchestrate.

I have successfully stored the activities information in the working memory table as requested. Now, I'll provide a concise summary of the findings to the user.

Final response:

I've found several exciting activities for your stay in New York City from Dec 26 to Jan 1. These include:

1. Kwanzaa Celebrations at various locations, including the Apollo Theater
2. Scenic winter train rides from NYC
3. Macy's Holiday Display at Herald Square
4. Lektrik: A Festival of Lights wit...

----sub-agent activity_finder output text:

I've found several exciting activities for your stay in New York City from Dec 26 to Jan 1. These include:

1. Kwanzaa Celebrations at various locations, including the Apollo Theater
2. Scenic winter train rides from NYC
3. Macy's Holiday Display at Herald Square
4. Lektrik: A Festival of Lights wit...

---- Step 2 ----

Took 5.2s, using 2432 tokens (in: 2210, out: 222) to complete prior action, observe, orchestrate.

Now that I have the activities, let me proceed with Task 2 and contact the restaurant scout for dining recommendations.

Using sub-agent collaborator: 'restaurant_scout [WFNRTHUFDG/D988EGPI1P]' passing input text:

Please find restaurants in New York City for a 25-year-old traveler staying near Times Square from Dec 26 to Jan 1. They prefer steak and seafood. Need recommendations for dinners and some lunches/breakfasts. Please store your findings in the working memory table 'trip-planner-2278f8ac-557a-471f-b57

---- Step 2.1 [using sub-agent name:restaurant_scout, id:WFNRTHUFDG/D988EGPI1P] ----

Took 3.6s, using 1401 tokens (in: 1227, out: 174) to complete prior action, observe, orchestrate.

To fulfill this request, I'll need to search for restaurants in New York City that match the traveler's preferences. I'll focus on steak and seafood restaurants near Times Square, considering both dinner and lunch/breakfast options. After gathering the information, I'll store it in the specified working memory table. Let's start by searching for relevant restaurants.

^@Using tool: web_search with these inputs:

---- Step 2.2 [using sub-agent name:restaurant_scout, id:WFnRTHUFDG/D988EGPI1P] ----

Took 6.5s, using 2082 tokens (in: 1918, out: 164) to complete prior action, observe, orchestrate.

Based on the search results, I've found some great options for steak and seafood restaurants near Times Square. Now, I'll compile this information and search for a few more breakfast and lunch options to provide a well-rounded list of recommendations. Then, I'll store all the gathered information in the specified working memory table.

Using tool: web_search with these inputs:

```
[{"name": "topic", "type": "string", "value": "general"}, {"name": "search_query", "type": "string", "value": "Best breakfast and lunch restaurants near Times Square, New York City"}]
```

--tool outputs:

Here are the top search results for the query 'Best breakfast and lunch restaurants near Times Square, New York City': {"query": "Best breakfast and lunch restaurants near Times Square, New York City", "follow_up_questions": null, "answer": null, "images": [], "results": [{"url": "https://www.yelp.com/search?..."}]}

---- Step 2.3 [using sub-agent name:restaurant_scout, id:WFnRTHUFDG/D988EGPI1P] ----

Took 17.5s, using 3359 tokens (in: 2529, out: 830) to complete prior action, observe, orchestrate.

Now that I have gathered information on steak, seafood, breakfast, and lunch options near Times Square, I'll compile a comprehensive list of restaurant recommendations and store it in the specified working memory table.

Using tool: set_value_for_key with these inputs:

```
[{"name": "value", "type": "string", "value": "{\n    \"dinner_options\": [\n        {\n            \"name\": \"Empire Steak House\", \n            \"address\": \"151 W 49th St, New York, NY 10019\", \n            \"cuisine\": \"Steakhouse, Seafood\", \n            \"description\": \"Top-rated NYC steakhouse renowned for USDA Prime Dry-Aged steaks, fresh seafood, and an extensive wine selection.\", \n            \"rating\": \"4.5/5\"\n        },\n        {\n            \"name\": \"Bourbon Steak\", \n            \"address\": \"1 W 46th St, New York, NY 10036\", \n            \"cuisine\": \"Steakhouse, American\", \n            \"description\": \"Upscale steakhouse offering prime cuts and innovative cocktails in a sophisticated setting.\", \n            \"rating\": \"4.4/5\"\n        },\n        {\n            \"name\": \"Wolfgang\\'s Steakhouse\", \n            \"address\": \"250 W 41st St, New York, NY 10036\", \n            \"cuisine\": \"Steakhouse\", \n            \"description\": \"Classic New York steakhouse known for dry-aged steaks and a refined atmosphere.\", \n            \"rating\": \"4.3/5\"\n        }\n    ], \n    \"lunch_breakfast_options\": [\n        {\n            \"name\": \"Times Square Diner\", \n            \"address\": \"807 8th Ave, New York, NY 10019\", \n            \"cuisine\": \"American, Diner\", \n            \"description\": \"Classic NYC diner offering hearty breakfast and lunch options in a casual setting.\", \n            \"rating\": \"4.2/5\"\n        },\n        {\n            \"name\": \"Junior\\'s Restaurant\", \n            \"address\": \"1515 Broadway, New York, NY 10036\", \n            \"cuisine\": \"American, Deli\", \n            \"description\": \"Iconic NYC restaurant famous for cheesecake, also serving breakfast and lunch classics.\", \n            \"rating\": \"4.1/5\"\n        },\n        {\n            \"name\": \"Friedmans\", \n            \"address\": \"132 W 31st St, New York, NY 10001\", \n            \"cuisine\": \"American, Breakfast & Brunch\", \n            \"description\": \"Popular spot for gluten-free options and comfort food, great for breakfast and lunch.\", \n            \"rating\": \"4.3/5\"\n        }\n    ]\n}"]
```



---- Step 2.4 [using sub-agent name:restaurant_scout, id:WFnRTHUFdg/D988EGPI1P] ----

Took 6.4s, using 4431 tokens (in: 4175, out: 256) to complete prior action, observe, orchestrate.

Great! I've successfully stored the restaurant recommendations in the specified working memory table. Now, I'll provide a concise summary of the recommendations to the user.

Final response:

I've found and stored several restaurant recommendations for your trip to New York City near Times Square. Here's a summary:

For dinner options focusing on steak and seafood:

1. Empire Steak House: Known for USDA Prime Dry-Aged steaks and fresh seafood.
2. Bourbon Steak: Upscale steakhouse with pri...

----sub-agent restaurant_scout output text:

I've found and stored several restaurant recommendations for your trip to New York City near Times Square. Here's a summary:

For dinner options focusing on steak and seafood:

1. Empire Steak House: Known for USDA Prime Dry-Aged steaks and fresh seafood.
2. Bourbon Steak: Upscale steakhouse with pri...

---- Step 3 ----

Took 5.5s, using 2911 tokens (in: 2651, out: 260) to complete prior action, observe, orchestrate.

Now that I have both activities and restaurant recommendations, I can proceed with Task 3. I'll ask the itinerary compiler to create a comprehensive day-by-day plan.

Using sub-agent collaborator: 'itinerary_compiler [BFKYBUYKYX/L0YDTCY7ZJ]' passing input text:

Please create a comprehensive day-by-day itinerary for a trip to New York City. The traveler is 25 years old, staying near Times Square from Dec 26 (11:00) to Jan 1 (17:00). Use the activities and restaurants stored in the working memory table 'trip-planner-2278f8ac-557a-471f-b57f-ea0b5ca4c8ed' unde

---- Step 3.1 [using sub-agent name:itinerary_compiler, id:BFKYBUYKYX/L0YDTCY7ZJ] ----

Took 4.4s, using 1294 tokens (in: 1098, out: 196) to complete prior action, observe, orchestrate.

To create a comprehensive day-by-day itinerary for the trip to New York City, I need to retrieve the stored activities and restaurants from the working memory table. Then, I'll organize this information into a daily schedule. After creating the itinerary, I'll delete the working memory table as requested. Let's start by retrieving the necessary information.

Using tool: get_key_value with these inputs:

```
[{'name': 'table_name', 'type': 'string', 'value': 'trip-planner-2278f8ac-557a-471f-b57f-ea0b5ca4c8ed'}, {'name': 'key', 'type': '
```

---- Step 3.2 [using sub-agent name:itinerary_compiler, id:BFKYBUYKYX/L0YDTCY7ZJ] ----

Took 3.2s, using 2012 tokens (in: 1878, out: 134) to complete prior action, observe, orchestrate.

Now that I have the activities, let's retrieve the restaurants information.

Using tool: get_key_value with these inputs:

```
[{"name": "table_name", "type": "string", "value": "trip-planner-2278f8ac-557a-471f-b57f-ea0b5ca4c8ed"}, {"name": "key", "type": "string", "value": "restaurants"}]
```

--tool outputs:

```
"{\n    \"dinner_options\": [\n        {\n            \"name\": \"Empire Steak House\", \n            \"address\": \"151 W 49th St, New York, NY 10019\", \n            \"cuisine\": \"Steakhouse, Seafood\", \n            \"description\": \"Top-rated NYC steakhouse renowned for USDA Prime Dry-Aged steaks, fresh seafood, and an extens...\"}\n    ]\n}
```

^@---- Step 3.3 [using sub-agent name:itinerary_compiler, id:BFKYBUYKYX/L0YDTCY7ZJ] ----

Took 18.9s, using 3718 tokens (in: 2792, out: 926) to complete prior action, observe, orchestrate.

Now that I have all the necessary information, I'll create a comprehensive day-by-day itinerary for the trip to New York City. I'll organize the activities and restaurants into a daily schedule, taking into account the traveler's age and the duration of their stay. After creating the itinerary, I'll delete the working memory table as requested.

Final response:

Here's a comprehensive day-by-day itinerary for your trip to New York City from Dec 26 (11:00) to Jan 1 (17:00):

Day 1 (Dec 26):

- 11:00 AM: Arrive and check-in to hotel near Times Square
- 1:00 PM: Lunch at Junior's Restaurant (1515 Broadway)
- 3:00 PM: Visit Macy's Holiday Display at Herald Squar...

----sub-agent itinerary_compiler output text:

Here's a comprehensive day-by-day itinerary for your trip to New York City from Dec 26 (11:00) to Jan 1 (17:00):

Day 1 (Dec 26):

- 11:00 AM: Arrive and check-in to hotel near Times Square
- 1:00 PM: Lunch at Junior's Restaurant (1515 Broadway)
- 3:00 PM: Visit Macy's Holiday Display at Herald Squar...



---- Step 4 ----

Took 14.3s, using 4672 tokens (in: 3758, out: 914) to complete prior action, observe, orchestrate.

The itinerary compiler has created a comprehensive plan and deleted the working memory table. I'll now send this final itinerary to the user.

Final response:

Here's your comprehensive day-by-day itinerary for your trip to New York City from Dec 26 (11:00) to Jan 1 (17:00):

Day 1 (Dec 26):

- 11:00 AM: Arrive and check-in to hotel near Times Square
- 1:00 PM: Lunch at Junior's Restaurant (1515 Broadway)
- 3:00 PM: Visit Macy's Holiday Display at Herald Sq...

Agent made a total of 14 LLM calls, using 38536 tokens (in: 33099, out: 5437), and took 135.2 total seconds

Here's your comprehensive day-by-day itinerary for your trip to New York City from Dec 26 (11:00) to Jan 1 (17:00):

Day 1 (Dec 26):

- 11:00 AM: Arrive and check-in to hotel near Times Square
- 1:00 PM: Lunch at Junior's Restaurant (1515 Broadway)
- 3:00 PM: Visit Macy's Holiday Display at Herald Square
- 6:00 PM: Dinner at Empire Steak House (151 W 49th St)
- 8:00 PM: Evening walk in Times Square

Day 2 (Dec 27):

- 9:00 AM: Breakfast at Best Bagel & Coffee (225 W 35th St)
- 11:00 AM: Winter Train Ride (various routes from NYC)
- 3:00 PM: Late lunch at Friedmans (132 W 31st St)
- 6:00 PM: Lektrik: A Festival of Lights
- 8:30 PM: Dinner at Bourbon Steak (1 W 46th St)

Day 3 (Dec 28):

- 9:00 AM: Breakfast at Times Square Diner (807 8th Ave)
- 11:00 AM: Observation Deck Visit (Empire State Building or Top of the Rock)
- 2:00 PM: Lunch at Junior's Restaurant
- 4:00 PM: Explore Central Park
- 7:00 PM: Dinner at Wolfgang's Steakhouse (250 W 41st St)

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- **Validating Bedrock agent**
- Exploring Bedrock fine-tune
- Q & A

Amazon Bedrock Workshop

▶ Prerequisites

Prompt Engineering

Text Generation

Knowledge Bases and RAG

Model Customization

Image and Video Generation

Applications

▼ Agents

[Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation](#)

Multi-Step Bedrock Agent using Jupyter Notebooks

Open Source With Bedrock

▼ AWS account access

[Open AWS console \(us-west-2\)](#) [Get AWS CLI credentials](#)

▼ Content preferences

Language

English

Exit event



Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation

Creating an AI agent with code interpretation capabilities using the Amazon Bedrock console

This section walks you through the process of creating an AI agent with code interpretation capabilities using the Amazon Bedrock console. This agent will be able to perform data analysis tasks and generate visualizations.

Prerequisites

- An AWS account with access to Amazon Bedrock
- Familiarity with the AWS Management Console
- Enable required models from Amazon Bedrock Model Access.

Instructions

1. Access Amazon Bedrock Console

- Sign in to the AWS Management Console
- Navigate to the Amazon Bedrock service
- Select "Agents" from the left navigation pane



Amazon Bedrock



Amazon Bedrock > Agents



Getting started

Overview

Providers

Foundation models

Model catalog [New](#)

Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

Image / Video

Builder tools

[Agents](#) 

Flows

Knowledge Bases

Prompt Management

Safeguards

Guardrails

Watermark detection

Inference and Assessment

Agents Info

▼ Overview



Prepare

Create your Agent by selecting a Foundation model, and adding Action groups. After creation you can test out the Agent in real-time and create multiple versions.



Deploy

Create and associate Aliases to deploy an Agent version in your application. Point an Alias to a specific version of your Agent to test it before deploying it to your client application.

Agents (0)

 Find Agents

[Delete](#)

[Edit](#)

[Create agent](#)

< 1 >



Name	Status	Description	Last updated
------	--------	-------------	--------------

No agents

No agents to display

Amazon Bedrock

Agents 

Overview

Prepare

Create your Agent by selecting a Foundation model, and adding Action groups. After creation you can test out the Agent in real-time and create multiple versions.

Deploy

Create and associate Aliases to deploy an Agent version in your application. Point an Alias to a specific version of your Agent to test it before deploying it to your client application.

Agents (0)

 Find Agents

    1  

Name	Status	Description	Last updated
No agents			
No agents to display			



Amazon Bedrock <

▼ Getting started

Overview

Providers

▼ Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

▼ Playgrounds

Chat / Text

Image / Video

▼ Builder tools

[Agents](#)

Flows

Knowledge Bases

Prompt Management

▼ Safeguards

Guardrails

Watermark detection

▼ Inference and Assessment



Amazon Bedrock > Agents

Agents Info

▼ Overview



Prepare

Create your Agent groups. After creating groups, you can publish multiple versions of your Agent.

Agents (0)



Name

Description - optional

Multi-agent collaboration

[Learn more about multi-agent collaboration](#) Enable multi-agent collaboration

If you have more than one agent, multi-agent collaboration allows this agent to associate the others as agent collaborators to orchestrate responses. You can change this later.

Cancel

Create

Use this page to deploy an Agent version in your application. Publish a new version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Version 1 of your Agent has been published. This is the first version of your Agent to test it before deploying it to production.

Amazon Bedrock Workshop

▶ Prerequisites

Prompt Engineering

Text Generation

Knowledge Bases and RAG

Model Customization

Image and Video Generation

Applications

▼ Agents

[Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation](#)

Multi-Step Bedrock Agent using Jupyter Notebooks

Open Source With Bedrock

▼ AWS account access

[Open AWS console \(us-west-2\)](#) [Get AWS CLI credentials](#)

▼ Content preferences

Language

English

Exit event

2. Create a New Agent

- Click on "Create Agent"
- Provide a name for your agent "DataAnalysisAssistant"
- Add a description (Optional) - "AI Agent with Code Interpreter enabled for Data analysis"
- Click "Create"

Create agent

Name

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Description - optional

The description can have up to 200 characters.

Multi-agent collaboration - Preview



Amazon Bedrock <

▼ Getting started

Overview

Providers

▼ Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

▼ Playgrounds

Chat / Text

Image / Video

▼ Builder tools

[Agents](#)

Flows

Knowledge Bases

Prompt Management

▼ Safeguards

Guardrails

Watermark detection

▼ Inference and Assessment

Amazon Bedrock > Agents

Agents Info

▼ Overview



Prepare

Create your Agent groups. After creating groups, you can create multiple versions of your Agent.

Agents (0)

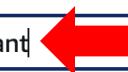


Name



Name

DataAnalysisAssistant



Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Description - optional

Enter description

The description can have up to 200 characters.

Multi-agent collaboration

[Learn more about multi-agent collaboration](#) Enable multi-agent collaboration

If you have more than one agent, multi-agent collaboration allows this agent to associate the others as agent collaborators to orchestrate responses. You can change this later.

Cancel

Create

uses to deploy an Agent version in your application. Create a new Agent version of your Agent to test it before deploying it to production.

[Delete](#) [Edit](#) [Create agent](#)

< 1 >

Last updated



Amazon Bedrock Workshop

▶ Prerequisites

Prompt Engineering

Text Generation

Knowledge Bases and RAG

Model Customization

Image and Video Generation

Applications

▼ Agents

[Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation](#)

Multi-Step Bedrock Agent using Jupyter Notebooks

Open Source With Bedrock

▼ AWS account access

[Open AWS console \(us-west-2\)](#)[Get AWS CLI credentials](#)

▼ Content preferences

Language

English

Exit event

2. Create a New Agent

- Click on "Create Agent"
- Provide a name for your agent "DataAnalysisAssistant"
- Add a description (Optional) - "AI Agent with Code Interpreter enabled for Data analysis" 
- Click "Create"

Create agent

Name

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Description - optional

The description can have up to 200 characters.

Multi-agent collaboration - Preview



Amazon Bedrock



Amazon Bedrock > Agents

Agents Info

▼ Overview



Prepare

Create your Agent groups. After creating groups, you can add multiple versions of your Agent.

▼ Getting started

Overview

Providers

▼ Foundation models

Model catalog NewMarketplace deployments New

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers Preview

▼ Playgrounds

Chat / Text

Image / Video

▼ Builder tools

Agents

Flows

Knowledge Bases

Prompt Management

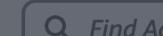
▼ Safeguards

Guardrails

Watermark detection

▼ Inference and Assessment

Agents (0)



Name



Create agent



Name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Description - optional

AI Agent with Code Interpreter enabled for Data analysis



The description can have up to 200 characters.

Multi-agent collaboration

[Learn more about multi-agent collaboration](#) Enable multi-agent collaboration

If you have more than one agent, multi-agent collaboration allows this agent to associate the others as agent collaborators to orchestrate responses. You can change this later.

Cancel

Create



Amazon Bedrock 

Getting started

[Overview](#)[Providers](#)

Foundation models

[Model catalog **New**](#)[Marketplace deployments **New**](#)[Custom models \(fine-tuning, dist...\)](#)[Imported models](#)[Prompt Routers **Preview**](#)

Playgrounds

[Chat / Text](#)[Image / Video](#)

Builder tools

[Agents](#)[Flows](#)[Knowledge Bases](#)[Prompt Management](#)

Safeguards

[Guardrails](#)[Watermark detection](#)

Inference and Assessment

✓ Agent: DataAnalysisAssistant was successfully created. 

[Amazon Bedrock](#) > [Agents](#) > [DataAnalysisAssistant](#) > Agent builder: DataAnalysisAssist...Agent builder [Manual](#)[Assistant](#)[Test](#)[Prepare](#)[Save](#)[Save and exit](#)

Agent details

Agent name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

AI Agent with Code Interpreter enabled for Data analysis

The description can have up to 200 characters.

Agent resource role

- Create and use a new service role
- Use an existing service role

AmazonBedrockExecutionRoleForAgents_IKF84K5V4SC

Select model

[Select model](#) 

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain

Test Agent

   Using ODT [Change](#)

Enter your message here

...

[Run](#)

Getting started

Overview

Providers

Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

Image / Video

Builder tools

Agents

Flows

Knowledge Bases

Prompt Management

Safeguards

Guardrails

Watermark detection

Inference and Assessment

Select model



Search available models and inference

 Bedrock Agents optimized [Learn more](#)

1. Categories

Model providers



2. Models

Models with access (2)

Claude 3 Sonnet v1 Legacy

Text & vision model | Max 200k tokens

Claude 3 Haiku v1

Text & vision model | Max 200k tokens

Models without access (7) [Request access](#)**Claude 3.5 Sonnet v2**

Text & vision model | Max 200k tokens

Claude 3.5 Haiku v1

Text model | Max 200k tokens

Claude Instant v1.2 Legacy

Text model | Max 100k tokens

Claude v2.1 Legacy

Text model | Max 200k tokens

3. Inference

On demand

On-demand

 ⓘ Can't find the model you are looking for? See all models [here](#)

Cancel

Apply

Run

aws Services Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 6757-9927-5846 ▾

Support

Amazon Bedrock <

Getting started

- Overview
- Providers

Foundation models

- Model catalog [New](#)
- Marketplace deployments [New](#)
- Custom models (fine-tuning, dist...)
- Imported models
- Prompt Routers [Preview](#)

Playgrounds

- Chat / Text
- Image / Video

Builder tools

- Agents
- Flows
- Knowledge Bases
- Prompt Management

Safeguards

- Guardrails
- Watermark detection

Inference and Assessment

Select model

Claude 3 Sonnet v1 ⓘ ⚙️ On-demand

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

Enter instructions

This instruction must have a minimum of 40 characters.

► Additional settings

Action groups (0) [Info](#)

Delete Add

Find action groups

Name Description State ⚙️ Last updated ▾

No action groups

No actions group to display

Add

Test Agent

Using ODT [Change](#)

Enter your message here

Run

⋮

Information icon

Help icon

Settings icon

File icon

Edit icon

Back icon

Forward icon

Information icon

Help icon

Settings icon

catalog.us-east-1.prod.workshops.aws/event/dashboard/en-US/workshop/150-agents/151-code-intpreter-agent

aws workshop studio

Amazon Bedrock Workshop

- ▶ Prerequisites
- Prompt Engineering
- Text Generation
- Knowledge Bases and RAG
- Model Customization
- Image and Video Generation
- Applications

▼ Agents

- [Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation](#)
- Multi-Step Bedrock Agent using Jupyter Notebooks
- Open Source With Bedrock

▼ AWS account access

- [Open AWS console \(us-west-2\)](#)
- [Get AWS CLI credentials](#)

▼ Content preferences

- Language [English](#)
- Exit event

3. Configure Agent Details

- In the Agent builder, go to the "Agent details" section
- For "Agent resource role", select "Create and use a new service role"
- Choose a foundation model optimized for agents (e.g., Claude 3.5 Sonnet v2)
- In "Instructions for the Agent", enter:

I can write, run, and debug Python code to help users with their queries. Always provide clear explanations of your process and results.

Amazon Bedrock > Agents > DataAnalysisAssistant > Agent builder: DataAnalysisAssistant

Agent builder [Info](#)

Manual Assistant Test Prepare Save Save and exit

Agent details

Agent name Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional The description can have up to 200 characters.

Agent resource role Create and use a new service role Use an existing service role



aws Services Search [Option+S] United States (Oregon) WSParticipantRole/Participant @ 6757-9927-5846 ▾

Support

Amazon Bedrock <

Getting started

- Overview
- Providers

Foundation models

- Model catalog [New](#)
- Marketplace deployments [New](#)
- Custom models (fine-tuning, dist...)
- Imported models
- Prompt Routers [Preview](#)

Playgrounds

- Chat / Text
- Image / Video

Builder tools

- Agents
- Flows
- Knowledge Bases
- Prompt Management

Safeguards

- Guardrails
- Watermark detection

Inference and Assessment

Select model

Claude 3 Sonnet v1 [i](#) [p](#)
On-demand

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

You are an AI assistant specialized in data analysis and visualization. You can write, run, and debug Python code to help users with their queries. Always provide clear explanations of your process and results.

This instruction must have a minimum of 40 characters.

Additional settings

Action groups (0) [Info](#) [Delete](#) [Add](#)

Find action groups

Name Description State Last updated

No action groups

No actions group to display [Add](#)

Test Agent

Using ODT [Change](#)

Enter your message here

Run

Amazon Bedrock 

▼ Getting started

[Overview](#)[Providers](#)

▼ Foundation models

[Model catalog !\[\]\(e144deadf8fb653921edc1933152e159_img.jpg\)](#)[Marketplace deployments !\[\]\(eca5aa4c3a85a7dbb72af21f75bf904e_img.jpg\)](#)[Custom models \(fine-tuning, dist...\)](#)[Imported models](#)[Prompt Routers !\[\]\(a80586c4f5aace3e67e6cb2abc35811b_img.jpg\)](#)

▼ Playgrounds

[Chat / Text](#)[Image / Video](#)

▼ Builder tools

[Agents](#)[Flows](#)[Knowledge Bases](#)[Prompt Management](#)

▼ Safeguards

[Guardrails](#)[Watermark detection](#)

▼ Inference and Assessment

[CloudShell](#)[Feedback](#)

Select model

**Claude 3 Sonnet v1**  

On-demand

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

You are an AI assistant specialized in data analysis and visualization. You can write, run, and debug Python code to help users with their queries. Always provide clear explanations of your process and results.

This instruction must have a minimum of 40 characters.

▼ Additional settings

Code Interpreter

Code Interpreter enables agents to handle tasks that involve writing, running, testing, and troubleshooting code in a secure environment.

 Enabled Disabled

User input

Select whether the agent can prompt additional information from the user when it does not have enough information to respond to an utterance.

 Enabled

Allow agent to ask the user clarifying questions to capture necessary inputs.

 Disabled

The selected foundation model within the Agent will make a best guess at invoking the appropriate action groups.

KMS key selection 

Your data is encrypted by default with a key that AWS owns and manages for you. To choose a different key, customize your encryption settings.

 Customize encryption settings (advanced)

To use the default key, clear this option.

Test Agent

Using ODT 

Enter your message here

...

Run



Support

Amazon Bedrock



Getting started

Overview

Providers

Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

Image / Video

Builder tools

Agents

Flows

Knowledge Bases

Prompt Management

Safeguards

Guardrails

Watermark detection

Inference and Assessment

Select model

Claude 3 Sonnet v1 [Edit](#) [Delete](#)
On-demand

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

You are an AI assistant specialized in data analysis and visualization. You can write, run, and debug Python code to help users with their queries. Always provide clear explanations of your process and results.

This instruction must have a minimum of 40 characters.

Additional settings

Code Interpreter

Code Interpreter enables agents to handle tasks that involve writing, running, testing, and troubleshooting code in a secure environment.

 Enabled Disabled

User input

Select whether the agent can prompt additional information from the user when it does not have enough information to respond to an utterance.

 Enabled

Allow agent to ask the user clarifying questions to capture necessary inputs.

 Disabled

The selected foundation model within the Agent will make a best guess at invoking the appropriate action groups.

KMS key selection | [Info](#)

Your data is encrypted by default with a key that AWS owns and manages for you. To choose a different key, customize your encryption settings.

 Customize encryption settings (advanced)

To use the default key, clear this option.

Test Agent

Using ODT [Change](#)



Support

Amazon Bedrock



Getting started

Overview

Providers

Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

Image / Video

Builder tools

Agents

Flows

Knowledge Bases

Prompt Management

Safeguards

Guardrails

Watermark detection

Inference and Assessment

CloudShell

Feedback

Select model

Claude 3 Sonnet v1 [Info](#) [Edit](#)
On-demand

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

You are an AI assistant specialized in data analysis and visualization. You can write, run, and debug Python code to help users with their queries. Always provide clear explanations of your process and results.

This instruction must have a minimum of 40 characters.

Additional settings

Code Interpreter

Code Interpreter enables agents to handle tasks that involve writing, running, testing, and troubleshooting code in a secure environment.

 Enabled Disabled

User input

Select whether the agent can prompt additional information from the user when it does not have enough information to respond to an utterance.

 Enabled

Allow agent to ask the user clarifying questions to capture necessary inputs.

 Disabled

The selected foundation model within the Agent will make a best guess at invoking the appropriate action groups.

KMS key selection | [Info](#)

Your data is encrypted by default with a key that AWS owns and manages for you. To choose a different key, customize your encryption settings.

 Customize encryption settings (advanced)

To use the default key, clear this option.

Test Agent

Using ODT [Change](#)

Amazon Bedrock



Getting started

[Overview](#)[Providers](#)

Foundation models

[Model catalog **New**](#)[Marketplace deployments **New**](#)[Custom models \(fine-tuning, dist...\)](#)[Imported models](#)[Prompt Routers **Preview**](#)

Playgrounds

[Chat / Text](#)[Image / Video](#)

Builder tools

[Agents](#)[Flows](#)[Knowledge Bases](#)[Prompt Management](#)

Safeguards

[Guardrails](#)[Watermark detection](#)

Inference and Assessment

Idle session timeout

You can configure how long a session is maintained when the user does not provide any input and the session is idle. Amazon Bedrock retains context information until a session ends.

600

Second(s)

By default, session idle timeout is 10 minutes, but you can specify any duration between 1 and 60 minutes.

Action groups (0) [Info](#)

[Delete](#)[Add](#) [Find action groups](#) [1](#) [Name](#)[Description](#)[State](#) [Last updated](#)**No action groups****No actions group to display**[Add](#)

Memory - **New**

Memory allows an agent to retain conversational contexts across multiple sessions and recall past actions and behaviors. [See supported models](#)

Enable session summarization

Enabling session summarization will generate a summary of each test session. [View session summaries](#)

 Enabled Disabled

Test Agent

[Using ODT **Change**](#)[Enter your message here](#)

...

[Run](#)



Amazon Bedrock



Amazon Bedrock > Agents > DataAnalysisAssistant > Agent builder: DataAnalysisAssist...

Getting started

Overview

Providers

Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

Image / Video

Builder tools

Agents

Flows

Knowledge Bases

Prompt Management

Safeguards

Guardrails

Watermark detection

Inference and Assessment

Agent builder [Info](#)

Manual

Assistant

Test

Prepare

Save

Close and exit

Agent details

Agent name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

AI Agent with Code Interpreter enabled for Data analysis

The description can have up to 200 characters.

Agent resource role

 Create and use a new service role Use an existing service role

AmazonBedrockExecutionRoleForAgents_IXFB4K5V4SC

Select model



Claude 3 Sonnet v1



On-demand

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

Test Agent

Using ODT [Change](#)

Amazon Bedrock 

Getting started

[Overview](#)[Providers](#)

Foundation models

[Model catalog !\[\]\(806e4b9747ea092b5900690a40846ea0_img.jpg\)](#)[Marketplace deployments !\[\]\(990e22d08bda883bbe98bf4051f3341d_img.jpg\)](#)[Custom models \(fine-tuning, dist...\)](#)[Imported models](#)[Prompt Routers !\[\]\(65eefab5d15eb350efd214b4562a9ec2_img.jpg\)](#)

Playgrounds

[Chat / Text](#)[Image / Video](#)

Builder tools

[Agents](#)[Flows](#)[Knowledge Bases](#)[Prompt Management](#)

Safeguards

[Guardrails](#)[Watermark detection](#)

Inference and Assessment

✓ Agent: DataAnalysisAssistant was successfully updated. Prepare the agent to keep its details up to date. 

[Amazon Bedrock](#) > [Agents](#) > [DataAnalysisAssistant](#) > Agent builder: DataAnalysisAssist...

Agent builder [Manual](#)[Assistant](#)[Test](#)[Prepare](#)[Save and exit](#)

Agent details

Agent name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

AI Agent with Code Interpreter enabled for Data analysis

The description can have up to 200 characters.

Agent resource role

 Create and use a new service role Use an existing service role

AmazonBedrockExecutionRoleForAgents_IXFB4K5V4SC

Test Agent

Using ODT [Change](#)

ⓘ Prepare the Agent to test the latest changes.

[Prepare](#)

Enter your message here

...

[Run](#)Claude 3 Sonnet v1  

On-demand

Amazon Bedrock



✓ Agent: DataAnalysisAssistant was successfully prepared.



Getting started

Overview

Providers

Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

Image / Video

Builder tools

Agents

Flows

Knowledge Bases

Prompt Management

Safeguards

Guardrails

Watermark detection

Inference and Assessment

[Amazon Bedrock](#) > [Agents](#) > [DataAnalysisAssistant](#) > Agent builder: DataAnalysisAssist...

Agent builder Info

Manual

Assistant

Test

Prepare

Save

Save and exit

Agent details

Agent name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

AI Agent with Code Interpreter enabled for Data analysis

The description can have up to 200 characters.

Agent resource role

- Create and use a new service role
- Use an existing service role

arn:aws:iam::675799275846:role/service-role/AmazonBedrockExecutionRoleFor...

Test Agent

Using ODT [Change](#)

Enter your message here

...

Run

bedrock-workshop-dryrun-20250319



Amazon Bedrock Workshop

▶ Prerequisites

Prompt Engineering

Text Generation

Knowledge Bases and RAG

Model Customization

Image and Video Generation

Applications

▼ Agents

[Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation](#)

Multi-Step Bedrock Agent using

▼ AWS account access

[Open AWS console \(us-west-2\)](#) [Get AWS CLI credentials](#)

▼ Content preferences

Language

English

Exit event

8. Test the Agent

- Once prepared, use the test window (on the right side panel) to interact with your agent
- Try the following prompt to test the code interpreter:

Using the customer satisfaction data provided below, create a bar graph showing the average overall rating for each hotel (H001, H002, H003).

```
date,hotel_id,overall_rating,room_cleanliness,staff_service,amenities,value_for_money,location
2023-01-05,H001,4.2,4.5,4.0,3.8,4.1,4.5
2023-01-12,H002,3.8,4.0,3.5,3.9,3.7,4.2
2023-01-20,H003,4.5,4.7,4.6,4.3,4.2,4.8
2023-02-03,H001,3.9,4.2,3.7,3.6,3.8,4.4
2023-02-15,H002,4.1,4.3,4.0,4.1,3.9,4.3
2023-02-28,H003,4.3,4.5,4.4,4.2,4.0,4.6
2023-03-10,H001,4.0,4.3,3.8,3.7,3.9,4.5
2023-03-22,H002,3.7,3.9,3.6,3.8,3.5,4.1
2023-04-05,H003,4.4,4.6,4.5,4.1,4.1,4.7
2023-04-18,H001,4.1,4.4,3.9,3.8,4.0,4.6
2023-05-01,H002,3.9,4.1,3.8,4.0,3.6,4.2
2023-05-15,H003,4.6,4.8,4.7,4.4,4.3,4.9
2023-06-02,H001,4.3,4.6,4.1,4.0,4.2,4.7
2023-06-20,H002,4.0,4.2,3.9,4.1,3.8,4.3
2023-06-30,H003,4.5,4.7,4.6,4.3,4.2,4.8
```

[Amazon Bedrock](#) > [Agents](#) > [DataAnalysisAssistant](#) > Agent builder: DataAnalysisAssistant

Agent builder Info

Manual

Assistant

Test

Prepare

Save

Save and exit

Agent details

Test Agent

Using ODT [Change](#)

Amazon Bedrock



Amazon Bedrock > Agents > DataAnalysisAssistant > Agent builder: DataAnalysisAssist...

Agent builder Info

[Manual](#)[Assistant](#)[Test](#)[Prepare](#)[Save](#)[Save and exit](#)

Agent details

Agent name

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

The description can have up to 200 characters.

Agent resource role

 Create and use a new service role Use an existing service role▼

Select model

 ⓘ ⓘ

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

Test Agent



Using ODT [Change](#)



Using the customer satisfaction data provided below, create a bar graph showing the average overall rating for each hotel

⋮

[Run](#)

Amazon Bedrock



Amazon Bedrock > Agents > DataAnalysisAssistant > Agent builder: DataAnalysisAssist...

Agent builder Info

Manual

Assistant

Test

Prepare

Save

Save and exit

Agent details

Agent name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

AI Agent with Code Interpreter enabled for Data analysis

The description can have up to 200 characters.

Agent resource role

 Create and use a new service role Use an existing service role

arn:aws:iam::675799275846:role/service-role/AmazonBedrockExecutionRoleFor...



Select model

Claude 3 Sonnet v1 i edit

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

Test Agent

i edit copy refresh nextUsing ODT Change

2023-02-
15,H002,4.1,4.3,4.0,4.1,3.9,4.3 2023-
02-28,H003,4.3,4.5,4.4,4.2,4.0,4.6
2023-03-
10,H001,4.0,4.3,3.8,3.7,3.9,4.5 2023-
03-22,H002,3.7,3.9,3.6,3.8,3.5,4.1
2023-04-
05,H003,4.4,4.6,4.5,4.1,4.1,4.7 2023-
04-18,H001,4.1,4.4,3.9,3.8,4.0,4.6
2023-05-
01,H002,3.9,4.1,3.8,4.0,3.6,4.2 2023-
05-15,H003,4.6,4.8,4.7,4.4,4.3,4.9
2023-06-
02,H001,4.3,4.6,4.1,4.0,4.2,4.7 2023-
06-20,H002,4.0,4.2,3.9,4.1,3.8,4.3
2023-06-
30,H003,4.5,4.7,4.6,4.3,4.2,4.8

Enter your message here

...

Run



Amazon Bedrock



Amazon Bedrock > Agents > DataAnalysisAssistant > Agent builder: DataAnalysisAssist...

Agent builder Info

Manual

Assistant

Test

Prepare

Save

Save and exit

Agent details

Agent name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

AI Agent with Code Interpreter enabled for Data analysis

The description can have up to 200 characters.

Agent resource role

 Create and use a new service role Use an existing service role

arn:aws:iam::675799275846:role/service-role/AmazonBedrockExecutionRoleFor...

Test Agent

Using ODT [Change](#)

2023-06-

30,H003,4.5,4.7,4.6,4.3,4.2,4.8



The bar graph showing the average overall rating for each hotel (H001, H002, H003) has been created. The x-axis displays the hotel IDs and the y-axis shows the average overall rating score. The graph has been saved to the file 'avg_ratings_plot.png'.

[avg_ratings_plot.png](#) [avg_ratings_plot.png](#)

[Show trace >](#)

Enter your message here

...

Run

 Claude 3 Sonnet v1 i edit

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

AWS Services Search WSParticipantRole/Participant @ 6757-9927-5846 ▼

Support

Amazon Bedrock <

Getting started

- Overview
- Providers

Foundation models

- Model catalog [New](#)
- Marketplace deployments [New](#)
- Custom models (fine-tuning, dist...)
- Imported models
- Prompt Routers [Preview](#)

Playgrounds

- Chat / Text
- Image / Video

Builder tools

- Agents
- Flows
- Knowledge Bases
- Prompt Management

Safeguards

- Guardrails
- Watermark detection

Inference and Assessment

Avg ratings plot (1).png

Average Overall Rating by Hotel

Hotel ID	Average Overall Rating
H001	~4.1
H002	~3.8
H003	~4.3

Using ODT Change

023-06-0,H003,4.5,4.7,4.6,4.3,4.2,4.8

The bar graph showing the average overall rating for each hotel (H001, H002, H003) has been created. The x-axis displays the hotel IDs and the y-axis shows the average overall rating score. The graph has been saved to the file 'avg_ratings_plot.png'.

avg_ratings_plot.png avg_ratings_plot.png

Show trace >

Claude 3 Sonnet v1 ⓘ ⚙️

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

Run

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Amazon Bedrock 

Getting started

[Overview](#)[Providers](#)

Foundation models

[Model catalog !\[\]\(3537deb254e618466019a43401d47143_img.jpg\)](#)[Marketplace deployments !\[\]\(4f3f00ab74b6d0936aa2be0ab3694b50_img.jpg\)](#)[Custom models \(fine-tuning, dist...\)](#)[Imported models](#)[Prompt Routers !\[\]\(5008dae5883a92f685eba95f082103ed_img.jpg\)](#)

Playgrounds

[Chat / Text](#)[Image / Video](#)

Builder tools

[Agents](#)[Flows](#)[Knowledge Bases](#)[Prompt Management](#)

Safeguards

[Guardrails](#)[Watermark detection](#)

Inference and Assessment

Test Agent

Using ODT [Change](#)

Using the customer satisfaction data provided below, create a bar graph showing the average overall rating for each hotel (H001, H002, H003). The graph should have the hotel IDs on the x-axis and the average overall rating on the y-axis.

date,hotel_id,overall_rating,room_cleanliness, staff_service,amenities,value_for_money,location
2023-01-05,H001,4.2,4.5,4.0,3.8,4.1,4.5
2023-01-12,H002,3.8,4.0,3.5,3.9,3.7,4.2
2023-01-20,H003,4.5,4.7,4.6,4.3,4.2,4.8
2023-02-03,H001,3.9,4.2,3.7,3.6,3.8,4.4
2023-02-15,H002,4.1,4.3,4.0,4.1,3.9,4.3
2023-02-28,H003,4.3,4.5,4.4,4.2,4.0,4.6
2023-03-10,H001,4.0,4.3,3.8,3.7,3.9,4.5
2023-03-22,H002,3.7,3.9,3.6,3.8,3.5,4.1
2023-04-05,H003,4.4,4.6,4.5,4.1,4.1,4.7
2023-04-18,H001,4.1,4.4,3.9,3.8,4.0,4.6
2023-05-01,H002,3.9,4.1,3.8,4.0,3.6,4.2
2023-05-15,H003,4.6,4.8,4.7,4.4,4.3,4.9
2023-06-02,H001,4.3,4.6,4.1,4.0,4.2,4.7
2023-06-20,H002,4.0,4.3,3.9,4.1,3.7,4.7

Enter your message here

...



Trace

Session summaries

Trace (3)

[Copy all](#)Request ID 20d2c2f1-af86-425d-b9b6-36832a3eaa29 

Routing Trace

Orchestration and Knowledge Base



Orchestration and Knowledge Base

After running a prompt in the test window, this orchestration trace allows you to explore the trace steps to understand the linear chain of thought used by the agent's orchestration prompt component. If a knowledge base was invoked, the trace also allows you to see how the results from the knowledge base were summarized to generate an observation that is used for for orchestrating the next step or generating the response.

▶ Trace step 1

[Show config](#)

▶ Trace step 2

[Show config](#)

▶ Trace step 3

[Show config](#)



Amazon Bedrock Workshop

▶ Prerequisites

Prompt Engineering

Text Generation

Knowledge Bases and RAG

Model Customization

Image and Video Generation

Applications

▼ Agents

[Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation](#)

Multi-Step Bedrock Agent using Jupyter Notebooks

Open Source With Bedrock

▼ AWS account access

[Open AWS console \(us-west-2\)](#)[Get AWS CLI credentials](#)

▼ Content preferences

Language

English

Exit event

H001

H002
Hotel ID

H003

9. Additional prompt testing

- Perform testing with below prompts for additional analysis and visualizations Using the customer satisfaction data provided above, perform the following tasks:
 1. Create a bar graph showing the average overall rating for each hotel (H001, H002, H003). The graph should have the hotel IDs on the x-axis and the average overall rating on the y-axis.
 2. Generate a heatmap that displays the correlation between different satisfaction factors (room_cleanliness, staff_service, amenities, value_for_money, location) and the overall_rating. 
 3. Calculate and print the average scores for each satisfaction factor across all hotels.
 4. Identify the hotel with the highest average overall rating and provide its average scores for each satisfaction factor.
 5. Create a line graph showing the trend of overall ratings over time for each hotel.
 6. Ensure that all graphs are clearly labeled and easy to read. After generating the visualizations and performing the analysis, provide a brief interpretation of the results, highlighting key findings and any notable trends or patterns in the data.

Conclusion

By following these steps, you've created an AI agent capable of performing data analysis tasks using the Code Interpreter tool directly through the Amazon Bedrock console. This agent can help with various data processing, analysis, and visualization tasks, making it a powerful tool for data-driven decision making.

[Previous](#)[Next](#)

Amazon Bedrock



Amazon Bedrock > Agents > DataAnalysisAssistant > Agent builder: DataAnalysisAssist...

Agent builder Info

[Manual](#)[Assistant](#)[Test](#)[Prepare](#)[Save](#)[Save and exit](#)

Agent details

Agent name

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

The description can have up to 200 characters.

Agent resource role

 Create and use a new service role Use an existing service role▼

Select model

**Claude 3 Sonnet v1**

Instructions for the Agent

Provide clear and specific instructions for the task the Agent will perform. You can also provide certain style and tone.

Test Agent



Using ODT [Change](#)

axis and the average overall rating on the y-axis.



The bar graph showing the average overall rating for each hotel (H001, H002, H003) has been created. The hotel IDs are on the x-axis and the average overall rating is on the y-axis. The graph has been saved to the file 'avg_ratings_plot.png'.

[avg_ratings_plot.png](#) [avg_ratings_plot.png](#)



[Show trace](#) >

Generate a heatmap that displays the correlation between different satisfaction factors (room_cleanliness, staff_service, amenities, value_for_money, location) and

[Run](#)

Amazon Bedrock



Amazon Bedrock > Agents > DataAnalysisAssistant > Agent builder: DataAnalysisAssist...

Agent builder Info

Manual

Assistant

Test

Prepare

Save

Save and exit

Agent details

Agent name

DataAnalysisAssistant

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 100 characters.

Agent description - optional

AI Agent with Code Interpreter enabled for Data analysis

The description can have up to 200 characters.

Agent resource role

 Create and use a new service role Use an existing service role

arn:aws:iam::675799275846:role/service-role/AmazonBedrockExecutionRoleFor...

Test Agent

Using ODT [Change](#)

The heatmap displaying the correlation between the different satisfaction factors and the overall_rating has been generated. The heatmap uses a color scale to show the correlation coefficients, with red indicating positive correlation and blue indicating negative correlation. The darker the color, the stronger the correlation. The heatmap has been saved to the file 'correlation_heatmap.png'.

[correlation_heatmap.png](#) [correlation_heatmap.png](#)

[Show trace >](#)

Enter your message here

...

Run

Getting started

Overview

Providers

Foundation models

Model catalog [New](#)Marketplace deployments [New](#)

Custom models (fine-tuning, dist...)

Imported models

Prompt Routers [Preview](#)

Playgrounds

Chat / Text

Image / Video

Builder tools

Agents

Flows

Knowledge Bases

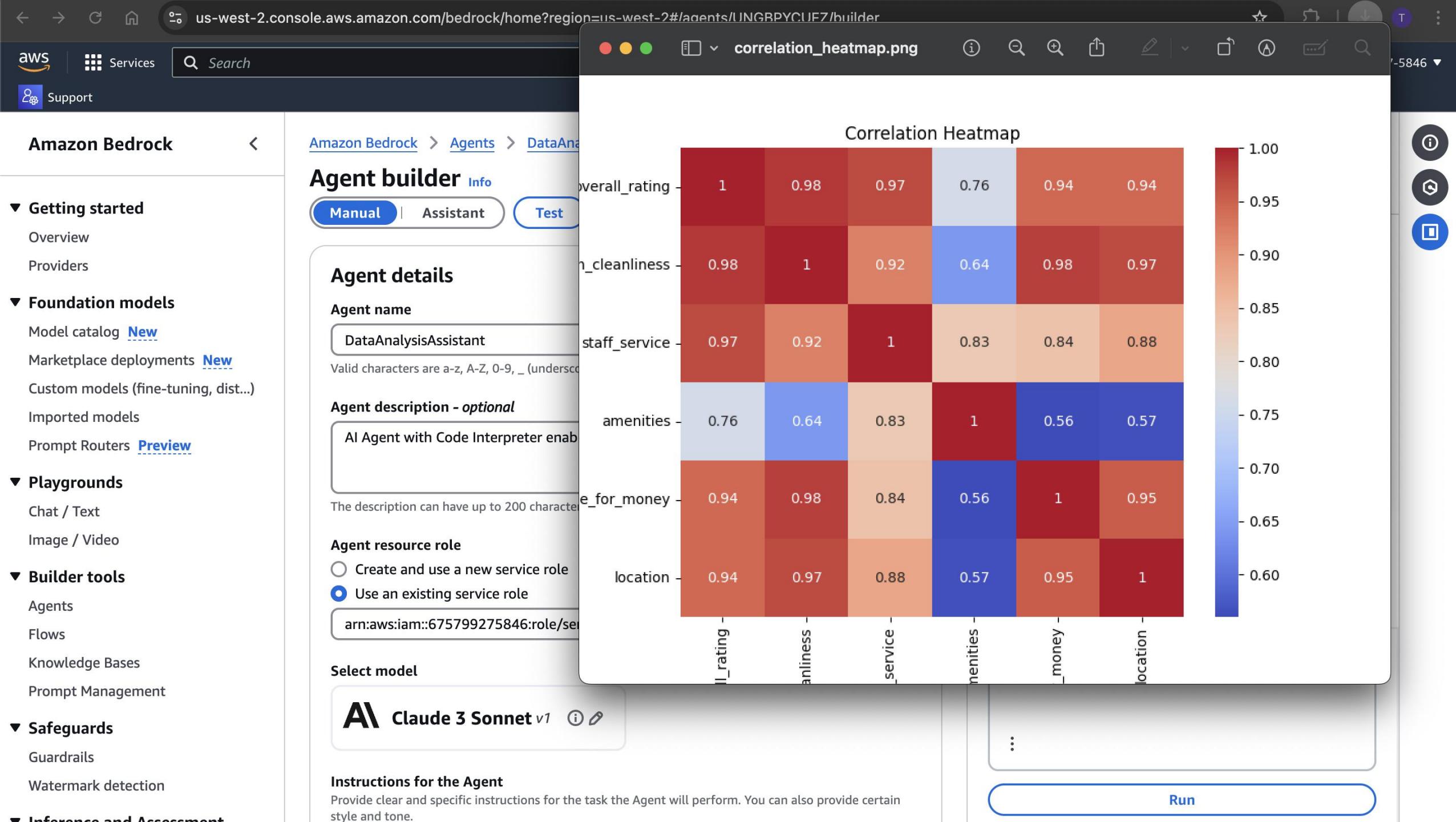
Prompt Management

Safeguards

Guardrails

Watermark detection

Inference and Assessment



Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- **Validating Bedrock agent**
- Exploring Bedrock fine-tune
- Q & A

Amazon Bedrock Workshop

- ▶ Prerequisites
- Prompt Engineering
- Text Generation
- Knowledge Bases and RAG
- Model Customization
- Image and Video Generation
- Applications
- ▼ Agents
 - Console Walkthrough of Configuring Bedrock Agent using for Code Interpretation
 - Multi-Step Bedrock Agent using Jupyter Notebooks**
 - Open Source With Bedrock

▶ AWS account access

▼ Content preferences

Language

English

Exit event



Multi-Step Bedrock Agent using Jupyter Notebooks

Overview

Generative AI applications often need to execute multistep tasks across company systems and data sources. Agents for Bedrock allows you to automatically orchestrate and analyze the task and break it down into the correct logical sequence using the FM's reasoning abilities. Agents automatically call the necessary APIs to transact with the company systems and processes to fulfill the request, determining along the way if they can proceed or if they need to gather more information. Using encryption in transit and at rest and IAM roles, agents provide secure access to enterprise data and APIs. With the Action Groups integration with AWS Lambda functions, agents lets you choose the implementation language for your API's connection. Thanks to the fully managed infrastructure provided, you don't have to worry about provisioning or managing infrastructure for your agent.

Target Audience

- Software Developers, Data Scientists, Solutions Architects and anyone else building Generative AI applications with access to internal APIs or Knowledge Bases

Challenges

Because agents architecture may include external components, debugging errors could be complex. This workshop will guide you through the debugging process of the most common challenges.

Sub-patterns

During this section, you will cover four modules:

+ + ↑ ↓ ✖ ✖

amazon-bedrock-workshop / 05_Agents /

Name	Modified
images	11m ago
kb_documents	11m ago
00_inline_agents.ipynb	11m ago
01_create_agent.ipynb	11m ago
02_associate_knowled...	5m ago
03_invoke_agent.ipynb	11m ago
04_clean_up_agent_r...	11m ago
agent.py	11m ago
knowledge_base.py	11m ago
README.md	11m ago
requirements.txt	11m ago

Launcher 01_create_agent.ipynb +

Markdown git

Notebook Cluster No Kernel

Agents for Amazon Bedrock - create agent

This notebook provides sample code for building an Agent for Amazon Bedrock that has an Action Group attached to it.

Use Case

We will create a restaurant assistant that allows customers to create, delete or get reservation information. The architecture looks as following:

The diagram illustrates the architecture for creating a restaurant assistant agent. It shows a flow from a customer (represented by a person icon) interacting with an Action Group Lambda function (represented by an orange Lambda icon). The Lambda function is connected to an Amazon DynamoDB table (represented by a purple icon). Arrows indicate the flow of data: 'customer' feeds into the Lambda function, and the Lambda function updates the DynamoDB table via 'feeds' and 'update' operations.

Select Kernel

Select kernel for: "01_create_agent.ipynb"

Python 3 (ipykernel)

Always start the preferred kernel No Kernel Select

Notebook Walk-through

In this notebook we will:

- Choose our Agent's underline foundation model
- Create a dynamoDB table to store the reservation details
- Create a lambda function that handles the restaurant bookings
- Create an agent
- Create an action group and associate it with the agent
- Test the agent invocation

Would you like to get notified about official Jupyter news?

[Open privacy policy](#) [Yes](#)

+ + ↑ ↶ ⬇ FILTER

amazon-bedrock-workshop / 05_Agents /

Name	Modified
images	11m ago
kb_documents	11m ago
00_inline_agents.ipynb	11m ago
01_create_agent.ipynb	11m ago
02_associate_knowled...	5m ago
03_invoke_agent.ipynb	11m ago
04_clean_up_agent_r...	11m ago
agent.py	11m ago
knowledge_base.py	11m ago
README.md	11m ago
requirements.txt	11m ago

Launcher x 01_create_agent.ipynb • +

Notebook Cluster Python 3 (ipykernel) ⚙

Agents for Amazon Bedrock - create agent

This notebook provides sample code for building an Agent for Amazon Bedrock that has an Action Group attached to it.

Use Case

We will create a restaurant assistant that allows customers to create, delete or get reservation information. The architecture looks as following:

```
graph LR; Customer((Customer)) -- "User question" --> RestaurantAgent[Restaurant Agent]; RestaurantAgent -- "LLM response" --> Customer; RestaurantAgent -- "task execution" --> ActionGroup[Action Group]; ActionGroup -- "invoke" --> LambdaFunction[Lambda function]; LambdaFunction -- "respond" --> ActionGroup; LambdaFunction -- "update" --> DynamoDB[Amazon DynamoDB]; ActionGroup -- "feeds" --> LambdaFunction;
```

Notebook Walk-through

In this notebook we will:

- Choose our Agent's underline foundation model
- Create a dynamoDB table to store the reservation details
- Create a lambda function that handles the restaurant bookings
- Create an agent
- Create an action group and associate it with the agent
- Test the agent invocation

Would you like to get notified about official Jupyter news? Open privacy policy Yes

Launcher x 01_create_agent.ipynb +

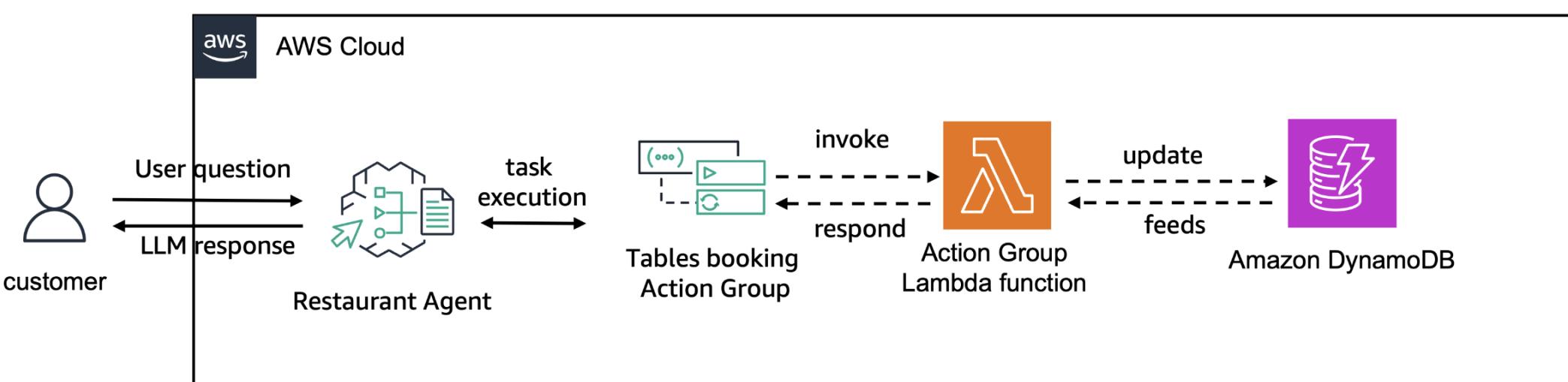
Notebook Cluster Python 3 (ipykernel)

Agents for Amazon Bedrock - create agent

This notebook provides sample code for building an Agent for Amazon Bedrock that has an Action Group attached to it.

Use Case

We will create a restaurant assistant that allows customers to create, delete or get reservation information. The architecture looks as following:



Notebook Walk-through

In this notebook we will:

- Choose our Agent's underline foundation model
- Create a dynamoDB table to store the reservation details
- Create a lambda function that handles the restaurant bookings



File Edit View Run Kernel Git Tabs Settings Help sagemakeruser / amazon-bedrock-workshop-space

Launcher + X -

Run Selected Cell ↗ ⇌ Run Selected Cell and Insert Below ↘ ⇌ Run Selected Cell 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...

Notebook Cluster Python 3 (ipykernel) ⚡

Create agent

Amazon Bedrock that has an Action Group attached to it.

customer → User question → Restaurant Agent → task execution → Action Group Lambda function → invoke → (Tables booking, ...)

Action Group Lambda function → respond → Restaurant Agent → LLM response → customer

Action Group Lambda function → update feeds → Amazon DynamoDB

Notebook Walk-through

In this notebook we will:

- Choose our Agent's underline foundation model
- Create a dynamoDB table to store the reservation details
- Create a lambda function that handles the restaurant bookings

Simple 0 \$ 2 main Python 3 (ipykernel) | Busy Initialized (additional servers needed) Instance MEM 50% ✓ Amazon Q Cookie Preferences Mode: Command Ln 1, Col 1

Setup

Before running the rest of this notebook, you'll need to run the cells below to ensure necessary libraries are installed

```
[1]: !pip install --upgrade -q -r requirements.txt
```

Let's now import the necessary libraries and initiate the required boto3 clients

```
[2]: import time
import boto3
import logging
import ipywidgets as widgets
import uuid

from agent import create_agent_role, create_lambda_role
from agent import create_dynamodb, create_lambda, invoke_agent_helper
```

```
[3]: #Clients
s3_client = boto3.client('s3')
sts_client = boto3.client('sts')
session = boto3.Session()
region = session.region_name
account_id = sts_client.get_caller_identity()["Account"]
bedrock_agent_client = boto3.client('bedrock-agent')
bedrock_agent_runtime_client = boto3.client('bedrock-agent-runtime')
logging.basicConfig(format='[%(asctime)s] %(process)s %(filename)s:%(lineno)d %(levelname)s - %(message)s', level=logging.INFO)
logger = logging.getLogger(__name__)
region, account_id
```

```
[3]: ('us-west-2', '364013687952')
```

Setting up Agent's information

We will now set the variables that define our agent:



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel) ○

Setting up Agent's information

We will now set the variables that define our agent:

- **agent_name**: provides the name of the agent to be created, in this case `booking-agent`
- **agent_description**: the description of the agent used to display the agents list on the console. This description is **not** part of the agent's prompts
- **agent_instruction**: the instructions of what the agent should and should not do. This description is part of the agent's prompt and is used during the agent's invocation
- **agent_action_group_name**: the action group name used on the definition of the agent's action, in this case `TableBookingsActionGroup`.
- **agent_action_group_description**: the description of the action group name used on the UI to list the action groups. This description is **not** used by the agent's prompts

```
[4]: suffix = f'{region}-{account_id}'
agent_name = 'booking-agent'
agent_bedrock_allow_policy_name = f'{agent_name}-ba'
agent_role_name = f'AmazonBedrockExecutionRoleForAgents_{agent_name}'

agent_description = "Agent in charge of a restaurants table bookings"
agent_instruction = """
You are a restaurant agent, helping clients retrieve information from their booking,
create a new booking or delete an existing booking
"""

agent_action_group_description = """
Actions for getting table booking information, create a new booking or delete an existing booking"""

agent_action_group_name = "TableBookingsActionGroup"
```

Select Foundation Model

Use this dropdown menu to select the underline model of your agent. You can find more information about the supported foundation models [here](#)

```
[5]: agent.foundation_model_selector = widgets.Dropdown(
    options=[('Claude 3 Sonnet', 'anthropic.claude-3-sonnet-20240229-v1:0'),
             ('Claude 3 Haiku', 'anthropic.claude-3-haiku-20240307-v1:0')]
```



bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge.ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.p X Restaurant_Childrens_Men X +

File Notebook Cluster Python 3 (ipykernel)

Select Foundation Model

Use this dropdown menu to select the underline model of your agent. You can find more information about the supported foundation models [here](#)

```
[5]: agent_foundation_model_selector = widgets.Dropdown(  
    options=[  
        ('Claude 3 Sonnet', 'anthropic.claude-3-sonnet-20240229-v1:0'),  
        ('Claude 3 Haiku', 'anthropic.claude-3-haiku-20240307-v1:0')  
    ],  
    value='anthropic.claude-3-sonnet-20240229-v1:0',  
    description='FM:',  
    disabled=False,  
)  
agent_foundation_model_selector
```

```
[5]: FM: Claude 3 Sonnet ▾
```

Let's confirm that the model has been selected correctly

```
[6]: agent_foundation_model = agent_foundation_model_selector.value  
agent_foundation_model  
[6]: 'anthropic.claude-3-sonnet-20240229-v1:0'
```

Creating DynamoDB table

Let's now create an [Amazon DynamoDB](#) table called `restaurant_bookings`. This table will store information about the reservations, including a `booking_id`, reservation `date`, the `name` of the person doing the reservation, the `hour` of the reservation and the number of guests as `num_guests`. To do so, we use the `create_dynamodb` function from the `agent.py` file. This function will support the creation of the table and its requirements (IAM roles and permissions).

```
[7]: table_name = 'restaurant_bookings'  
create_dynamodb(table_name)  
  
Creating table restaurant_bookings...  
Table restaurant_bookings created successfully!
```



Creating DynamoDB table

Let's now create an [Amazon DynamoDB](#) table called `restaurant_bookings`. This table will store information about the reservations, including a `booking_id`, reservation `date`, the `name` of the person doing the reservation, the `hour` of the reservation and the number of guests as `num_guests`. To do so, we use the `create_dynamodb` function from the `agent.py` file. This function will support the creation of the table and its requirements (IAM roles and permissions).

```
[7]: table_name = 'restaurant_bookings'  
create_dynamodb(table_name)
```

```
Creating table restaurant_bookings...  
Table restaurant_bookings created successfully!
```

Creating Lambda Function

Next we will create the [AWS Lambda](#) function that executes the actions for our agent. This lambda function will have 3 actions:

- `get_booking_details(booking_id)` : returns the details of a booking based on the booking id
- `create_booking(date, name, hour, num_guests)` : creates a new booking for the restaurant
- `delete_booking(booking_id)` : deletes an existent booking based on the booking id

The `lambda_handler` receives the `event` from the agent and the `event` contains information about the `function` to be executed and its `parameters`.

A `functionResponse` is returned by the lambda function with the response body having a `TEXT` field.

You can find more information on how to set your agent lambda function [here](#).

Let's first write the code of the lambda function to the `lambda_function.py` file

```
[8]: %%writefile lambda_function.py  
import json  
import uuid  
import boto3  
  
dynamodb = boto3.resource('dynamodb')  
table = dynamodb.Table('restaurant_bookings')
```



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel) ○

Creating Lambda Function

Next we will create the [AWS Lambda](#) function that executes the actions for our agent. This lambda function will have 3 actions:

- `get_booking_details(booking_id)` : returns the details of a booking based on the booking id
- `create_booking(date, name, hour, num_guests)` : creates a new booking for the restaurant
- `delete_booking(booking_id)` : deletes an existent booking based on the booking id

The `lambda_handler` receives the `event` from the agent and the `event` contains information about the `function` to be executed and its `parameters`.

A `functionResponse` is returned by the lambda function with the response body having a `TEXT` field.

You can find more information on how to set your agent lambda function [here](#).

Let's first write the code of the lambda function to the `lambda_function.py` file

```
[8]: %%writefile lambda_function.py
import json
import uuid
import boto3

dynamodb = boto3.resource('dynamodb')
table = dynamodb.Table('restaurant_bookings')

def get_named_parameter(event, name):
    """
    Get a parameter from the lambda event
    """
    return next(item for item in event['parameters'] if item['name'] == name)['value']

def get_booking_details(booking_id):
    """
    Retrieve details of a restaurant booking
    Args:
        booking_id (str): The ID of the booking to retrieve.
    Returns:
        dict: The details of the booking.
    """
    response = table.get_item(
        Key={'id': str(uuid.uuid4()), 'booking_id': booking_id}
    )
    return response['Item']

def create_booking(date, name, hour, num_guests):
    """
    Create a new booking for the restaurant.
    Args:
        date (str): The date of the booking.
        name (str): The name of the guest.
        hour (int): The hour of the booking.
        num_guests (int): The number of guests.
    Returns:
        dict: The details of the newly created booking.
    """
    booking_id = str(uuid.uuid4())
    item = {
        'id': str(uuid.uuid4()),
        'booking_id': booking_id,
        'date': date,
        'name': name,
        'hour': hour,
        'num_guests': num_guests
    }
    table.put_item(Item=item)
    return item

def delete_booking(booking_id):
    """
    Delete an existing booking.
    Args:
        booking_id (str): The ID of the booking to delete.
    Returns:
        None
    """
    table.delete_item(
        Key={'id': str(uuid.uuid4()), 'booking_id': booking_id}
    )
```



bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge.ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.p X Restaurant_Childrens_Men X +

File + X Markdown git

Notebook Cluster Python 3 (ipykernel) ⚙️

```
[8]: %writefile lambda_function.py
import json
import uuid
import boto3

dynamodb = boto3.resource('dynamodb')
table = dynamodb.Table('restaurant_bookings')

def get_named_parameter(event, name):
    """
    Get a parameter from the lambda event
    """
    return next(item for item in event['parameters'] if item['name'] == name)['value']

def get_booking_details(booking_id):
    """
    Retrieve details of a restaurant booking

    Args:
        booking_id (string): The ID of the booking to retrieve
    """
    try:
        response = table.get_item(Key={'booking_id': booking_id})
        if 'Item' in response:
            return response['Item']
        else:
            return {'message': f'No booking found with ID {booking_id}'}
    except Exception as e:
        return {'error': str(e)}

def create_booking(date, name, hour, num_guests):
    """
    Create a new restaurant booking

    Args:
        date (string): The date of the booking
        name (string): The name of the guest
        hour (int): The hour of the booking
        num_guests (int): The number of guests
    """
    # Create a new item in the dynamoDB table
    item = {
        'date': date,
        'name': name,
        'hour': hour,
        'num_guests': num_guests
    }
```



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel)

```
def create_booking(date, name, hour, num_guests):
    """
    Create a new restaurant booking

    Args:
        date (string): The date of the booking
        name (string): Name to identify your reservation
        hour (string): The hour of the booking
        num_guests (integer): The number of guests for the booking
    """
    try:
        booking_id = str(uuid.uuid4())[:8]
        table.put_item(
            Item={
                'booking_id': booking_id,
                'date': date,
                'name': name,
                'hour': hour,
                'num_guests': num_guests
            }
        )
        return {'booking_id': booking_id}
    except Exception as e:
        return {'error': str(e)}

def delete_booking(booking_id):
    """
    Delete an existing restaurant booking

    Args:
        booking_id (str): The ID of the booking to delete
    """
    try:
        response = table.delete_item(Key={'booking_id': booking_id})
        if response['ResponseMetadata']['HTTPStatusCode'] == 200:
            return {'message': f'Booking with ID {booking_id} deleted successfully'}
    except Exception as e:
        return {'error': str(e)}
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel)

```
def delete_booking(booking_id):
    """
    Delete an existing restaurant booking

    Args:
        booking_id (str): The ID of the booking to delete
    """

    try:
        response = table.delete_item(Key={'booking_id': booking_id})
        if response['ResponseMetadata']['HTTPStatusCode'] == 200:
            return {'message': f'Booking with ID {booking_id} deleted successfully'}
        else:
            return {'message': f'Failed to delete booking with ID {booking_id}'}
    except Exception as e:
        return {'error': str(e)}

def lambda_handler(event, context):
    # get the action group used during the invocation of the lambda function
    actionGroup = event.get('actionGroup', '')

    # name of the function that should be invoked
    function = event.get('function', '')

    # parameters to invoke function with
    parameters = event.get('parameters', [])

    if function == 'get_booking_details':
        booking_id = get_named_parameter(event, "booking_id")
        if booking_id:
            response = str(get_booking_details(booking_id))
            responseBody = {'TEXT': {'body': json.dumps(response)}}
        else:
            responseBody = {'TEXT': {'body': 'Missing booking_id parameter'}}

    elif function == 'create_booking':
        date = get_named_parameter(event, "date")
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel) ○

```
def lambda_handler(event, context):
    # get the action group used during the invocation of the lambda function
    actionGroup = event.get('actionGroup', '')

    # name of the function that should be invoked
    function = event.get('function', '')

    # parameters to invoke function with
    parameters = event.get('parameters', [])

    if function == 'get_booking_details':
        booking_id = get_named_parameter(event, "booking_id")
        if booking_id:
            response = str(get_booking_details(booking_id))
            responseBody = {'TEXT': {'body': json.dumps(response)}}
        else:
            responseBody = {'TEXT': {'body': 'Missing booking_id parameter'}}

    elif function == 'create_booking':
        date = get_named_parameter(event, "date")
        name = get_named_parameter(event, "name")
        hour = get_named_parameter(event, "hour")
        num_guests = get_named_parameter(event, "num_guests")

        if date and hour and num_guests:
            response = str(create_booking(date, name, hour, num_guests))
            responseBody = {'TEXT': {'body': json.dumps(response)}}
        else:
            responseBody = {'TEXT': {'body': 'Missing required parameters'}}

    elif function == 'delete_booking':
        booking_id = get_named_parameter(event, "booking_id")
        if booking_id:
            response = str(delete_booking(booking_id))
            responseBody = {'TEXT': {'body': json.dumps(response)}}
        else:
            responseBody = {'TEXT': {'body': 'Missing booking_id parameter'}}
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel) ⚙️

Next we create the function requirements for IAM role and policies using the support function `create_lambda_role` and create the lambda using the support function `create_lambda` both from the `agent.py` file

```
[9]: lambda_iam_role = create_lambda_role(agent_name, table_name)

[10]: lambda_function_name = f'{agent_name}-lambda'

[11]: lambda_function = create_lambda(lambda_function_name, lambda_iam_role)
```

Creating Agent

Now that we have created the dynamoDB table and lambda function, let's create our Agent.

To do so, we first need to create an agent role and its required policies. Let's do so using the `create_agent_role` function from the `agent.py` file.

```
[12]: agent_role = create_agent_role(agent_name, agent.foundation_model)

[13]: agent_role

[13]: {'Role': {'Path': '/',
   'RoleName': 'AmazonBedrockExecutionRoleForAgents_booking-agent',
   'RoleId': 'AROAVJQHICSLFEQPG3TL',
   'Arn': 'arn:aws:iam::364013687952:role/AmazonBedrockExecutionRoleForAgents_booking-agent',
   'CreateDate': datetime.datetime(2025, 3, 19, 16, 24, 29, tzinfo=tzlocal()),
   'AssumeRolePolicyDocument': {'Version': '2012-10-17',
     'Statement': [{'Effect': 'Allow',
       'Principal': {'Service': 'bedrock.amazonaws.com'},
       'Action': 'sts:AssumeRole'}]}},
   'ResponseMetadata': {'RequestId': '413ee14e-ee8c-4171-a5ac-818cb5ad28f0',
   'HTTPStatusCode': 200,
   'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:24:29 GMT',
     'x-amzn-requestid': '413ee14e-ee8c-4171-a5ac-818cb5ad28f0',
     'content-type': 'text/xml',
     'content-length': '853'},
   'RetryAttempts': 0}}
```



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel) ○

Creating Agent

Now that we have created the dynamoDB table and lambda function, let's create our Agent.

To do so, we first need to create an agent role and its required policies. Let's do so using the `create_agent_role` function from the `agent.py` file.

```
[12]: agent_role = create_agent_role(agent_name, agent.foundation_model)

[13]: agent_role

[13]: {'Role': {'Path': '/',
   'RoleName': 'AmazonBedrockExecutionRoleForAgents_booking-agent',
   'RoleId': 'AROAVJQHICSLFEQPG3TL',
   'Arn': 'arn:aws:iam::364013687952:role/AmazonBedrockExecutionRoleForAgents_booking-agent',
   'CreateDate': datetime.datetime(2025, 3, 19, 16, 24, 29, tzinfo=tzlocal()),
   'AssumeRolePolicyDocument': {'Version': '2012-10-17',
      'Statement': [{{'Effect': 'Allow',
         'Principal': {'Service': 'bedrock.amazonaws.com'},
         'Action': 'sts:AssumeRole'}}]},
   'ResponseMetadata': {'RequestId': '413ee14e-ee8c-4171-a5ac-818cb5ad28f0',
   'HTTPStatusCode': 200,
   'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:24:29 GMT',
      'x-amzn-requestid': '413ee14e-ee8c-4171-a5ac-818cb5ad28f0',
      'content-type': 'text/xml',
      'content-length': '853'},
   'RetryAttempts': 0}}
```

With the Agent IAM role created, we can now use the boto3 function `create_agent` to create our agent.

On the agent creation, all you need to provide is the agent name, foundation model and instruction. We will associate an action group to the agent once it has been created

```
[14]: response = bedrock_agent_client.create_agent(
    agentName=agent_name,
    agentResourceRoleArn=agent_role['Role']['Arn'],
    description=agent_description,
    idleSessionTTLInSeconds=1800,
    foundationModel=agent.foundation_model,
```



bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.p X Restaurant_Childrens_Men X +

File + X D Markdown git

Notebook Cluster Python 3 (ipykernel) ⚙️

On the agent creation, all you need to provide is the agent name, foundation model and instruction. We will associate an action group to the agent once it has been created

```
[14]: response = bedrock_agent_client.create_agent(  
        agentName=agent_name,  
        agentResourceRoleArn=agent_role['Role']['Arn'],  
        description=agent_description,  
        idleSessionTTLInSeconds=1800,  
        foundationModel=agent_foundation_model,  
        instruction=agent_instruction,  
)  
response
```

```
[14]: {'ResponseMetadata': {'RequestId': '45c175d2-338e-4e34-8718-79517e3a8224',  
    'HTTPStatusCode': 202,  
    'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:24:40 GMT',  
        'content-type': 'application/json',  
        'content-length': '726',  
        'connection': 'keep-alive',  
        'x-amzn-requestid': '45c175d2-338e-4e34-8718-79517e3a8224',  
        'x-amz-apigw-id': 'HrpnVFtkPHcEg0g',  
        'x-amzn-trace-id': 'Root=1-67daefc8-0f76910f1e82692313590916'},  
    'RetryAttempts': 0},  
    'agent': {'agentArn': 'arn:aws:bedrock:us-west-2:364013687952:agent/VMG6HC5N8I',  
        'agentCollaboration': 'DISABLED',  
        'agentId': 'VMG6HC5N8I',  
        'agentName': 'booking-agent',  
        'agentResourceRoleArn': 'arn:aws:iam::364013687952:role/AmazonBedrockExecutionRoleForAgents_booking-agent',  
        'agentStatus': 'CREATING',  
        'createdAt': datetime.datetime(2025, 3, 19, 16, 24, 40, 274174, tzinfo=tzlocal()),  
        'description': 'Agent in charge of a restaurants table bookings',  
        'foundationModel': 'anthropic.claude-3-sonnet-20240229-v1:0',  
        'idleSessionTTLInSeconds': 1800,  
        'instruction': '\nYou are a restaurant agent, helping clients retrieve information from their booking,\ncreate a new booking or delete an existing booking\n\n',  
        'orchestrationType': 'DEFAULT',  
        'updatedAt': datetime.datetime(2025, 3, 19, 16, 24, 40, 274174, tzinfo=tzlocal())}}
```

Now that our agent has been created, we will retrieve the `agentId`. It will be used to associate the action group to the agent in our next step.



File Edit View Run Kernel Git Tabs Settings Help sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

[15]: agent_id = response['agent']['agentId']
print("The agent id is:",agent_id)

The agent id is: VMG6HC5N8I

▼ Create Agent Action Group

now that we have created the agent, let's create an [Action Group](#) and associate with the agent. The action group will allow our agent to execute the booking tasks. To do so, we will "inform" our agent about the existent functionalities using a [function schema](#) defined in [JSON](#) format.

The function schema requires the function `name`, `description` and `parameters` to be provided. Each parameter has a parameter name, description, type and a boolean flag indicating if the parameter is required.

Let's define the functions [JSON](#) as `agent_functions`

```
[16]: agent_functions = [  
    {  
        'name': 'get_booking_details',  
        'description': 'Retrieve details of a restaurant booking',  
        'parameters': {  
            "booking_id": {  
                "description": "The ID of the booking to retrieve",  
                "required": True,  
                "type": "string"  
            }  
        }  
    },  
    {  
        'name': 'create_booking',  
        'description': 'Create a new restaurant booking',  
        'parameters': {  
            "date": {  
                "description": "The date of the booking in the format YYYY-MM-DD",  
                "required": True,  
                "type": "string"  
            }  
        }  
    }]
```

A screenshot of a Jupyter Notebook interface. The top bar shows the URL: t3yfm3qipfvu992.studio.us-west-2.sagemaker.aws/jupyterlab/default/lab?. The top menu includes File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. The right side of the top bar shows the user sagemakeruser / amazon-bedrock-workshop-space. Below the menu, there are several tabs: bedrock_basics.ipynb, 01_create_agent.ipynb (selected), 02_associate_knowledge_ipynb, 03_invoke_agent.ipynb, Restaurant_Dinner_Menu.pdf, and Restaurant_Childrens_Men.pdf. The toolbar below the tabs includes icons for file operations like new, open, save, and copy, along with Markdown, git, and a plus sign. On the far right of the toolbar are Notebook, Cluster, Python 3 (ipykernel), and a settings gear icon.

▼ Create Agent Action Group

now that we have created the agent, let's create an [Action Group](#) and associate with the agent. The action group will allow our agent to execute the booking tasks. To do so, we will "inform" our agent about the existent functionalities using a [function schema](#) defined in [JSON](#) format.

The function schema requires the function `name`, `description` and `parameters` to be provided. Each parameter has a parameter name, description, type and a boolean flag indicating if the parameter is required.

Let's define the functions [JSON](#) as `agent_functions`

```
[16]: agent_functions = [
    {
        'name': 'get_booking_details',
        'description': 'Retrieve details of a restaurant booking',
        'parameters': {
            "booking_id": {
                "description": "The ID of the booking to retrieve",
                "required": True,
                "type": "string"
            }
        }
    },
    {
        'name': 'create_booking',
        'description': 'Create a new restaurant booking',
        'parameters': {
            "date": {
                "description": "The date of the booking in the format YYYY-MM-DD",
                "required": True,
                "type": "string"
            },
            "name": {
                "description": "Name to identify your reservation",
                "required": True,
                "type": "string"
            }
        }
    }
]
```



bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge.ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.p X Restaurant_Childrens_Men X +

File + X D C Markdown git

Notebook Cluster Python 3 (ipykernel) ⚙

```
{  
    'name': 'create_booking',  
    'description': 'Create a new restaurant booking',  
    'parameters': {  
        "date": {  
            "description": "The date of the booking in the format YYYY-MM-DD",  
            "required": True,  
            "type": "string"  
        },  
        "name": {  
            "description": "Name to identify your reservation",  
            "required": True,  
            "type": "string"  
        },  
        "hour": {  
            "description": "The hour of the booking in the format HH:MM",  
            "required": True,  
            "type": "string"  
        },  
        "num_guests": {  
            "description": "The number of guests for the booking",  
            "required": True,  
            "type": "integer"  
        }  
    },  
    {  
        'name': 'delete_booking',  
        'description': 'Delete an existing restaurant booking',  
        'parameters': {  
            "booking_id": {  
                "description": "The ID of the booking to delete",  
                "required": True,  
                "type": "string"  
            }  
        }  
    },  
},  
]  
]
```



bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge.ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.p X Restaurant_Childrens_Men X +

File + X D Markdown C git

Notebook Cluster Python 3 (ipykernel) ⚙

```
{  
    'name': 'delete_booking',  
    'description': 'Delete an existing restaurant booking',  
    'parameters': {  
        'booking_id': {  
            'description': "The ID of the booking to delete",  
            'required': True,  
            'type': "string"  
        }  
    },  
}
```

Now we can use the `create_agent_action_group` function from the boto3 SDK to create the action group

```
[17]: # Pause to make sure agent is created  
time.sleep(30)  
  
# Now, we can configure and create an action group here:  
agent_action_group_response = bedrock_agent_client.create_agent_action_group(  
    agentId=agent_id,  
    agentVersion='DRAFT',  
    actionGroupExecutor={  
        'lambda': lambda_function['FunctionArn']  
    },  
    actionGroupName=agent_action_group_name,  
    functionSchema={  
        'functions': agent_functions  
    },  
    description=agent_action_group_description  
)
```

```
[18]: agent_action_group_response
```

```
[18]: {'ResponseMetadata': {'RequestId': '20c7639c-2ec3-433b-a83f-73be916dece9',  
    'HTTPStatusCode': 200,  
    'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:25:10 GMT',  
        'content-type': 'application/json'}}
```



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel) ○

Now we can use the `create_agent_action_group` function from the boto3 SDK to create the action group

```
[17]: # Pause to make sure agent is created
time.sleep(30)

# Now, we can configure and create an action group here:
agent_action_group_response = bedrock_agent_client.create_agent_action_group(
    agentId=agent_id,
    agentVersion='DRAFT',
    actionGroupExecutor={
        'lambda': lambda_function['FunctionArn']
    },
    actionGroupName=agent_action_group_name,
    functionSchema={
        'functions': agent_functions
    },
    description=agent_action_group_description
)
```

```
[18]: agent_action_group_response
```

```
[18]: {'ResponseMetadata': {'RequestId': '20c7639c-2ec3-433b-a83f-73be916dece9',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:25:10 GMT',
    'content-type': 'application/json',
    'content-length': '1490',
    'connection': 'keep-alive',
    'x-amzn-requestid': '20c7639c-2ec3-433b-a83f-73be916dece9',
    'x-amz-apigw-id': 'HrpsFEG4PHcEdmg=',
    'x-amzn-trace-id': 'Root=1-67daefe6-6e466e1d33baa36f6886ad34'},
  'RetryAttempts': 0},
  'agentActionGroup': {'actionGroupExecutor': {'lambda': 'arn:aws:lambda:us-west-2:364013687952:function:booking-agent-lambda'},
    'actionGroupId': 'MPD4SARU7L',
    'actionGroupName': 'TableBookingsActionGroup',
    'actionGroupState': 'ENABLED',
    'agentId': 'VMG6HC5N8I',
    'agentVersion': 'DRAFT'}}
```



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge_ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.pdf X Restaurant_Childrens_Men.pdf X +

Notebook Cluster Python 3 (ipykernel) ⚙️

Allowing bedrock to invoke lambda function

The last requirement is to add the [resource-based policy](#) to allow bedrock to invoke the action group lambda function.

```
[19]: # Create allow to invoke permission on lambda
lambda_client = boto3.client('lambda')
try:
    response = lambda_client.add_permission(
        FunctionName=lambda_function_name,
        StatementId=f'allow_bedrock_{agent_id}',
        Action='lambda:InvokeFunction',
        Principal='bedrock.amazonaws.com',
        SourceArn=f"arn:aws:bedrock:{region}:{account_id}:agent/{agent_id}",
    )
    print(response)
except Exception as e:
    print(e)
```

```
{'ResponseMetadata': {'RequestId': '637cb201-518a-4fbb-bb8c-0251a8e959ef', 'HTTPStatusCode': 201, 'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:25:10 GMT', 'content-type': 'application/json', 'content-length': '359', 'connection': 'keep-alive', 'x-amzn-requestid': '637cb201-518a-4fbb-bb8c-0251a8e959ef'}, 'RetryAttempts': 0}, 'Statement': '{"Sid":"allow_bedrock_VMG6HC5N8I","Effect":"Allow","Principal":{"Service":"bedrock.amazonaws.com"},"Action":"lambda:InvokeFunction","Resource":"arn:aws:lambda:us-west-2:364013687952:function:booking-agent-lambda","Condition":{"ArnLike":{"AWS:SourceArn":"arn:aws:bedrock:us-west-2:364013687952:agent/VMG6HC5N8I"}}}'}
```

Preparing agent

Before invoking the agent we need to prepare it. Preparing your agent will package all its components, including the security configurations. It will bring the agent into a state where it can be tested in runtime. We will use the `prepare_agent` function from the `boto3` sdk to prepare our agent.

```
[20]: response = bedrock_agent_client.prepare_agent(
    agentId=agent_id
)
print(response)
# Pause to make sure agent is prepared
time.sleep(30)
```

```
{'ResponseMetadata': {'RequestId': 'ee10e01d-f9d4-4924-8cc0-31b62d502564', 'HTTPStatusCode': 202, 'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:25:11 GMT', 'content-type': 'application/json', 'content-length': '359', 'connection': 'keep-alive', 'x-amzn-requestid': 'ee10e01d-f9d4-4924-8cc0-31b62d502564'}, 'RetryAttempts': 0}}
```

bedrock_basics.ipynb X 01_create_agent.ipynb X 02_associate_knowledge.ipynb X 03_invoke_agent.ipynb X Restaurant_Dinner_Menu.p X Restaurant_Childrens_Men X +

File + X □ ▢ ▣ ▤ ▥ ▦ ▨ ▩ git

Notebook Cluster Python 3 (ipykernel) ⚙



Preparing agent

Before invoking the agent we need to prepare it. Preparing your agent will package all its components, including the security configurations. It will bring the agent into a state where it can be tested in runtime. We will use the `prepare_agent` function from the boto3 sdk to prepare our agent.

```
[20]: response = bedrock_agent_client.prepare_agent(  
        agentId=agent_id  
)  
print(response)  
# Pause to make sure agent is prepared  
time.sleep(30)
```

```
{'ResponseMetadata': {'RequestId': 'ee10e01d-f9d4-4924-8cc0-31b62d502564', 'HTTPStatusCode': 202, 'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:25:11 GMT', 'content-type': 'application/json', 'content-length': '119', 'connection': 'keep-alive', 'x-amzn-requestid': 'ee10e01d-f9d4-4924-8cc0-31b62d502564', 'x-amz-apigw-id': 'HrpsIGbDPHcEGSQ=', 'x-amzn-trace-id': 'Root=1-67daefe6-3ccee22e77e5924268f68099'}, 'RetryAttempts': 0}, 'agentId': 'VMG6HC5N8I', 'agentStatus': 'PREPARING', 'agentVersion': 'DRAFT', 'preparedAt': datetime.datetime(2025, 3, 19, 16, 25, 11, 135568, tzinfo=tzlocal())}
```

Invoking Agent

Now that our Agent is ready to be used, let's test it. To do so we will use the `invoke_agent` function from the boto3 Bedrock runtime client.

To invoke an agent, you have to refer to its alias. You can create a new alias, or you can use the test alias to invoke your `DRAFT` agent. The test alias used to invoke the draft agent is `TSTALIASID` and it will work with any agent.

We will use the support function called `invoke_agent_helper` from the `agents.py` support file to allow us to invoke the agent with or without trace enabled and with or without session state. We will get into more details about those concepts in the `03_invoke_agent.ipynb` notebook.

```
[21]: alias_id = 'TSTALIASID'
```

```
[22]: %time  
session_id:str = str(uuid.uuid1())  
query = "Hi, I am Anna. I want to create a booking for 2 people, at 8pm on the 5th of May 2024."  
response = invoke_agent_helper(query, session_id, agent_id, alias_id)  
print(response)
```

Your booking for 2 people on May 5th, 2024 at 8pm has been created successfully. The booking ID is b51760cd.
CPU times: user 15.9 ms, sys: 600 µs, total: 16.5 ms



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

Launcher 01_create_agent.ipynb bedrock_basics.ipynb +

Notebook Cluster Python 3 (ipykernel)

Invoking Agent

Now that our Agent is ready to be used, let's test it. To do so we will use the `invoke_agent` function from the `boto3` Bedrock runtime client.

To invoke an agent, you have to refer to its alias. You can create a new alias, or you can use the test alias to invoke your `DRAFT` agent. The test alias used to invoke the draft agent is `TSTALIASID` and it will work with any agent.

We will use the support function called `invoke_agent_helper` from the `agents.py` support file to allow us to invoke the agent with or without trace enabled and with or without session state. We will get into more details about those concepts in the `03_invoke_agent.ipynb` notebook.

```
[21]: alias_id = 'TSTALIASID'
```

```
[22]: %%time
session_id:str = str(uuid.uuid1())
query = "Hi, I am Anna. I want to create a booking for 2 people, at 8pm on the 5th of May 2024."
response = invoke_agent_helper(query, session_id, agent_id, alias_id)
print(response)
```

Your booking for 2 people on May 5th, 2024 at 8pm has been created successfully. The booking ID is b51760cd.
CPU times: user 15.9 ms, sys: 600 µs, total: 16.5 ms
Wall time: 7.01 s

Next Steps

Before moving to the next notebook, let's store a couple of variables to continue working the the same notebook.

Next we will update our agent to associate a knowledge base containing the menus for our restaurant. We will then test the agent invocations and clean up all the created resources.

```
[23]: %store agent_id
%store agent_role
%store lambda_iam_role
%store agent_name
%store suffix
%store region
%store agent.foundation_model
```



us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio-landing

warn Support

Services

DynamoDB Managed NoSQL Database

Top features Tables Imports from S3 Explore Items Clusters Reserved Capacity

Amazon DocumentDB Fully-managed MongoDB-compatible database service

Athena Serverless interactive analytics service

Features

Show more

Settings

- DynamoDB feature

Clusters

- DynamoDB feature

Subnet Groups

- DynamoDB feature

Were these results helpful?

Yes No

United States (Oregon) WSParticipantRole/Participant @ 3640-1368-7952

Learn more

(US)

SageMaker Studio, you pay only for what you use. Authoring, training and hosting is

DynamoDB



Dashboard

Tables

Explore items

PartiQL editor

Backups

Exports to S3

Imports from S3

Integrations New

Reserved capacity

Settings

▼ DAX

Clusters

Subnet groups

Parameter groups

Events

Dashboard

Favorite tables

[View all tables](#)

Find favorite tables

< 1 > ⚙️

Table name	Status	Created at (UTC)
No favorite tables		

To get started, click the star icon on the tables page or table details page to favorite a table.

Alarms (0) [Info](#)[Manage in CloudWatch](#)

Find alarms

< 1 > ⚙️

Alarm name	Status
No custom alarms	

DAX clusters (0) [Info](#)[View details](#)

Find clusters

< 1 > ⚙️

Cluster name	Status
No DAX clusters	

Create resources

Create an Amazon DynamoDB table for fast and predictable database performance at any scale. [Learn more](#)

[Create table](#)

Amazon DynamoDB Accelerator (DAX) is a fully-managed, highly-available, in-memory caching service for DynamoDB. [Learn more](#)

[Create DAX cluster](#)

What's new

MAR 19 [Amazon DynamoDB Accelerator \(DAX\) SDK for JavaScript version 3 is now available](#)

MAR 19 [Amazon DynamoDB Accelerator \(DAX\) SDK for Go version 2 is now available](#)

MAR 17 [Amazon DynamoDB zero-ETL integration with Amazon Redshift now generally available](#)



us-west-2.console.aws.amazon.com/dynamodbv2/home?region=us-west-2#tables

aws Support

DynamoDB > Tables

Tables (3) Info

Find tables Any tag key Any tag value

Actions Delete Create table

Name	Status	Partition key	Sort key	Indexes	Re
BedrockChatStack-DatabaseConversationTable03F3FD7A-60PX4XSOH81H	Active	PK (S)	SK (S)	3	0
BedrockChatStack-DatabaseWebSocketSessionTable2302422E-1ZDH4YWR58DR	Active	ConnectionId (S)	MessagePartId (N)	0	0
restaurant_bookings	Active	booking_id (S)	-	0	0

A red arrow points to the "restaurant_bookings" table row.

us-west-2.console.aws.amazon.com/dynamodbv2/home?region=us-west-2#table?name=restaurant_bookings

aws Support

DynamoDB > Tables > restaurant_bookings

Tables (3)

- Any tag key
- Any tag value

Find tables

1

BedrockChatStack-
DatabaseConversationT
able03F3FD7A-
60PX4XSOH81H

BedrockChatStack-
DatabaseWebsocketSess
ionTable2302422E-
1ZDH4YWR58DR

restaurant_bookings

Actions

Explore table items

restaurant_bookings

Overview

Indexes

Monitor

Global tables

Backups

Exports and streams

Protect your DynamoDB table from accidental writes and deletes

When you turn on point-in-time recovery (PITR), DynamoDB backs up your table data automatically so that you can restore to any given second in the preceding 1 to 35 days.

Additional charges apply. [Learn more](#)

Edit PITR

General information

Partition key: booking_id (String)

Sort key: -

Capacity mode: On-demand

Table status: Active

Alarms: No active alarms

Point-in-time recovery (PITR): Off

Resource-based policy: Not active

Additional info

Items summary

Get live item count

DynamoDB updates the following information approximately every six hours.

Item count

Table size

Average item size

Support

DynamoDB

Dashboard

Tables

Explore items

PartiQL editor

Backups

Exports to S3

Imports from S3

Integrations New

Reserved capacity

Settings

▼ DAX

Clusters

Subnet groups

Parameter groups

Events

Tables (3)

Any tag key

Any tag value

Find tables

<

1

>

 BedrockChatStack-
DatabaseConversationTa
ble03F3FD7A-
60PX4XSOH81H BedrockChatStack-
DatabaseWebsocketSessi
onTable2302422E-
1ZDH4YWR58DR restaurant_bookings

restaurant_bookings

Autopreview

View table details

▼ Scan or query items

 Scan Query

Select a table or index

Table - restaurant_bookings

Select attribute projection

All attributes

► Filters - optional

Run

Reset

✓ Completed · Items returned: 1 · Items scanned: 1 · Efficiency: 100% · RCUs consumed: 2

Table: restaurant_bookings - Items returned (1)

Scan started on March 20, 2025, 00:27:57



Actions ▾

Create item



booking_id (String)	date	hour	name	num_guests
b51760ed	2024-05-05	20:00	Anna	2



DynamoDB

Dashboard

Tables

Explore items

PartiQL editor

Backups

Exports to S3

Imports from S3

Integrations [New](#)

Reserved capacity

Settings

▼ DAX

Clusters

Subnet groups

Parameter groups

Events

Any tag key

Any tag value

Find tables

< 1 >

BedrockChatStack-
DatabaseConversationTa
ble03F3FD7A-
60PX4XSOH81HBedrockChatStack-
DatabaseWebsocketSessi
onTable2302422E-
1ZDH4YWR58DR

restaurant_bookings



▼ Scan or query items

 Scan Query

Select a table or index

Table - restaurant_bookings

Select attribute projection

All attributes

► Filters - optional

Completed · Items returned: 1 · Items scanned: 1 · Efficiency: 100% · RCUs consumed: 2



Table: restaurant_bookings - Items returned (1)



Actions ▾

Create item

Scan started on March 20, 2025, 00:27:57

< 1 >

 booking_id (String) ▾ | date ▾ | hour ▾ | name ▾ | num_guests ▾ b51760cd ← 2024-05-05 20:00 Anna 2

← → C ⌂ us-west-2.console.aws.amazon.com/dynamodbv2/home?region=us-west-2#edit-item?itemMode=2&pk=b51760cd&route=ROUTE_ITEM_EXPLORER&table=rest... ☆ | T : |

aws | Search [Option+S] | United States (Oregon) | WSParticipantRole/Participant @ 3640-1368-7952

Support

DynamoDB > Explore items: restaurant_bookings > Edit item

Edit item

You can add, remove, or edit the attributes of an item. You can nest attributes inside other attributes up to 32 levels deep. [Learn more](#)

Attributes			Add new attribute ▾
Attribute name	Value	Type	
booking_id - Partition key	b51760cd	String	
date	2024-05-05	String	Remove
hour	20:00	String	Remove
name	Anna	String	Remove
num_guests	2	String	Remove

[Cancel](#) [Save](#) [Save and close](#)





Next Steps

Before moving to the next notebook, let's store a couple of variables to continue working the the same notebook.

Next we will update our agent to associate a knowledge base containing the menus for our restaurant. We will then test the agent invocations and clean up all the created resources.

```
[23]: %store agent_id  
%store agent_role  
%store lambda_iam_role  
%store agent_name  
%store suffix  
%store region  
%store agent.foundation_model  
%store account_id  
%store alias_id  
%store table_name  
%store lambda_function  
%store lambda_function_name  
%store agent_action_group_response  
%store agent_functions
```

```
Stored 'agent_id' (str)  
Stored 'agent_role' (dict)  
Stored 'lambda_iam_role' (dict)  
Stored 'agent_name' (str)  
Stored 'suffix' (str)  
Stored 'region' (str)  
Stored 'agent.foundation_model' (str)  
Stored 'account_id' (str)  
Stored 'alias_id' (str)  
Stored 'table_name' (str)  
Stored 'lambda_function' (dict)  
Stored 'lambda_function_name' (str)  
Stored 'agent_action_group_response' (dict)  
Stored 'agent_functions' (list)
```

+ + ↑ ↻ ✖ 滤

Launcher 01_create_agent.ipynb 02_associate_knowledge_base_to_agent.ipynb bedrock_basics.ipynb +

Notebook Cluster No Kernel

/ amazon-bedrock-workshop / 05_Agents /

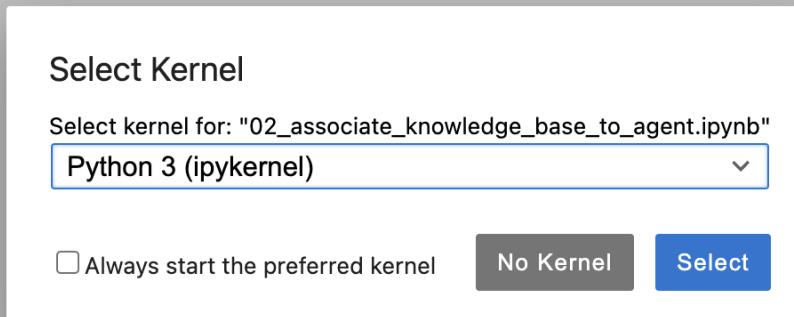
Name	Modified
images	21m ago
kb_documents	21m ago
00_inline_agents.ipynb	21m ago
01_create_agent.ipynb	1m ago
02_associate_knowledge_base_to_agent.ipynb	21m ago
03_invoke_agent.ipynb	21m ago
04_clean_up_agent_response.ipynb	21m ago
agent.py	21m ago
knowledge_base.py	21m ago
lambda_function.py	7m ago
README.md	29m ago
requirements.txt	29m ago

Agents for Amazon Bedrock - Associate Knowledge Base to Agent

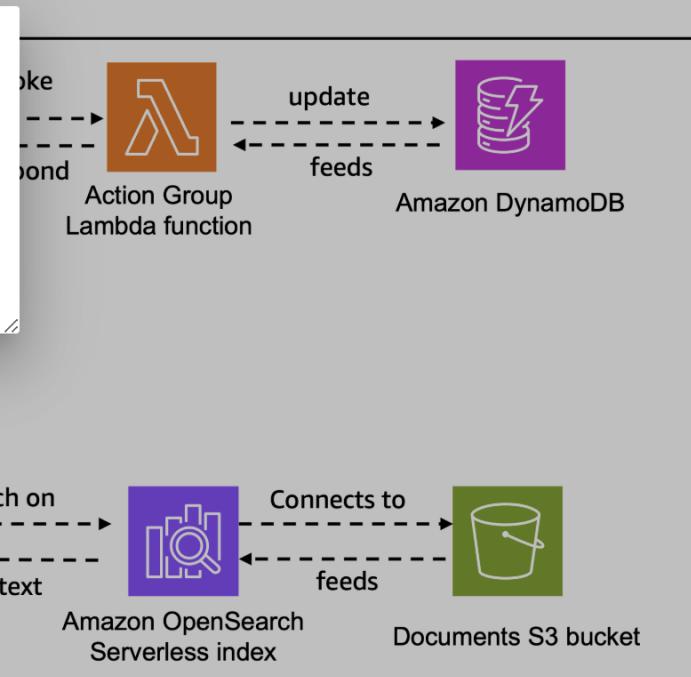
This notebook provides sample code for associating a Knowledge Base for Amazon Bedrock to an existent Agent for Amazon Bedrock that has an Action Group attached to it.

Use Case

We will update our restaurant assistant to allow customers to also ask questions about the restaurant menus. The architecture looks as following:



A modal dialog titled "Select Kernel" with the sub-instruction "Select kernel for: "02_associate_knowledge_base_to_agent.ipynb"" and a dropdown menu showing "Python 3 (ipykernel)". Below the dropdown are two buttons: "No Kernel" and "Select". To the left of the dialog is a small icon of a person labeled "customer".



The architecture diagram illustrates the flow of data between a customer, a Restaurant Agent, and various AWS services. The Restaurant Agent interacts with an LLM (Large Language Model) and a Menus Knowledge Base. It also connects to an Amazon OpenSearch Serverless index and a Documents S3 bucket via feeds. An Action Group Lambda function is triggered by the Restaurant Agent, which updates an Amazon DynamoDB table. The Menus Knowledge Base performs a search on context provided by the Restaurant Agent.

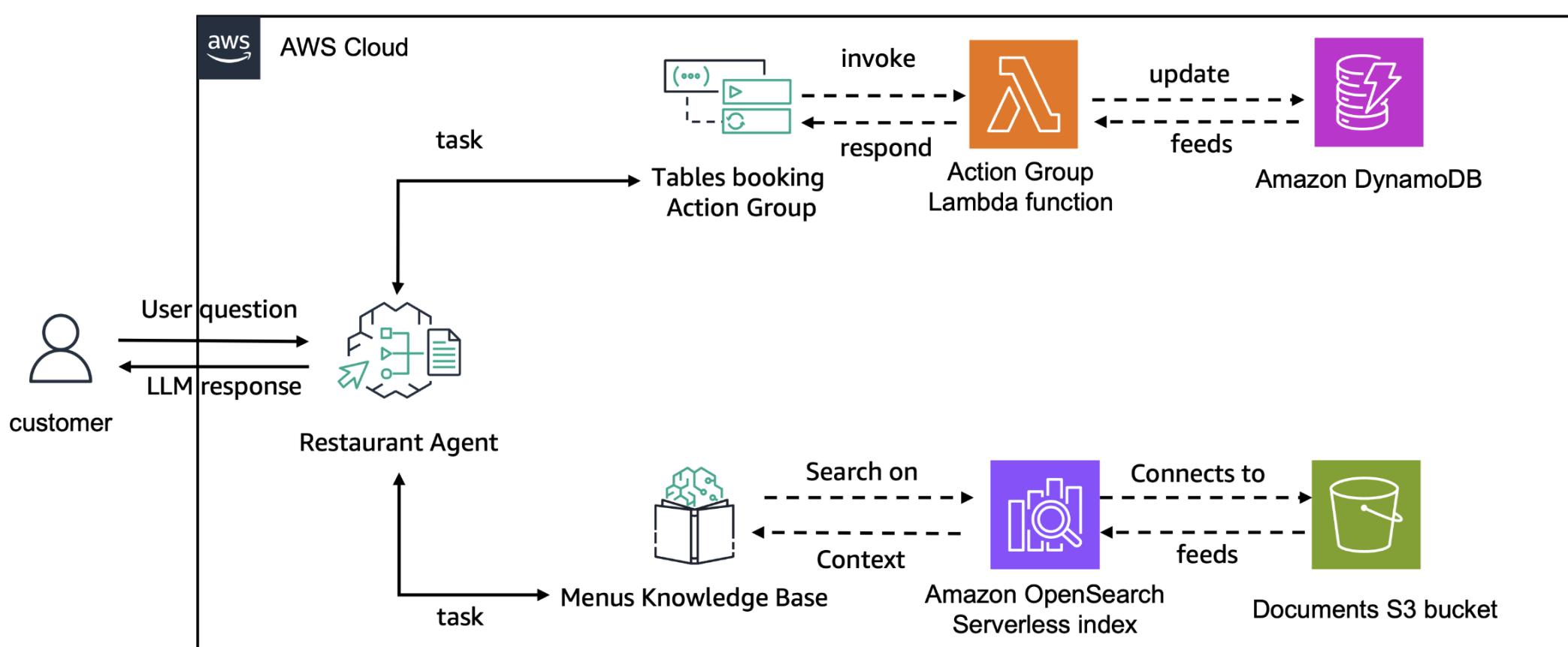
Notebook Walk-through

Agents for Amazon Bedrock - Associate Knowledge Base to Agent

This notebook provides sample code for associating a Knowledge Base for Amazon Bedrock to an existent Agent for Amazon Bedrock that has an Action Group attached to it.

Use Case

We will update our restaurant assistant to allow customers to also ask questions about the restaurant menus. The architecture looks as following:



File Edit View Run Kernel Git Tabs Settings Help sagemakeruser / amazon-bedrock-workshop-space

Launcher + X -

Run Selected Cell ↗ ⇌ Run Selected Cell and Insert Below ↘ ⇌ Run Selected Cell 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...

Associate Knowledge Base to Agent

We can associate a Knowledge Base for Amazon Bedrock to an existent Agent for Amazon Bedrock that has an Action Group attached to it.

We will use the Restaurant Agent example to ask questions about the restaurant menus. The architecture looks as following:

```

graph LR
    subgraph AWS_Cloud [AWS Cloud]
        AG[Action Group Lambda function] -- invoke --> AG_Box[Tables booking Action Group]
        AG_Box -- respond --> AG
        AG -- update --> DDB[Amazon DynamoDB]
        DDB -- feeds --> AG
        AG -- feeds --> OSE[Amazon OpenSearch Serverless index]
        OSE -- connects_to --> S3[Documents S3 bucket]
        OSE -- feeds --> KB[Menus Knowledge Base]
        KB -- search_on --> OSE
    end
    subgraph Customer [customer]
        User((User)) -- "User question" --> LLM[Restaurant Agent]
        LLM -- "LLM response" --> User
    end
    AG_Box -- task --> LLM
    LLM -- task --> KB

```

The diagram illustrates the architecture of the Restaurant Agent system. It consists of several components:

- AWS Cloud:** Contains an **Action Group Lambda function** (represented by an orange box with a Lambda icon), an **Amazon DynamoDB** (represented by a purple box with a lightning bolt icon), and an **Amazon OpenSearch Serverless index** (represented by a green box with a magnifying glass icon).
- Customer:** Represented by a user icon.
- Restaurant Agent:** Represented by a green brain-like icon.
- Knowledge Base:** Represented by a blue book icon.
- S3 Bucket:** Represented by a yellow bucket icon.

The flow of data is as follows:

- User** sends a **User question** to the **Restaurant Agent**.
- Restaurant Agent** sends an **LLM response** back to the **User**.
- Restaurant Agent** sends a **task** to the **Action Group Lambda function**.
- The **Action Group Lambda function** sends an **invoke** request to the **Tables booking Action Group**.
- The **Tables booking Action Group** sends a **respond** response back to the **Action Group Lambda function**.
- The **Action Group Lambda function** sends an **update** request to the **Amazon DynamoDB**.
- The **Amazon DynamoDB** sends **feeds** to the **Action Group Lambda function**.
- The **Action Group Lambda function** sends **feeds** to the **Amazon OpenSearch Serverless index**.
- The **Amazon OpenSearch Serverless index** sends **connects to** and **feeds** to the **Documents S3 bucket**.
- The **Amazon OpenSearch Serverless index** sends **search on Context** requests to the **Menus Knowledge Base**.
- The **Menus Knowledge Base** sends **feeds** to the **Amazon OpenSearch Serverless index**.
- The **Amazon OpenSearch Serverless index** sends **task** requests to the **Knowledge Base**.
- The **Knowledge Base** sends **task** requests back to the **Restaurant Agent**.

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_kb

Code git

Notebook Cluster Python 3 (ipykernel)



Pre-requisites

Before starting this lab, we need to load the variables that we stored in the previous notebook.

```
[1]: %store -r
```

```
[2]: agent_id
```

```
[2]: 'VMG6HC5N8I'
```

Let's now import the necessary libraries and initiate the required boto3 clients

```
[3]: from knowledge_base import KnowledgeBasesForAmazonBedrock
from agent import invoke_agent_helper
import boto3
import os
import time
import json
```

```
[4]: iam_client = boto3.client('iam')
bedrock_agent_client = boto3.client('bedrock-agent')
bedrock_agent_runtime_client = boto3.client('bedrock-agent-runtime')
s3_client = boto3.client('s3')
```

Setting up Knowledge Base Information

We will now set the variables that define our Knowledge Base:

- **knowledge_base_name**: provides the name of the Knowledge Base to be created, in this case `booking-agent-kb`
- **knowledge_base_description**: the description of the knowledge base used to display the agents list on the console. This description is **not** part of the agent's prompts
- **bucket_name**: name of the bucket containing the Knowledge Base documents

```
[5]: knowledge_base_name = f'{agent_id}-kb'
knowledge_base_description = "Knowledge Base containing the restaurant menu's collection"
```



Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_kb

Code git

Notebook Cluster Python 3 (ipykernel)



Setting up Knowledge Base Information

We will now set the variables that define our Knowledge Base:

- **knowledge_base_name**: provides the name of the Knowledge Base to be created, in this case `booking-agent-kb`
- **knowledge_base_description**: the description of the knowledge base used to display the agents list on the console. This description is **not** part of the agent's prompts
- **bucket_name**: name of the bucket containing the Knowledge Base documents

```
[5]: knowledge_base_name = f'{agent_name}-kb'  
knowledge_base_description = "Knowledge Base containing the restaurant menu's collection"  
bucket_name = f'{agent_name}-{suffix}'
```

Creating Knowledge Base for Amazon Bedrock

We will now go to create a Knowledge Base for Amazon Bedrock and its requirements including:

- [Amazon OpenSearch Serverless](#) for the vector database
- [AWS IAM](#) roles and permissions
- [Amazon S3](#) bucket to store the knowledge base documents

To create the knowledge base and its dependencies, we will use the `BedrockKnowledgeBase` support class, available in this folder. It allows you to create a new knowledge base, ingest documents to the knowledge base data source and delete the resources after you are done working with this lab.

Note that creation of the Amazon OpenSearch Serverless collection can take several minutes. You can use the Amazon OpenSearch Serverless console to monitor creation progress.

```
[6]: kb = KnowledgeBasesForAmazonBedrock()  
kb_id, ds_id = kb.create_or_retrieve_knowledge_base(knowledge_base_name, knowledge_base_description, bucket_name)
```

Creating KB booking-agent-kb
=====

Step 1 - Creating or retrieving booking-agent-us-west-2-364013687952 S3 bucket for Knowledge Base documents
Creating bucket booking-agent-us-west-2-364013687952
=====

Step 2 - Creating Knowledge Base Execution Role (AmazonBedrockExecutionRoleForKnowledgeBase_590) and Policies
=====



```
[6]: kb = KnowledgeBasesForAmazonBedrock()
kb_id, ds_id = kb.create_or_retrieve_knowledge_base(knowledge_base_name, knowledge_base_description, bucket_name)
```

Creating KB booking-agent-kl

Step 1 – Creating or retrieving booking-agent-us-west-2-364013687952 S3 bucket for Knowledge Base documents
Creating bucket booking-agent-us-west-2-364013687952

Step 2 – Creating Knowledge Base Execution Role (AmazonBedrockExecutionRoleForKnowledgeBase_590) and Policies

Step 3 – Creating OSS encryption, network and data access policies

Step 4 – Creating OSS Collection (this step takes a couple of minutes to complete)

```
        'x-amzn-requestid': '0911b6e1-b6f9-48c5-9b03-2738e96aecca'},
    'HTTPStatusCode': 200,
    'RequestId': '0911b6e1-b6f9-48c5-9b03-2738e96aecca',
    'RetryAttempts': 0},
'createCollectionDetail': { 'arn': 'arn:aws:aoss:us-west-2:364013687952:collection/ge0eptasvycaah1zobga',
    'createdDate': 1742401951722,
    'id': 'ge0eptasvycaah1zobga',
    'kmsKeyArn': 'auto',
    'lastModifiedDate': 1742401951722,
    'name': 'booking-agent-kb-590',
    'standbyReplicas': 'ENABLED',
    'status': 'CREATING',
    'ttype': 'VECTORSEARCH'}}}
```

ge0eptasvycaah1zobga.us-west-2.aoss.amazonaws.co

Creating collection.

Creating collection...
Creating collection...

Creating collection.....

Creating collection.

Creating collection.....

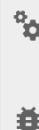
Creating collection.....

Creating collection.....

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_k

Code git

Notebook Cluster Python 3 (ipykernel)



```
Collection successfully created:  
[ { 'arn': 'arn:aws:aoss:us-west-2:364013687952:collection/ge0eptasvycaah1zobga',  
  'collectionEndpoint': 'https://ge0eptasvycaah1zobga.us-west-2.aoss.amazonaws.com',  
  'createdDate': 1742401951722,  
  'dashboardEndpoint': 'https://ge0eptasvycaah1zobga.us-west-2.aoss.amazonaws.com/_dashboards',  
  'id': 'ge0eptasvycaah1zobga',  
  'kmsKeyArn': 'auto',  
  'lastModifiedDate': 1742402158543,  
  'name': 'booking-agent-kb-590',  
  'standbyReplicas': 'ENABLED',  
  'status': 'ACTIVE',  
  'type': 'VECTORSEARCH'}]
```

Opensearch serverless arn: arn:aws:iam::364013687952:policy/AmazonBedrockOSSPolicyForKnowledgeBase_590

Sleeping for a minute to ensure data access rules have been enforced

=====

Step 5 – Creating OSS Vector Index

```
[2025-03-19 16:37:04,344] p610 {base.py:258} INFO - PUT https://ge0eptasvycaah1zobga.us-west-2.aoss.amazonaws.com:443/booking-agent-kb-index-590 [status:200 request:1.144s]
```

Creating index:

```
{ 'acknowledged': True,  
  'index': 'booking-agent-kb-index-590',  
  'shards_acknowledged': True}
```

=====

Step 6 – Creating Knowledge Base

```
{'type': 'VECTOR', 'vectorKnowledgeBaseConfiguration': {'embeddingModelArn': 'arn:aws:bedrock:us-west-2::foundation-model/amazon.titan-embed-text-v2:0'}}  
{ 'createdAt': datetime.datetime(2025, 3, 19, 16, 38, 4, 504202, tzinfo=tzlocal()),  
  'description': "Knowledge Base containing the restaurant menu's collection",  
  'knowledgeBaseArn': 'arn:aws:bedrock:us-west-2:364013687952:knowledge-base/RLF5M8ABK4',  
  'knowledgeBaseConfiguration': { 'type': 'VECTOR',  
    'vectorKnowledgeBaseConfiguration': { 'embeddingModelArn': 'arn:aws:bedrock:us-west-2::foundation-model/amazon.titan-embed-text-v2:0'}},  
  'knowledgeBaseId': 'RLF5M8ABK4',  
  'name': 'booking-agent-kb',  
  'roleArn': 'arn:aws:iam::364013687952:role/AmazonBedrockExecutionRoleForKnowledgeBase_590',  
  'status': 'CREATING',  
  'storageConfiguration': { 'opensearchServerlessConfiguration': { 'collectionArn': 'arn:aws:aoss:us-west-2:364013687952:collection/ge0eptasvycaah1zobga',  
    'fieldMapping': { 'metadataField': 'text-metadata',  
      'textField': 'text',  
      'vectorField': 'vector'}},  
    'openSearchServerlessArn': 'arn:aws:aoss:us-west-2:364013687952:open-search-serverless/kb-590'}}
```



Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_kb + Notebook Cluster Python 3 (ipykernel) ⚙

Step 6 – Creating Knowledge Base

```
{'type': 'VECTOR', 'vectorKnowledgeBaseConfiguration': {'embeddingModelArn': 'arn:aws:bedrock:us-west-2::foundation-model/amazon.titan-embed-text-v2:0'}},  
{'createdAt': datetime.datetime(2025, 3, 19, 16, 38, 4, 504202, tzinfo=tzlocal()),  
 'description': "Knowledge Base containing the restaurant menu's collection",  
 'knowledgeBaseArn': 'arn:aws:bedrock:us-west-2:364013687952:knowledge-base/RLF5M8ABK4',  
 'knowledgeBaseConfiguration': { 'type': 'VECTOR',  
 'vectorKnowledgeBaseConfiguration': { 'embeddingModelArn': 'arn:aws:bedrock:us-west-2::foundation-model/amazon.titan-embed-te  
 xt-v2:0'}},  
 'knowledgeBaseId': 'RLF5M8ABK4',  
 'name': 'booking-agent-kb',  
 'roleArn': 'arn:aws:iam::364013687952:role/AmazonBedrockExecutionRoleForKnowledgeBase_590',  
 'status': 'CREATING',  
 'storageConfiguration': { 'opensearchServerlessConfiguration': { 'collectionArn': 'arn:aws:aoss:us-west-2:364013687952:collection/ge0eptasvycaah1zobga',  
 'fieldMapping': { 'metadataField': 'text-metadata',  
 'textField': 'text',  
 'vectorField': 'vector'},  
 'vectorIndexName': 'booking-agent-kb-index-590'},  
 'type': 'OPENSEARCH_SERVERLESS'},  
 'updatedAt': datetime.datetime(2025, 3, 19, 16, 38, 4, 504202, tzinfo=tzlocal())}  
{'createdAt': datetime.datetime(2025, 3, 19, 16, 38, 5, 479427, tzinfo=tzlocal()),  
 'dataDeletionPolicy': 'RETAIN',  
 'dataSourceConfiguration': { 's3Configuration': { 'bucketArn': 'arn:aws:s3:::booking-agent-us-west-2-364013687952',  
 'type': 'S3'},  
 'dataSourceId': 'XYK5AZWKSV',  
 'description': "Knowledge Base containing the restaurant menu's collection",  
 'knowledgeBaseId': 'RLF5M8ABK4',  
 'name': 'booking-agent-kb',  
 'status': 'AVAILABLE',  
 'updatedAt': datetime.datetime(2025, 3, 19, 16, 38, 5, 479427, tzinfo=tzlocal()),  
 'vectorIngestionConfiguration': { 'chunkingConfiguration': { 'chunkingStrategy': 'FIXED_SIZE',  
 'fixedSizeChunkingConfiguration': { 'maxTokens': 512, 'overlapPercentage': 20}}}}}
```

We now upload the knowledge base documents to S3

```
[7]: def upload_directory(path, bucket_name):  
     for root, dirs, files in os.walk(path):  
         for file in files:  
             file_path = os.path.join(root, file)
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

[7]:

```
def upload_directory(path, bucket_name):
    for root, dirs, files in os.walk(path):
        for file in files:
            file_to_upload = os.path.join(root, file)
            print(f"uploading file {file_to_upload} to {bucket_name}")
            s3_client.upload_file(file_to_upload, bucket_name, file)

upload_directory("kb_documents", bucket_name)
```

uploading file kb_documents/Restaurant_Childrens_Menu.pdf to booking-agent-us-west-2-364013687952
uploading file kb_documents/Restaurant_Dinner_Menu.pdf to booking-agent-us-west-2-364013687952
uploading file kb_documents/Restaurant_week_specials.pdf to booking-agent-us-west-2-364013687952
uploading file kb_documents/.ipynb_checkpoints/Restaurant_Childrens_Menu-checkpoint.pdf to booking-agent-us-west-2-364013687952
uploading file kb_documents/.ipynb_checkpoints/Restaurant_Dinner_Menu-checkpoint.pdf to booking-agent-us-west-2-364013687952
uploading file kb_documents/.ipynb_checkpoints/Restaurant_week_specials-checkpoint.pdf to booking-agent-us-west-2-364013687952

And ingest the documents to the knowledge base

[8]:

```
# ensure that the kb is available
i_status = ['CREATING', 'DELETING', 'UPDATING']
while bedrock_agent_client.get_knowledge_base(knowledgeBaseId=kb_id)['knowledgeBase']['status'] in i_status:
    time.sleep(10)

# sync knowledge base
kb.synchronize_data(kb_id, ds_id)
```

```
{ 'dataSourceId': 'XYK5AZWKS',  
  'ingestionJobId': 'EGB0BBJY0L',  
  'knowledgeBaseId': 'RLF5M8ABK4',  
  'startedAt': datetime.datetime(2025, 3, 19, 16, 39, 7, 13349, tzinfo=tzlocal()),  
  'statistics': { 'numberOfDocumentsDeleted': 0,  
                 'numberOfDocumentsFailed': 0,  
                 'numberOfDocumentsScanned': 0,  
                 'numberOfMetadataDocumentsModified': 0,  
                 'numberOfMetadataDocumentsScanned': 0,  
                 'numberOfModifiedDocumentsIndexed': 0,  
                 'numberOfNewDocumentsIndexed': 0},  
  'status': 'STARTING',
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_kb.ipynb +

Notebook Cluster Python 3 (ipykernel)

[8]: # ensure that the kb is available
i_status = ['CREATING', 'DELETING', 'UPDATING']
while bedrock_agent_client.get_knowledge_base(knowledgeBaseId=kb_id)['knowledgeBase']['status'] in i_status:
 time.sleep(10)

sync knowledge base
kb.synchronize_data(kb_id, ds_id)

{ 'dataSourceId': 'XYK5AZWKS',
 'ingestionJobId': 'EGB0BBJY0L',
 'knowledgeBaseId': 'RLF5M8ABK4',
 'startedAt': datetime.datetime(2025, 3, 19, 16, 39, 7, 13349, tzinfo=tzlocal()),
 'statistics': { 'numberOfDocumentsDeleted': 0,
 'numberOfDocumentsFailed': 0,
 'numberOfDocumentsScanned': 0,
 'numberOfMetadataDocumentsModified': 0,
 'numberOfMetadataDocumentsScanned': 0,
 'numberOfModifiedDocumentsIndexed': 0,
 'numberOfNewDocumentsIndexed': 0},
 'status': 'STARTING',
 'updatedAt': datetime.datetime(2025, 3, 19, 16, 39, 7, 13349, tzinfo=tzlocal())}
{ 'dataSourceId': 'XYK5AZWKS',
 'ingestionJobId': 'EGB0BBJY0L',
 'knowledgeBaseId': 'RLF5M8ABK4',
 'startedAt': datetime.datetime(2025, 3, 19, 16, 39, 7, 13349, tzinfo=tzlocal()),
 'statistics': { 'numberOfDocumentsDeleted': 0,
 'numberOfDocumentsFailed': 0,
 'numberOfDocumentsScanned': 6,
 'numberOfMetadataDocumentsModified': 0,
 'numberOfMetadataDocumentsScanned': 0,
 'numberOfModifiedDocumentsIndexed': 0,
 'numberOfNewDocumentsIndexed': 6},
 'status': 'COMPLETE',
 'updatedAt': datetime.datetime(2025, 3, 19, 16, 39, 18, 820154, tzinfo=tzlocal())}
.....

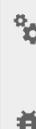
Testing Knowledge Base

Let's now test that the created knowledge base works as expected. To do so, we first retrieve the knowledge base id

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_base.ipynb +

Code git

Notebook Cluster Python 3 (ipykernel)



Testing Knowledge Base

Let's now test that the created knowledge base works as expected. To do so, we first retrieve the knowledge base id

Next we can use the `RetrieveAndGenerate` API from boto3 to retrieve the context for the question from the knowledge base and generate the final response

```
[9]: response = bedrock_agent_runtime_client.retrieve_and_generate(
    input={
        "text": "Which are the 5 mains available in the childrens menu?"}, retrieveAndGenerateConfiguration={ "type": "KNOWLEDGE_BASE", "knowledgeBaseConfiguration": { "knowledgeBaseId": kb_id, "modelArn": "arn:aws:bedrock:{}::foundation-model{}".format(region, agent_foundation_model), "retrievalConfiguration": { "vectorSearchConfiguration": { "number Of Results": 5 } } } }
```



```
print(response['output']['text'], end='\n'*2)
```

The 5 mains available in the children's menu are:

1. Mini Cheeseburgers – Small beef patties topped with cheese, served on mini buns. Allergens: Dairy, Gluten. Not suitable for vegetarians.
2. Fish Sticks – Breaded fish sticks served with tartar sauce. Allergens: Gluten (in the breading), possible Soy. Not suitable for vegetarians.
3. Grilled Cheese Sandwich – Melted cheese between slices of buttered bread, grilled to perfection. Allergens: Dairy, Gluten. Suitable for vegetarians.
4. Spaghetti with Marinara Sauce – Kid-friendly spaghetti noodles topped with tomato marinara sauce. Allergens: Gluten. Suitable for vegetarians.
5. Mini Pita Pizza – Small pita bread topped with tomato sauce, cheese, and favorite toppings. Allergens: Dairy, Gluten. Suitable for vegetarians.



File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_base_to_ag... +

Notebook Cluster Python 3 (ipykernel) ⚙️

The 5 mains available in the children's menu are:

1. Mini Cheeseburgers – Small beef patties topped with cheese, served on mini buns. Allergens: Dairy, Gluten. Not suitable for vegetarians.
2. Fish Sticks – Breaded fish sticks served with tartar sauce. Allergens: Gluten (in the breading), possible Soy. Not suitable for vegetarians.
3. Grilled Cheese Sandwich – Melted cheese between slices of buttered bread, grilled to perfection. Allergens: Dairy, Gluten. Suitable for vegetarians.
4. Spaghetti with Marinara Sauce – Kid-friendly spaghetti noodles topped with tomato marinara sauce. Allergens: Gluten. Suitable for vegetarians.
5. Mini Pita Pizza – Small pita bread topped with tomato sauce, cheese, and favorite toppings. Allergens: Dairy, Gluten. Suitable for vegetarians.

Updating Agent role to allow Knowledge Base Retrieve and Retrieve and Generate queries

Now that our Knowledge Base is working, we will associate it with the agent. To do so, we first need to update the agent role to allow for retrieval from context from our knowledge base

```
[10]: kb_policies_statements = [
    {
        "Sid": "QueryKB",
        "Effect": "Allow",
        "Action": [
            "bedrock:Retrieve",
            "bedrock:RetrieveAndGenerate"
        ],
        "Resource": [
            f"arn:aws:bedrock:{region}:{account_id}:knowledge-base/{kb_id}"
        ]
    }
]
bedrock_agent_kb_policy_statement = {
    "Version": "2012-10-17",
    "Statement": kb_policies_statements
}
bedrock_agent_kb_policy_json = json.dumps(bedrock_agent_kb_policy_statement)
kb_policy_name = f"{agent_name}-kb-{kb_id}"
agent_kb_policy = iam_client.create_policy(
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

+ + + + +

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_base_to_ag... +

Code git

Notebook Cluster Python 3 (ipykernel)

`print(response['output']['text'],end='\n'*2)`

The 5 mains available in the children's menu are:

1. Mini Cheeseburgers – Small beef patties topped with cheese, served on mini buns. Allergens: Dairy, Gluten. Not suitable for vegetarians.
2. Fish Sticks – Breaded fish sticks served with tartar sauce. Allergens: Gluten (in the breading), possible Soy. Not suitable for vegetarians.
3. Grilled Cheese Sandwich – Melted cheese between slices of buttered bread, grilled to perfection. Allergens: Dairy, Gluten. Suitable for vegetarians.
4. Spaghetti with Marinara Sauce – Kid-friendly spaghetti noodles topped with tomato marinara sauce. Allergens: Gluten. Suitable for vegetarians.
5. Mini Pita Pizza – Small pita bread topped with tomato sauce, cheese, and favorite toppings. Allergens: Dairy, Gluten. Suitable for vegetarians.

Updating Agent role to allow Knowledge Base Retrieve and Retrieve and Generate queries

Now that our Knowledge Base is working, we will associate it with the agent. To do so, we first need to update the agent role to allow for retrieval from context from our knowledge base

```
[10]: kb_policies_statements = [
    {
        "Sid": "QueryKB",
        "Effect": "Allow",
        "Action": [
            "bedrock:Retrieve",
            "bedrock:RetrieveAndGenerate"
        ],
        "Resource": [
            f"arn:aws:bedrock:{region}:{account_id}:knowledge-base/{kb_id}"
        ]
    }
]
```

File Edit View Run Kernel Git Tabs Settings Help sagemakeruser / amazon-bedrock-workshop-space

+ Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_base_to_ag... +

Name Modified

PDF Restaurant_Childrens_... 21m ago

PDF Restaurant_Dinner_Me... 41m ago

PDF Restaurant_week_spe... 41m ago

```
print(response['output']['text'],end='\n'*2)
```

The 5 mains available in the children's menu are:

1. Mini Cheeseburgers – Small beef patties topped with cheese, served on mini buns. Allergens: Dairy, Gluten. Not suitable for vegetarians.
2. Fish Sticks – Breaded fish sticks served with tartar sauce. Allergens: Gluten (in the breading), possible Soy. Not suitable for vegetarians.
3. Grilled Cheese Sandwich – Melted cheese between slices of buttered bread, grilled to perfection. Allergens: Dairy, Gluten. Suitable for vegetarians.
4. Spaghetti with Marinara Sauce – Kid-friendly spaghetti noodles topped with tomato marinara sauce. Allergens: Gluten. Suitable for vegetarians.
5. Mini Pita Pizza – Small pita bread topped with tomato sauce, cheese, and favorite toppings. Allergens: Dairy, Gluten. Suitable for vegetarians.

Updating Agent role to allow Knowledge Base Retrieve and Retrieve and Generate queries

Now that our Knowledge Base is working, we will associate it with the agent. To do so, we first need to update the agent role to allow for retrieval from context from our knowledge base

```
[10]: kb_policies_statements = [
    {
        "Sid": "QueryKB",
        "Effect": "Allow",
        "Action": [
            "bedrock:Retrieve",
            "bedrock:RetrieveAndGenerate"
        ],
        "Resource": [
            f"arn:aws:bedrock:{region}:{account_id}:knowledge-base/{kb_id}"
        ]
    }
]
```

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / amazon-bedrock-workshop-space

+ C ↻

/ ... / 05_Agents / kb_documents /

Name	Modified
Restaurant_Childrens_Menu.pdf	42 minutes ago
Restaurant_Dinner_Menu.pdf	42 minutes ago
Restaurant_week_specials.pdf	42 minutes ago

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_base_to_ag... +

Code git

Notebook Cluster Python 3 (ipykernel) +

print(response['output']['text'],end='\n'*2)

The 5 mains available in the children's menu are:

1. Mini Cheeseburgers – Small beef patties topped with cheese, served on mini buns. Allergens: Dairy, Gluten. Not suitable for vegetarians.
2. Fish Sticks – Breaded fish sticks served with tartar sauce. Allergens: Gluten (in the breading), possible Soy. Not suitable for vegetarians.
3. Grilled Cheese Sandwich – Melted cheese between slices of buttered bread, grilled to perfection. Allergens: Dairy, Gluten. Suitable for vegetarians.
4. Spaghetti with Marinara Sauce – Kid-friendly spaghetti noodles topped with tomato marinara sauce. Allergens: Gluten. Suitable for vegetarians.
5. Mini Pita Pizza – Small pita bread topped with tomato sauce, cheese, and favorite toppings. Allergens: Dairy, Gluten. Suitable for vegetarians.

Updating Agent role to allow Knowledge Base Retrieve and Retrieve and Generate queries

Now that our Knowledge Base is working, we will associate it with the agent. To do so, we first need to update the agent role to allow for retrieval from context from our knowledge base

```
[10]: kb_policies_statements = [
    {
        "Sid": "QueryKB",
        "Effect": "Allow",
        "Action": [
            "bedrock:Retrieve",
            "bedrock:RetrieveAndGenerate"
        ],
        "Resource": [
            "arn:aws:bedrock:{region}:{account_id}:knowledge-base/{kb_id}"
        ]
    }
]
```

...

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowled Restaurant_Childrens_

baf1f916-a9f3-4616-9d22-d672d2b... 1 / 3 - 100% + ⌂ ⌂

The Regrettable Experience — Children's Menu

Entrees:

1. CHICKEN NUGGETS

- Description:** Crispy chicken nuggets served with a side of ketchup or ranch dressing.
- Allergens:** Gluten (in the coating), possible Soy.
- Suitable for Vegetarians:** No

2. MACARONI AND CHEESE

- Description:** Classic macaroni pasta smothered in creamy cheese sauce.
- Allergens:** Dairy, Gluten.
- Suitable for Vegetarians:** Yes

3. MINI CHEESE QUESADILLAS

- Description:** Small flour tortillas filled with melted cheese, served with a mild salsa.
- Allergens:** Dairy, Gluten.
- Suitable for Vegetarians:** Yes

4. PEANUT BUTTER AND BANANA SANDWICH

- Description:** Peanut butter and banana slices on whole wheat bread.
- Allergens:** Nuts (peanut), Gluten.
- Suitable for Vegetarians:** Yes (if using vegetarian peanut butter)

5. VEGGIE PITA POCKETS

- Description:** Mini whole wheat pita pockets filled with hummus, cucumber, and cherry tomatoes.
- Allergens:** Gluten, possible Soy.
- Suitable for Vegetarians:** Yes

Mains:

Simple 0 \$ 3 main Instance MEM 57% ✓ Amazon Q

Cookie Preferences 0 Restaurant_Childrens_Menu.pdf 1

+ ⌂ ⌄ ⌅ ⌆

/ ... / 05_Agents / kb_documents /

Name	Modified
Restaurant_Childrens_Menu.pdf	42 minutes ago
Restaurant_Dinner_Menu.pdf	42 minutes ago
Restaurant_week_specials.pdf	42 minutes ago

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowled Restaurant_Childrens_

baf1f916-a9f3-4616-9d22-d672d2b... 1 / 3 - 100% + ⌂ ⌆

Mains: ←

1. MINI CHEESEBURGERS

- Description:** Small beef patties topped with cheese, served on mini buns.
- Allergens:** Dairy, Gluten.
- Suitable for Vegetarians:** No

2. FISH STICKS

- Description:** Breaded fish sticks served with tartar sauce.
- Allergens:** Gluten (in the breading), possible Soy.
- Suitable for Vegetarians:** No

3. GRILLED CHEESE SANDWICH

- Description:** Melted cheese between slices of buttered bread, grilled to perfection.
- Allergens:** Dairy, Gluten.
- Suitable for Vegetarians:** Yes

4. SPAGHETTI WITH MARINARA SAUCE

- Description:** Kid-friendly spaghetti noodles topped with tomato marinara sauce.
- Allergens:** Gluten.

Simple 0 \$ 3 main Instance MEM 57% ✓ Amazon Q

Cookie Preferences

Restaurant_Childrens_Menu.pdf 1

+

... / 05_Agents / kb_documents /

Name	Modified
Restaurant_Childrens_Menu.pdf	42 minutes ago
Restaurant_Dinner_Menu.pdf	42 minutes ago
Restaurant_week_specials.pdf	42 minutes ago

Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledg Restaurant_Childrens_

baf1f916-a9f3-4616-9d22-d672d2b...

2 / 3 | - 100% + |

The Regaleable Experience — Children's Menu

3. GRILLED CHEESE SANDWICH

- Description:** Melted cheese between slices of buttered bread, grilled to perfection.
- Allergens:** Dairy, Gluten.
- Suitable for Vegetarians:** Yes

4. SPAGHETTI WITH MARINARA SAUCE

- Description:** Kid-friendly spaghetti noodles topped with tomato marinara sauce.
- Allergens:** Gluten.
- Suitable for Vegetarians:** Yes

5. MINI PITA PIZZA

- Description:** Small pita bread topped with tomato sauce, cheese, and favorite toppings.
- Allergens:** Dairy, Gluten.
- Suitable for Vegetarians:** Yes

Desserts:

1. MINI ICE CREAM SUNDAE

- Description:** A scoop of vanilla ice cream with chocolate sauce and rainbow sprinkles.
- Allergens:** Dairy.
- Suitable for Vegetarians:** Yes

2. FRUIT KABOBS

- Description:** Fresh fruit chunks on skewers for a fun and healthy dessert.
- Allergens:** None.
- Suitable for Vegetarians:** Yes

3. CHOCOLATE CHIP COOKIE BITES

- Description:** Bite-sized chocolate chip cookies.
- Allergens:** Dairy, Gluten.



Launcher x bedrock_basics.ipynb x 01_create_agent.ipynb x 02_associate_knowledge.ipynb x Restaurant_Childrens_Men x +

A horizontal row of small, light-gray icons representing various software applications and tools, including a file folder, a plus sign, a pair of scissors, a clipboard, a play button, a square, a circular arrow, a double right arrow, the word "Code", a downward arrow, a clock, the word "git", and a blue folder icon.

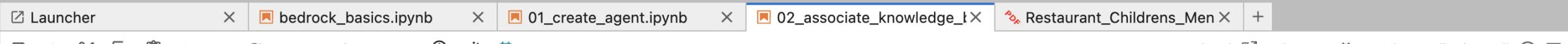
Notebook Cluster Python 3 (ipykernel)

Updating Agent role to allow Knowledge Base Retrieve and Retrieve and Generate queries

Now that our Knowledge Base is working, we will associate it with the agent. To do so, we first need to update the agent role to allow for retrieval from context from our knowledge base

```
[10]: kb_policies_statements = [
    {
        "Sid": "QueryKB",
        "Effect": "Allow",
        "Action": [
            "bedrock:Retrieve",
            "bedrock:RetrieveAndGenerate"
        ],
        "Resource": [
            f"arn:aws:bedrock:{region}:{account_id}:knowledge-base/{kb_id}"
        ]
    }
]
bedrock_agent_kb_policy_statement = {
    "Version": "2012-10-17",
    "Statement": kb_policies_statements
}
bedrock_agent_kb_policy_json = json.dumps(bedrock_agent_kb_policy_statement)
kb_policy_name = f"{agent_name}-kb-{kb_id}"
agent_kb_policy = iam_client.create_policy(
    PolicyName=kb_policy_name,
    PolicyDocument=bedrock_agent_kb_policy_json
)
iam_client.attach_role_policy(
    RoleName=agent_role['Role']['RoleName'],
    PolicyArn=agent_kb_policy['Policy']['Arn']
)

[10]: {'ResponseMetadata': {'RequestId': '412cc961-3d5e-4baa-8e71-7d4469a2f541',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:40:08 GMT',
    'x-amzn-requestid': '412cc961-3d5e-4baa-8e71-7d4469a2f541',
    'content-type': 'text/xml'}}
```



The screenshot shows a Jupyter Notebook interface with several tabs open at the top: 'Launcher', 'bedrock_basics.ipynb', '01_create_agent.ipynb', '02_associate_knowledge_base.ipynb' (the active tab), and 'Restaurant_Childrens_Men.ipynb'. Below the tabs is a toolbar with icons for file operations like new, open, save, and run, along with 'Code', 'git', and a plus sign for adding new cells. On the right side, there are buttons for 'Notebook', 'Cluster', 'Python 3 (ipykernel)', and a gear icon for settings.

Associate Knowledge Base with Agent

Finally, we can associate the new knowledge base with the agent using the `AssociateAgentKnowledgeBase` API from boto3

```
[11]: #time.sleep(10)
response = bedrock_agent_client.associate_agent_knowledge_base(
    agentId=agent_id,
    agentVersion='DRAFT',
    description='Access the knowledge base when customers ask about the plates in the menu.',
    knowledgeBaseId=kb_id,
    knowledgeBaseState='ENABLED'
)
```

Preparing Agent

after updating our agent, we need to prepare it again to package all its new components

```
[12]: response = bedrock_agent_client.prepare_agent(
    agentId=agent_id
)
print(response)
# Pause to make sure agent is prepared
intermediate_agent_status = ['CREATING', 'PREPARING', 'UPDATING', 'VERSIONING']
while bedrock_agent_client.get_agent(agentId=agent_id)['agent']['agentStatus'] in intermediate_agent_status:
    time.sleep(10)
```

```
{'ResponseMetadata': {'RequestId': 'c183d089-05b6-4eb4-9d29-e2d8588833b9', 'HTTPStatusCode': 202, 'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:40:09 GMT', 'content-type': 'application/json', 'content-length': '119', 'connection': 'keep-alive', 'x-amzn-requestid': 'c183d089-05b6-4eb4-9d29-e2d8588833b9', 'x-amz-apigw-id': 'Hrr4dHXIPhCEWyA=', 'x-amzn-trace-id': 'Root=1-67daf368-3f7e3f53239a39257590a838'}, 'RetryAttempts': 0}, 'agentId': 'VMG6HC5N8I', 'agentStatus': 'PREPARING', 'agentVersion': 'DRAFT', 'preparedAt': datetime.datetime(2025, 3, 19, 16, 40, 9, 144684, tzinfo=tzlocal())}
```

Invoking Agent

Now that our Agent has been updated, let's test it again. To do so we will again use the `invoke_agent` function from the boto3 Bedrock runtime client.

We will use the same support function called `invoke_agent_helper` to allow us to invoke the agent with or without trace enabled and with or without session state. We will get into more



Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_ipynb Restaurant_Childrens_Men +

Notebook Cluster Python 3 (ipykernel)

Preparing Agent

after updating our agent, we need to prepare it again to package all its new components

```
[12]: response = bedrock_agent_client.prepare_agent(  
        agentId=agent_id  
)  
print(response)  
# Pause to make sure agent is prepared  
intermediate_agent_status = ['CREATING', 'PREPARING', 'UPDATING', 'VERSIONING']  
while bedrock_agent_client.get_agent(agentId=agent_id)['agent']['agentStatus'] in intermediate_agent_status:  
    time.sleep(10)
```

```
{'ResponseMetadata': {'RequestId': 'c183d089-05b6-4eb4-9d29-e2d8588833b9', 'HTTPStatusCode': 202, 'HTTPHeaders': {'date': 'Wed, 19 Mar 2025 16:40:09 GMT', 'content-type': 'application/json', 'content-length': '119', 'connection': 'keep-alive', 'x-amzn-requestid': 'c183d089-05b6-4eb4-9d29-e2d8588833b9', 'x-amz-apigw-id': 'Hrr4dHXIPHcEWyA=', 'x-amzn-trace-id': 'Root=1-67daf368-3f7e3f53239a39257590a838'}, 'RetryAttempts': 0}, 'agentId': 'VMG6HC5N8I', 'agentStatus': 'PREPARING', 'agentVersion': 'DRAFT', 'preparedAt': datetime.datetime(2025, 3, 19, 16, 40, 9, 144684, tzinfo=tzlocal())}
```

Invoking Agent

Now that our Agent has been updated, let's test it again. To do so we will again use the `invoke_agent` function from the boto3 Bedrock runtime client.

We will use the same support function called `invoke_agent_helper` to allow us to invoke the agent with or without trace enabled and with or without session state. We will get into more details about those concepts in the `03_invoke_agent.ipynb` notebook.

Now we can test it by asking a question where the answer is available in the knowledge base documents

```
[13]: %%time  
import uuid  
session_id:str = str(uuid.uuid1())  
query = "What are the entrees in the childrens menu?"  
response = invoke_agent_helper(query, session_id, agent_id, alias_id)  
print(response)
```



Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_ipynb Restaurant_Childrens_Men.ipynb

Code git

Notebook Cluster Python 3 (ipykernel)



Invoking Agent

Now that our Agent has been updated, let's test it again. To do so we will again use the `invoke_agent` function from the boto3 Bedrock runtime client.

We will use the same support function called `invoke_agent_helper` to allow us to invoke the agent with or without trace enabled and with or without session state. We will get into more details about those concepts in the `03_invoke_agent.ipynb` notebook.

Now we can test it by asking a question where the answer is available in the knowledge base documents

```
[13]: %%time
import uuid
session_id:str = str(uuid.uuid1())
query = "What are the entrees in the childrens menu?"
response = invoke_agent_helper(query, session_id, agent_id, alias_id)
print(response)
```

The children's menu entrees at The Regrettable Experience include:

- Chicken nuggets (allergens: gluten, possible soy)
- Macaroni and cheese (allergens: dairy, gluten)
- Mini cheese quesadillas (allergens: dairy, gluten)
- Peanut butter and banana sandwich (allergens: nuts, gluten)
- Veggie pita pockets (allergens: gluten, possible soy)
- Mini cheeseburgers (allergens: dairy, gluten)
- Fish sticks (allergens: gluten, possible soy)
- Grilled cheese sandwich (allergens: dairy, gluten)
- Spaghetti with marinara sauce (allergens: gluten)
- Mini pita pizza (allergens: dairy, gluten)

CPU times: user 17.7 ms, sys: 1.42 ms, total: 19.1 ms
Wall time: 13.5 s

Next Steps



Launcher bedrock_basics.ipynb 01_create_agent.ipynb 02_associate_knowledge_ipynb Restaurant_Childrens_Men

Code git

Notebook Cluster Python 3 (ipykernel)



Next Steps

Before moving to the next notebook, let's store a couple of variables to continue working the the same notebook.

Next we will test our agent

```
[14]: %store kb_id  
%store knowledge_base_name  
%store knowledge_base_description  
%store kb_policy_name  
%store bucket_name  
%store knowledge_base_name
```

```
Stored 'kb_id' (str)  
Stored 'knowledge_base_name' (str)  
Stored 'knowledge_base_description' (str)  
Stored 'kb_policy_name' (str)  
Stored 'bucket_name' (str)  
Stored 'knowledge_base_name' (str)
```



Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- Validating Bedrock agent
- Exploring Bedrock fine-tune
- Q & A

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,  
}
```

<https://www.youtube.com/watch?v=YY9N7sDoP30>



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 MasterCard or VISA, so he uses his VISA card. He is given room 507, a nonsmoking room with a queen-size bed.

<https://www.youtube.com/watch?v=YY9N7sDoP30>



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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://www.youtube.com/watch?v=YY9N7sDoP30>



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.



+ New chat

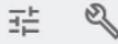
Query documents ⓘ

Index

canvas-blogs-EN



a Titan Express



Based on the information provided for the client and their goals and risk profile, recommend funds to invest on, and explain why.

Client Background: A 38-year-old married couple with moderate monthly expenses and a mortgage. They aim to retire comfortably in 20 years and want to support clean energy initiatives. They have a medium risk tolerance.

Goals: To retire comfortably and invest in clean energy.

Risk Profile: Medium

Recommendation:

Based on the client's goals and risk profile, I recommend a mix of equity and bond funds. Here are the details: Equity Funds:

1. Vanguard Total Stock Market ETF (VTI): This ETF tracks the performance of the entire U.S. stock market. It has a low expense ratio of 0.04% and a diversified portfolio of over 3,000 stocks. VTI provides exposure to a wide range of sectors, including technology, healthcare, and finance, which aligns with the client's clean energy interests.
2. iShares ESG Aware MSCI USA ETF (ESGU): This ETF invests in companies that meet ESG criteria, such as sustainability and social responsibility. It has a low expense ratio of 0.05% and a diversified portfolio of over 1,000 stocks. ESGU provides exposure to companies that are contributing to the transition to a greener economy, which aligns with the client's goals. Bond Funds:
3. Vanguard Total Bond Market ETF (BND): This ETF tracks the performance of the entire U.S. bond market. It has a low expense ratio of 0.05% and a diversified portfolio of over 3,000 bonds. BND provides stability and income, which can help the client achieve their retirement goals.
4. iShares ESG Aware MSCI USA Bond ETF (ESGB): This ETF invests in companies that meet ESG criteria, such as sustainability and social responsibility. It has a low expense ratio of 0.05% and a diversified portfolio of over 1,000 bonds. ESGB provides exposure to companies that are contributing to the transition to a greener economy, which aligns with the client's goals.

Explanation: Equity Funds: VTI and ESGU are both excellent choices for equity funds. VTI provides broad exposure to the U.S. stock market, which is a



Ask me anything...





+ New chat

Query documents ⓘ

Index
canvas-blogs-EN

FinancialPortfolio



Titan Express

Based on the information provided for the client and their goals and risk profile, recommend funds to invest on, and explain why.

Client Background: A 38-year-old married couple with moderate monthly expenses and a mortgage. They aim to retire comfortably in 20 years and want to support clean energy initiatives. They have a medium risk tolerance.

Goals: To retire comfortably and invest in clean energy.

Risk Profile: Medium

Recommendation:

Fund Allocation: 40% Clean Energy Impact Fund, 30% Early Retirement Portfolio, 30% Global Diversification Portfolio

Customised Results



Based on the information provided, I recommend the following funds to invest in:

Vanguard Total Stock Market Index Fund Admiral Shares: This fund provides broad exposure to the U.S. stock market and has a low expense ratio of 0.04%. It is a good choice for investors with a medium risk tolerance.

iShares ESG Aware MSCI USA ETF: This fund invests in companies that have strong environmental, social, and governance (ESG) practices. It has a low expense ratio of 0.05% and is a good choice for investors who want to support clean energy initiatives.

Vanguard Total International Stock Index Fund Admiral Shares: This fund provides exposure to stocks from around the world and has a low expense ratio of 0.04%. It is a good choice for investors with a medium risk tolerance who want to diversify their portfolio.

Generic Results

Ask me anything...



https://github.com/aws-samples/amazon-bedrock-workshop/tree/main/03_Model_customization

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

default-20240221t112169 / Personal Studio

00_setup.ipynb 01_fine-tuning-titan- X 02_fine-tuning_llama X 03_continued_pretrai X 1-TIGFT-customizatio X 2-TIGFT-provisioned- X

+

Filter files by name

/ amazon-bedrock-workshop-20240712 / 03_Model_customization /

Name	Last Modified
data	18 hours ago
fine-tuning-datasets	a day ago
=7,	a day ago
00_setup.ipynb	2 hours ago
• 01_fine-tuning-titan-lite.ipynb	2 hours ago
• 02_fine-tuning_llama2.ipynb	2 hours ago
• 03_continued_pretraining_titan...	2 hours ago
04_cleanup.ipynb	a day ago
README.md	a day ago

Setup for running customization notebooks both for fine-tuning and continued pre-training using Amazon Bedrock

In this notebook, we will create set of roles and an s3 bucket which will be used for other notebooks in this module.

This notebook should work well with the Data Science 3.0 , Python 3 , and ml.c5.2xlarge kernel in SageMaker Studio

Prerequisites

Custom job role

The notebook allows you to either create a Bedrock role for running customization jobs in the [Create IAM customisation job role](#) section or you can skip this section and create Bedrock Service role for customization jobs following [instructions on managing permissions for customization jobs](#). If you want to using an existing custom job role please edit the variable `customization_role` and also ensure it has access to the S3 bucket which is created containing the dataset.

Create IAM Pre-requisites

This notebook requires permissions to:

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

If you are running this notebook without an Admin role, make sure that your role include the following managed policies:

Search fo X | Search fo X | DevRel - X | Isengard X | Amazon X | Amazon X | 00_setup X | Domain- X | Domain- X | The Ama X | +

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.py X 2-TIGFT-provisioned-throughp X

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

Setup

Install and import all the needed libraries and dependencies to complete this notebook.

Warning: Please ignore error messages related to pip's dependency resolver.

```
[2]: !pip install --upgrade pip
%pip install --no-build-isolation --force-reinstall \
    "boto3>=1.28.57" \
    "awscli>=1.29.57" \
    "botocore>=1.31.57"
!pip install -qU --force-reinstall langchain typing_extensions pypdf urllib3==2.1.0
!pip install -qU ipywidgets>=7,<8
!pip install jsonlines
!pip install datasets==2.15.0
!pip install pandas==2.1.3
!pip install matplotlib==3.8.2
```

Requirement already satisfied: pip in /opt/conda/lib/python3.10/site-packages (24.1.2)
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusable.
It is recommended to use a virtual environment instead: <https://pip.pypa.io/warnings/venv>. Use the --root-user-action option if you know what you are doing and want to suppress this warning.
Collecting boto3>=1.28.57
Using cached boto3-1.34.143-py3-none-any.whl.metadata (6.6 kB)
Collecting awscli>=1.29.57
Using cached awscli-1.33.25-py3-none-any.whl.metadata (11 kB)
Collecting botocore>=1.31.57
Using cached botocore-1.34.143-py3-none-any.whl.metadata (5.7 kB)
Collecting jmespath<2.0.0,>=0.7.1 (from boto3>=1.28.57)
Using cached 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)

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

[3]: # restart kernel for packages to take effect
from IPython.core.display import HTML
HTML("<script>Jupyter.notebook.kernel.restart()</script>")

[3]:

[4]: import warnings
warnings.filterwarnings('ignore')
import json
import os
import sys
import boto3
import time
import pprint
from datasets import load_dataset
import random
import jsonlines

[5]: session = boto3.session.Session()
region = session.region_name
sts_client = boto3.client('sts')
account_id = sts_client.get_caller_identity()["Account"]
s3_suffix = f"{region}-{account_id}"
bucket_name = f"bedrock-customization-{s3_suffix}"
s3_client = boto3.client('s3')
bedrock = boto3.client(service_name="bedrock")
bedrock_runtime = boto3.client(service_name="bedrock-runtime")
iam = boto3.client('iam', region_name=region)

[6]: import uuid
suffix = str(uuid.uuid4())
role_name = "BedrockRole-" + suffix
s3_bedrock_finetuning_access_policy="BedrockPolicy-" + suffix
customization_role = f"arn:aws:iam::{account_id}:role/{role_name}"

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.py X 2-TIGFT-provisioned-throughp X

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

Testing boto3 connection

We will list the foundation models to test the bot3 connection and make sure bedrock client has been successfully created.

```
for model in bedrock.list.foundation_models(
    byCustomizationType="FINE_TUNING")["modelSummaries"]:
    for key, value in model.items():
        print(key, ":", value)
    print("----\n")

modelArn : arn:aws:bedrock:us-east-1::foundation-model/amazon.titan-image-generator-v1:0
modelId : amazon.titan-image-generator-v1:0
modelName : Titan Image Generator G1
providerName : Amazon
inputModalities : ['TEXT', 'IMAGE']
outputModalities : ['IMAGE']
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : ['PROVISIONED']
modelLifecycle : {'status': 'ACTIVE'}
-----

modelArn : arn:aws:bedrock:us-east-1::foundation-model/amazon.titan-text-lite-v1:0:4k
modelId : amazon.titan-text-lite-v1:0:4k
modelName : Titan Text G1 - Lite
providerName : Amazon
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING', 'CONTINUED_PRE_TRAINING']
inferenceTypesSupported : ['PROVISIONED']
modelLifecycle : {'status': 'ACTIVE'}
-----
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.py X 2-TIGFT-provisioned-throughp X

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

```
modelArn : arn:aws:bedrock:us-east-1::foundation-model/cohere.command-light-text-v14:7:4k
modelId : cohere.command-light-text-v14:7:4k
modelName : Command Light
providerName : Cohere
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : ['PROVISIONED']
modelLifecycle : {'status': 'ACTIVE'}
-----

modelArn : arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-13b-v1:0:4k ←
modelId : meta.llama2-13b-v1:0:4k
modelName : Llama 2 13B
providerName : Meta
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : []
modelLifecycle : {'status': 'LEGACY'}
-----

modelArn : arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-70b-v1:0:4k
modelId : meta.llama2-70b-v1:0:4k
modelName : Llama 2 70B
providerName : Meta
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : []
modelLifecycle : {'status': 'LEGACY'}
```

Search fo X | Search fo X | DevRel - X | Isengard X | Amazon X | Amazon X | 00_setup X | Domain- X | Domain- X | The Ama X | +

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughp X

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

Create s3 bucket

In this step we will create a s3 bucket, which will be used to store data for fine-tuning and continued pre-training notebooks.

```
[8]: # Create S3 bucket for knowledge base data source
s3bucket = s3_client.create_bucket(
    Bucket=bucket_name,
    ## Uncomment the following if you run into errors
    # CreateBucketConfiguration={
    #     'LocationConstraint':region,
    # },
)
```

Creating role and policies required to run customization jobs with Amazon Bedrock

This JSON object defines the trust relationship that allows the bedrock service to assume a role that will give it the ability to talk to other required AWS services. The conditions set restrict the assumption of the role to a specific account ID and a specific component of the bedrock service (model_customization_jobs)

```
[9]: ROLE_DOC = f"""
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "bedrock.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "{account_id}"
                }
            }
        }
    ]
}
```

Search fo X | Search fo X | DevRel - X | Isengard X | Amazon X | Amazon X | 00_setup X | Domain- X | Domain- X | The Ama X | +

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughp X

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

Creating role and policies required to run customization jobs with Amazon Bedrock

This JSON object defines the trust relationship that allows the bedrock service to assume a role that will give it the ability to talk to other required AWS services. The conditions set restrict the assumption of the role to a specific account ID and a specific component of the bedrock service (model_customization_jobs)

```
[9]: ROLE_DOC = f"""{{
    "Version": "2012-10-17",
    "Statement": [
        {{
            "Effect": "Allow",
            "Principal": {{
                "Service": "bedrock.amazonaws.com"
            }},
            "Action": "sts:AssumeRole",
            "Condition": {{
                "StringEquals": {{
                    "aws:SourceAccount": "{account_id}"
                }},
                "ArnEquals": {{
                    "aws:SourceArn": "arn:aws:bedrock:{region}:{account_id}:model-customization-job/*"
                }}
            }}
        }}
    ]
}}"""

```

This JSON object defines the permissions of the role we want bedrock to assume to allow access to the S3 bucket that we created that will hold our fine-tuning datasets and allow certain bucket and object manipulations.

```
[10]: ACCESS_POLICY_DOC = f"""{{
    "Version": "2012-10-17",
    "Statement": [
        {{
            "Effect": "Allow",
            "Principal": "bedrock.amazonaws.com",
            "Action": "s3:ListBucket",
            "Resource": "arn:aws:s3:::{bucket_name}"
        }},
        {{
            "Effect": "Allow",
            "Principal": "bedrock.amazonaws.com",
            "Action": "s3:PutObject",
            "Resource": "arn:aws:s3:::{bucket_name}/*"
        }},
        {{
            "Effect": "Allow",
            "Principal": "bedrock.amazonaws.com",
            "Action": "s3:GetObject",
            "Resource": "arn:aws:s3:::{bucket_name}/*"
        }},
        {{
            "Effect": "Allow",
            "Principal": "bedrock.amazonaws.com",
            "Action": "s3:DeleteObject",
            "Resource": "arn:aws:s3:::{bucket_name}/*"
        }}
    ]
}}"""

```

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

default-20240221t112169 / Personal Studio



00_setup.ipynb X | 01_fine-tuning-titan-lite.ipynb X | 02_fine-tuning_llama2.ipynb X | 03_continued_pretraining_tita X | 1-TIGFT-customization-job.ip X | 2-TIGFT-provisioned-throughp X

F + X D C ▶ git \$ ⌂

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

```
[10]: ACCESS_POLICY_DOC = f"""{{
    "Version": "2012-10-17",
    "Statement": [
        {{
            "Effect": "Allow",
            "Action": [
                "s3:AbortMultipartUpload",
                "s3:DeleteObject",
                "s3:PutObject",
                "s3:GetObject",
                "s3:GetBucketAcl",
                "s3:GetBucketNotification",
                "s3>ListBucket",
                "s3:PutBucketNotification"
            ],
            "Resource": [
                "arn:aws:s3:::{bucket_name}",
                "arn:aws:s3:::{bucket_name}/*"
            ]
        }}
    ]
}}"""
```

```
[11]: response = iam.create_role(
    RoleName=role_name,
    AssumeRolePolicyDocument=ROLE_DOC,
    Description="Role for Bedrock to access S3 for finetuning",
)
pprint.pprint(response)

{'Role': {'Path': '/',
          'RoleName': 'BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
          'RoleId': 'AROARMEM2CNOOMCZTH05I',
          'Arn': 'arn:aws:iamp::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

[11]:

```
response = iam.create_role(
    RoleName=role_name,
    AssumeRolePolicyDocument=ROLE_DOC,
    Description="Role for Bedrock to access S3 for finetuning",
)
pprint.pp(response)
```

{'Role': {'Path': '/',
 'RoleName': 'BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
 'RoleId': 'AROARMEM2CNOOMCZTH05I',
 'Arn': 'arn:aws:iam::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
 'CreateDate': datetime.datetime(2024, 7, 12, 5, 35, 7, tzinfo=tzlocal()),
 'AssumeRolePolicyDocument': {'Version': '2012-10-17',
 'Statement': [{'Effect': 'Allow',
 'Principal': {'Service': 'bedrock.amazonaws.com'},
 'Action': 'sts:AssumeRole',
 'Condition': {'StringEquals': {'aws:SourceAccount': '094784590684'},
 'ArnEquals': {'aws:SourceArn': 'arn:aws:bedrock:us-east-1:094784590684:model-customization-job/*'}}]}},
 'ResponseMetadata': {'RequestId': '8b973123-4d38-47f2-9340-a5bd2907c10b',
 'HTTPStatusCode': 200,
 'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 05:35:07 GMT',
 'x-amzn-requestid': '8b973123-4d38-47f2-9340-a5bd2907c10b',
 'content-type': 'text/xml',
 'content-length': '1812'},
 'RetryAttempts': 0}}

[12]:

```
role_arn = response["Role"]["Arn"]
pprint.pp(role_arn)
```

'arn:aws:iam::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac'

[13]:

```
response = iam.create_policy(
    PolicyName=s3_bedrock_finetuning_access_policy,
    PolicyDocument=ACCESS_POLICY_DOC,
)
pprint.pp(response)
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

[14]:

```
00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.py X 2-TIGFT-provisioned-throughp X
Cluster Data Science 3.0 Python 3 4 vCPU + 8 GiB Share ⚡
```

```
        )
pprint.pp(response)

{'Policy': {'PolicyName': 'BedrockPolicy-257549b9-3aff-4c30-86e9-530c1416e6ac',
  'PolicyId': 'ANPARMEM2CNOPFWJXG3E2',
  'Arn': 'arn:aws:iam::094784590684:policy/BedrockPolicy-257549b9-3aff-4c30-86e9-530c1416e6ac',
  'Path': '/',
  'DefaultVersionId': 'v1',
  'AttachmentCount': 0,
  'PermissionsBoundaryUsageCount': 0,
  'IsAttachable': True,
  'CreateDate': datetime.datetime(2024, 7, 12, 5, 35, 7, tzinfo=tzlocal()),
  'UpdateDate': datetime.datetime(2024, 7, 12, 5, 35, 7, tzinfo=tzlocal())},
 'ResponseMetadata': {'RequestId': '8bfffbb0b-4903-4692-82e5-de0f1a92d5ce',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 05:35:07 GMT',
    'x-amzn-requestid': '8bfffbb0b-4903-4692-82e5-de0f1a92d5ce',
    'content-type': 'text/xml',
    'content-length': '835'},
  'RetryAttempts': 0}}
```

[14]:

```
policy_arn = response["Policy"]["Arn"]
pprint.pp(policy_arn)
```

```
'arn:aws:iam::094784590684:policy/BedrockPolicy-257549b9-3aff-4c30-86e9-530c1416e6ac'
```

[15]:

```
iam.attach_role_policy(
    RoleName=role_name,
    PolicyArn=policy_arn,
)
```

[15]:

```
{'ResponseMetadata': {'RequestId': '081e20e5-d600-4e22-9ae1-758abafc53c8',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 05:35:07 GMT',
    'x-amzn-requestid': '081e20e5-d600-4e22-9ae1-758abafc53c8',
    'content-type': 'text/xml',
    'content-length': '212'}}
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughp X

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

Prepare CNN news article dataset for fine-tuning job and evaluation

The dataset that will be used is a collection of new articles from CNN and the associated highlights from that article. More information can be found at [huggingface](https://huggingface.co/datasets/cnn_dailymail):
https://huggingface.co/datasets/cnn_dailymail

#Load cnn dataset from huggingface
dataset = load_dataset("cnn_dailymail", "3.0.0")

View the structure of the dataset

[17]: print(dataset)

```
DatasetDict({  
    train: Dataset({  
        features: ['article', 'highlights', 'id'],  
        num_rows: 287113  
    })  
    validation: Dataset({  
        features: ['article', 'highlights', 'id'],  
        num_rows: 13368  
    })  
    test: Dataset({  
        features: ['article', 'highlights', 'id'],  
        num_rows: 11490  
    })  
})
```

Prepare the Fine-tuning Dataset In this example, we are using a .jsonl dataset following example format:

```
{"prompt": "", "completion": ""}
```

The following is an example item for a question-answer task: {"prompt": "prompt is AWS", "completion": "it's Amazon Web Services"}

A screenshot of a web browser window titled "Amazon WorkSpaces". The address bar shows a URL starting with "d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb". The browser interface includes standard navigation buttons (back, forward, search) and a tab bar with several open tabs.

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughp X

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

[18]: instruction='''Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.'''

instruction:

Summarize the news article provided below.

input:

...

For the 'completion' component we will attach the word "response" and new lines together with the summary/highlights of the article. The transformation of each datapoint is performed with the code below

```
[19]: datapoints_train=[]
for dp in dataset['train']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_train.append(temp_dict)
```

An example of a processed datapoint can be printed below

```
[20]: print(datapoints_train[4]['prompt'])
```

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.py X 2-TIGFT-provisioned-throughp X

An example of a processed datapoint can be printed below

```
print(datapoints_train[4]['prompt'])
```

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

(CNN) -- The National Football League has indefinitely suspended Atlanta Falcons quarterback Michael Vick without pay, officials with the league said Friday. NFL star Michael Vick is set to appear in court Monday. A judge will have the final say on a plea deal. Earlier, Vick admitted to participating in a dogfighting ring as part of a plea agreement with federal prosecutors in Virginia. "Your admitted conduct was not only illegal, but also cruel and reprehensible. Your team, the NFL, and NFL fans have all been hurt by your actions," NFL Commissioner Roger Goodell said in a letter to Vick. Goodell said he would review the status of the suspension after the legal proceedings are over. In papers filed Friday with a federal court in Virginia, Vick also admitted that he and two co-conspirators killed dogs that did not fight well. Falcons owner Arthur Blank said Vick's admissions describe actions that are "incomprehensible and unacceptable." The suspension makes "a strong statement that conduct which tarnishes the good reputation of the NFL will not be tolerated," he said in a statement. Watch what led to Vick's suspension ». Goodell said the Falcons could "assert any claims or remedies" to recover \$22 million of Vick's signing bonus from the 10-year, \$130 million contract he signed in 2004, according to The Associated Press. Vick said he would plead guilty to one count of "Conspiracy to Travel in Interstate Commerce in Aid of Unlawful Activities and to Sponsor a Dog in an Animal Fighting Venture" in a plea agreement filed at U.S. District Court in Richmond, Virginia. The charge is punishable by up to five years in prison, a \$250,000 fine, "full restitution, a special assessment and 3 years of supervised release," the plea deal said. Federal prosecutors agreed to ask for the low end of the sentencing guidelines. "The defendant will plead guilty because t

The same processing is done for the validation and test sets as well.

```
[21]: datapoints_valid=[]
for dp in dataset['validation']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_valid.append(temp_dict)
```

```
[22]: datapoints_test=[]
```

The screenshot shows a Jupyter notebook interface within Amazon SageMaker Studio. On the left, there's a file browser sidebar with a search bar and a list of files in the current directory:

- 00_setup.ipynb X
- 01_fine-tunin X
- 02_fine-tunin X
- test-cnn-10.js X
- train-cnn-5K.j X
- 03_continued X
- 1-TIGFT-custc X
- 2-TIGFT-provi X

The main area displays two numbered JSON objects representing prompts for a model to generate text completions:

```
1 {"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\ninstruction:\n\nSummarize the news article provided below.\n\ninput:\n\n(CNN) -- Slim Whitman, the high-pitched yodeling country music singer whose productive career spanned decades, died Wednesday, his son-in-law told CNN. He was 90. The singer-songwriter, born Ottis Dewey Whitman Jr., died of heart failure at Orange Park Medical Center in Florida, Roy Beagle said. Whitman gained fame in Europe as well as the United States. \"Love Song of the Waterfall,\" which a Country Music Television biography calls his \"breakthrough\" hit, was released in the early '50s. His next single \"Indian Love Call\" brought him stardom, according to the bio. \"Whitman joined the Grand Ole Opry, and then went to Britain in 1956 as the first country singer to play the London Palladium. Throughout the late '50s and early '60s, he had a string of British hits, including 'Tumbling Tumbleweeds,' 'Unchain My Heart,' and 'I'll Take You Home Again Kathleen,'\" the bio said. He gained cult status after he filmed a TV commercial that touted a released of his top greatest hits, a compilation that was a great success. \"Between 1980 and 1984, Whitman had a small run of minor hits, highlighted by 1980's number 15 hit 'When.' In the late '80s, he returned to television-marketed albums, releasing Slim Whitman: Best Loved Favorites in 1989 and 20 Precious Memories in 1991. During the '90s, Whitman recorded infrequently but continued to tour successfully, particularly in Europe and Australia,\" CMT said. His music was featured in the 1996 film \"Mars Attacks!\" In the film, the sound of Whitman's \"Indian Love Call\" made the heads of invading Martians explode. People we've lost in 2013: The lives they lived .", "completion": "response:\n\nCountry music singer Slim Whitman died at 90 Wednesday .\nThe cause was heart failure .\nThe yodeling singer gained fame in Europe as well as in the U.S."}
```

```
2 {"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\ninstruction:\n\nSummarize the news article provided below.\n\ninput:\n\nA young schoolgirl was horrified to discover six metal nails in her chicken and chips from the family's favourite local takeaway. Mollie Howe, 11, had been just about to tuck in to her tea when she spotted the inch-long nails in the £1.30 meal. Her furious mother, Sharon Howe, 39, has now blasted the Red Rooster fast food outlet, in Dagenham, Essex, which served her daughter the dangerous meal. Mollie Howe (left) and her mother Sharon Howe (right) were horrified to discover six nails in her takeaway from Red Rooster in Dagenham, Essex . Ms Howe said: 'I'm just so glad Mollie didn't eat it - who knows what might have happened? She came home, opened it up and there were these six nails - just like the ones you hammer into the wall. 'It's even more worrying because my dad could have eaten it as he often goes there too. He's 81 and lives on his own.' Ms Howe, of Dagenham, Essex, added that her family had been regular customers at the Red Rooster takeaway for years - but none of them will be going back. 'It's almost an everyday thing because it's cheap and cheerful. But I definitely won't be going there again - it just isn't worth the risk,' said the mother-of-two. The lethal inch-long metal nails were discovered nestled amongst the 11-year-old's chicken and chips . Maumand Khan, who works at Red Rooster (pictured) apologised and said he was baffled as to how the nails had got into the schoolgirl's food . 'None of my family and friends will be using it either. 'I just want people to be aware of what happened because it could have been a lot worse if poor Mollie had started eating and swallowed those nails.' Ms Howe said she was so outraged that she immediately went to confront staff at Red Rooster with her son Reece Hare, 21. 'I was so angry that my son wouldn't even let me go in there to see them,' she admitted. 'He went in instead and told them about it and they just wanted to offer him more food - it's not good enough.' Maumand Khan, who works at the takeaway, said he was baffled as to how the nails had got into Mollie's food. Mr Khan said: 'We are terribly sorry, but we can't understand how this has happened. 'We don't have that kind of nail anywhere in the shop .' He said they had not had any other complaints of this nature since the Red Rooster opened back in 2000. Furious Ms Howe ."}

A small blue message bubble icon is visible in the bottom right corner.


```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

The same processing is done for the validation and test sets as well.

```
[21]: datapoints_valid=[]
for dp in dataset['validation']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_valid.append(temp_dict)

[22]: datapoints_test=[]
for dp in dataset['test']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_test.append(temp_dict)
```

Here we define some helper functions to process our datapoints further by modifying the number of datapoints we want to include in each set and the max string length of the datapoints we want to include. The final function will convert our datasets into JSONL files.

```
[23]: def dp_transform(data_points,num_dps,max_dp_length):
    lines=[]
    for dp in data_points:
        if len(dp['prompt']+dp['completion'])<=max_dp_length:
            lines.append(dp)
    random.shuffle(lines)
    lines=lines[:num_dps]
    return lines

[24]: def jsonl_converter(dataset,file_name):
    print(file_name)
    with jsonlines.open(file_name, 'w') as writer:
        for line in dataset:
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughp X

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

Process data partitions. Every LLM may have different input token limits and what string of characters represents a token is defined by a particular model's vocabulary. For simplicity, we have restricted each datapoint to be <=3,000 characters.

```
[25]: train=dp_transform(datapoints_train,5000,3000)
validation=dp_transform(datapoints_valid,999,3000)
test=dp_transform(datapoints_test,10,3000)
```

Create local directory for datasets

Please note that your training dataset for fine-tuning cannot be greater than 10K records, and validation dataset has a maximum limit of 1K records.

```
[26]: dataset_folder="fine-tuning-datasets"
train_file_name="train-cnn-5K.jsonl"
validation_file_name="validation-cnn-1K.jsonl"
test_file_name="test-cnn-10.jsonl"
!mkdir fine-tuning-datasets
abs_path=os.path.abspath(dataset_folder)

mkdir: cannot create directory ‘fine-tuning-datasets’: File exists
```

Create JSONL format datasets for Bedrock fine-tuning job

```
[27]: jsonl_converter(train,f'{abs_path}/{train_file_name}')
jsonl_converter(validation,f'{abs_path}/{validation_file_name}')
jsonl_converter(test,f'{abs_path}/{test_file_name}')

/root/amazon-bedrock-workshop-20240712/03_Model_customization/fine-tuning-datasets/train-cnn-5K.jsonl
/root/amazon-bedrock-workshop-20240712/03_Model_customization/fine-tuning-datasets/validation-cnn-1K.jsonl
/root/amazon-bedrock-workshop-20240712/03_Model_customization/fine-tuning-datasets/test-cnn-10.jsonl
```

Upload datasets to s3 bucket

A screenshot of a web browser window titled "Amazon WorkSpaces". The address bar shows a URL starting with "d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb". The browser interface includes standard navigation buttons (back, forward, search) and a tab bar with several open tabs.

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughp X

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

Upload datasets to s3 bucket

Uploading both training and test dataset. We will use the training and validation datasets for fine-tuning the model. The test dataset will be used for evaluating the performance of the model on an unseen input.

```
[28]: s3_client.upload_file(f'{abs_path}/{train_file_name}', bucket_name, f'fine-tuning-datasets/train/{train_file_name}')
s3_client.upload_file(f'{abs_path}/{validation_file_name}', bucket_name, f'fine-tuning-datasets/validation/{validation_file_name}')
s3_client.upload_file(f'{abs_path}/{test_file_name}', bucket_name, f'fine-tuning-datasets/test/{test_file_name}'')
```

```
[29]: s3_train_uri=f's3://{bucket_name}/fine-tuning-datasets/train/{train_file_name}'
s3_validation_uri=f's3://{bucket_name}/fine-tuning-datasets/validation/{validation_file_name}'
s3_test_uri=f's3://{bucket_name}/fine-tuning-datasets/test/{test_file_name}''
```

Storing variables to be used in other notebooks.

Please make sure to use the same kernel as used for 00_setup.ipynb for other notebooks on fine-tuning and continued pre-training.

```
[30]: %store role_arn
%store bucket_name
%store role_name
%store policy_arn
%store s3_train_uri
%store s3_validation_uri
%store s3_test_uri
```

```
Stored 'role_arn' (str)
Stored 'bucket_name' (str)
Stored 'role_name' (str)
Stored 'policy_arn' (str)
Stored 's3_train_uri' (str)
Stored 's3_validation_uri' (str)
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-X 02_fine-tuning_llamaX 03_continued_pretrainX 1-TIGFT-customizatioX 2-TIGFT-provisioned-X

+

Filter files by name

/ amazon-bedrock-workshop-20240712 / 03_Model_customization /

Name	Last Modified
data	18 hours ago
fine-tuning-datasets	a day ago
=7,	a day ago
00_setup.ipynb	2 hours ago
01_fine-tuning-lite.ipynb	2 hours ago
02_fine-tuning_llama2.ipynb	2 hours ago
03_continued_pretraining_titan...	2 hours ago
04_cleanup.ipynb	a day ago
README.md	a day ago

A red arrow points to the "02_fine-tuning_llama2.ipynb" file in the list.

Fine-tune Meta Llama2 13B model provided by Amazon Bedrock: End-to-End

In this notebook we demonstrate using Boto3 sdk for the fine-tuning and provisioning of Llama2 13B model in Bedrock. You can also do this through the Bedrock Console.

Warning: This module cannot be executed in Workshop Studio Accounts, and you will have to run this notebook in your own account.

A Summarization Use Case

In this notebook, we build an end-to-end workflow for fine-tuning and evaluating the Foundation Models (FMs) in Amazon Bedrock. We choose Meta Llama 2 13B as our FM to perform the customization through fine-tuning, we then create provisioned throughput of the fine-tuned model, test the provisioned model invocation, and finally evaluate the fine-tuned model performance using fmeval on the summarization accuracy metrics including METEOR, ROUGE, and BERT scores. We have defined these scores in the Evaluate the Provisioned Custom Model section below.

This notebook should work well with the Data Science 3.0, Python 3, and ml.c5.2xlarge kernel in SageMaker Studio

Prerequisites

- Make sure you have executed 00_setup.ipynb notebook.
- Make sure you are using the same kernel and instance as 00_setup.ipynb notebook.

A Search fo X | A Search fo X | A DevRel - X | Isengard X | Amazon X | Amazon X | 02_fine-t X | A Domain- X | Domain- X | The Ama X | +

← → C ⌂ d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb ⌂ T :

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughput.ip X Cluster Data Science 3.0 Python 3 4 vCPU + 8 GiB Share ⚙

Fine-tune Meta Llama2 13B model provided by Amazon Bedrock: End-to-End

In this notebook we demonstrate using Boto3 sdk for the fine-tuning and provisioning of Llama2 13B model in Bedrock. You can also do this through the Bedrock Console.

Warning: This module cannot be executed in Workshop Studio Accounts, and you will have to run this notebook in your own account.

A Summarization Use Case

In this notebook, we build an end-to-end workflow for fine-tuning and evaluating the Foundation Models (FMs) in Amazon Bedrock. We choose Meta Llama 2 13B as our FM to perform the customization through fine-tuning, we then create provisioned throughput of the fine-tuned model, test the provisioned model invocation, and finally evaluate the fine-tuned model performance using fmeval on the summarization accuracy metrics including METEOR, ROUGE, and BERT scores. We have defined these scores in the Evaluate the Provisioned Custom Model section below.

This notebook should work well with the Data Science 3.0, Python 3, and ml.c5.2xlarge kernel in SageMaker Studio

Prerequisites

- Make sure you have executed 00_setup.ipynb notebook.
- Make sure you are using the same kernel and instance as 00_setup.ipynb notebook.

In this notebook we demonstrate using Boto3 sdk for the fine-tuning and provisioning of Llama2 13B model in Bedrock. You can also do this through the Bedrock Console.

Warning: This notebook will create provisioned throughput for testing the fine-tuned model. Therefore, please make sure to delete the provisioned throughput as mentioned in the last section of the notebook, otherwise you will be charged for it, even if you are not using it.

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_titan.ipynb X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughput.ipynb X

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

Setup

Install and import all the needed libraries and dependencies to complete this notebook.

Please ignore error messages related to pip's dependency resolver.

```
[2]: # # install the fmeval package for foundation model evaluation
!rm -Rf ~/.cache/pip/*
!pip install tokenizers==0.12.1
!pip install -qU fmeval==0.3.0

Collecting tokenizers==0.12.1
  Downloading tokenizers-0.12.1-cp310-cp310-manylinux_2_12_x86_64.manylinux2010_x86_64.whl.metadata (6.5 kB)
  Downloading tokenizers-0.12.1-cp310-cp310-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (6.6 MB)
    6.6/6.6 MB 39.9 MB/s eta 0:00:00:00:01

Installing collected packages: tokenizers
  Attempting uninstall: tokenizers
    Found existing installation: tokenizers 0.19.1
    Uninstalling tokenizers-0.19.1:
      Successfully uninstalled tokenizers-0.19.1
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 conflict
s.
transformers 4.42.4 requires tokenizers<0.20,>=0.19, but you have tokenizers 0.12.1 which is incompatible.
Successfully installed tokenizers-0.12.1
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusable.
It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppress this warning.
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 conflict
s.
sphinx 7.2.6 requires docutils<0.21,>=0.18.1, but you have docutils 0.16 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusable.
It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppress this warning.
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

[3]:

```
# restart kernel for packages to take effect
from IPython.core.display import HTML
HTML("<script>Jupyter.notebook.kernel.restart()</script>")
```

[3]:

[4]:

```
## Fetching variables from `00_setup.ipynb` notebook.
%store -r role_arn
%store -r s3_train_uri
%store -r s3_validation_uri
%store -r s3_test_uri
%store -r bucket_name
```

[5]:

```
import pprint
pprint.pprint(role_arn)
pprint.pprint(s3_train_uri)
pprint.pprint(s3_validation_uri)
pprint.pprint(s3_test_uri)
pprint.pprint(bucket_name)
```

'arn:aws:iam::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac'
's3://bedrock-customization-us-east-1-094784590684/fine-tuning-datasets/train/train-cnn-5K.jsonl'
's3://bedrock-customization-us-east-1-094784590684/fine-tuning-datasets/validation/validation-cnn-1K.jsonl'
's3://bedrock-customization-us-east-1-094784590684/fine-tuning-datasets/test/test-cnn-10.jsonl'
'bedrock-customization-us-east-1-094784590684'

[6]:

```
import warnings
warnings.filterwarnings('ignore')
import json
import os
import sys
import boto3
import pandas as pd
from matplotlib import pyplot as plt
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

[00_setup.ipynb] [01_fine-tuning-titan-lite.ipynb] [02_fine-tuning_llama2.ipynb] [03_continued_pretraining_tita...] [1-TIGFT-customization-job.ip...] [2-TIGFT-provisioned-throughp...]

[7]: session = boto3.Session()
region = session.region_name
sts_client = boto3.client('sts')
s3_client = boto3.client('s3')
aws_account_id = sts_client.get_caller_identity()["Account"]
bedrock = boto3.client(service_name="bedrock")
bedrock_runtime = boto3.client(service_name="bedrock-runtime")

[8]: test_file_name = "test-cnn-10.jsonl"
data_folder = "fine-tuning-datasets"

Create the Fine-Tuning Job

Note: Fine-tuning job will take around 60mins to complete with 5K records.

Meta Llama2 customization hyperparameters:

- epochs : The number of iterations through the entire training dataset and can take up any integer values in the range of 1-10, with a default value of 2.
- batchSize : The number of samples processed before updating model parameters and can take up any integer values in the range of 1-64, with a default value of 1.
- learningRate : The rate at which model parameters are updated after each batch which can take up a float value between 0.0-1.0 with a default value set to 1.00E-5.
- learningRateWarmupSteps : The number of iterations over which the learning rate is gradually increased to the specified rate and can take any integer value between 0-250 with a default value of 5.

For guidelines on setting hyper-parameters refer to the guidelines provided [here](#)

[9]: from datetime import datetime
ts = datetime.now().strftime("%Y-%m-%d-%H-%M-%S")

Amazon SageMaker Studio Classic default-20240221t112169 / Personal Studio

File Edit View Run Kernel Git Tabs Settings Help

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughp X

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

Create the Fine-Tuning Job

Note: Fine-tuning job will take around 60mins to complete with 5K records.

Meta Llama2 customization hyperparameters:

- epochs : The number of iterations through the entire training dataset and can take up any integer values in the range of 1-10, with a default value of 2.
- batchSize : The number of samples processed before updating model parameters and can take up any integer values in the range of 1-64, with a default value of 1.
- learningRate : The rate at which model parameters are updated after each batch which can take up a float value between 0.0-1.0 with a default value set to 1.00E-5.
- learningRateWarmupSteps : The number of iterations over which the learning rate is gradually increased to the specified rate and can take any integer value between 0-250 with a default value of 5.

For guidelines on setting hyper-parameters refer to the guidelines provided [here](#)

```
[9]: from datetime import datetime
ts = datetime.now().strftime("%Y-%m-%d-%H-%M-%S")

# Choose the foundation model you want to customize and provide ModelId(find more about model reference at https://docs.aws.amazon.com/bedrock/latest/userguide/bedrock-reference.html)
base_model_id = "meta.llama2-13b-v1:0:4k"

# Select the customization type from "FINE_TUNING" or "CONTINUED_PRE_TRAINING".
customization_type = "FINE_TUNING"

# Specify the roleArn for your customization job
customization_role = role_arn

# Create a customization job name
customization_job_name = f"llama2-finetune-sm-test-model-{ts}"
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

[9]:

```
from datetime import datetime
ts = datetime.now().strftime("%Y-%m-%d-%H-%M-%S")

# Choose the foundation model you want to customize and provide ModelId(find more about model reference at https://docs.aws.amazon.com/bedrock/latest/userguide/bedrock-reference.html)
base_model_id = "meta.llama2-13b-v1:0:4k"

# Select the customization type from "FINE_TUNING" or "CONTINUED_PRE_TRAINING".
customization_type = "FINE_TUNING"

# Specify the roleArn for your customization job
customization_role = role_arn

# Create a customization job name
customization_job_name = f"llama2-finetune-sm-test-model-{ts}"

# Create a customized model name for your fine-tuned Llama2 model
custom_model_name = f"llama2-finetune-{ts}"

# Define the hyperparameters for fine-tuning Llama2 model
hyper_parameters = {
    "epochCount": "2",
    "batchSize": "1",
    "learningRate": "0.00005",
}

# Specify your data path for training, validation(optional) and output
training_data_config = {"s3Uri": s3_train_uri}

# # uncomment the below section if you have validation dataset and provide the s3 uri for it.
validation_data_config = [
    {
        "s3Uri": s3_validation_uri
    }
]
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughp X

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

```
output_data_config = {"s3Uri": f's3://{bucket_name}/outputs/output-{custom_model_name}'}

# # Create the customization job
bedrock.create_model_customization_job(
    customizationType=customization_type,
    jobName=customization_job_name,
    customModelName=custom_model_name,
    roleArn=customization_role,
    baseModelIdentifier=base_model_id,
    hyperParameters=hyper_parameters,
    trainingDataConfig=training_data_config,
    validationDataConfig=validation_data_config,
    outputDataConfig=output_data_config
)
```

[9]: { 'ResponseMetadata': { 'RequestId': '814513c4-65b0-41bc-b83e-5abae996cf6c',
'HTTPStatusCode': 201,
'HTTPHeaders': { 'date': 'Fri, 12 Jul 2024 05:39:59 GMT',
'content-type': 'application/json',
'content-length': '112',
'connection': 'keep-alive',
'x-amzn-requestid': '814513c4-65b0-41bc-b83e-5abae996cf6c'},
'RetryAttempts': 0},
'jobArn': 'arn:aws:bedrock:us-east-1:094784590684:model-customization-job/meta.llama2-13b-v1:0:4k/hj6q5t4jcg91' }

Check Customization Job Status

```
[10]: import time
fine_tune_job = bedrock.get_model_customization_job(jobIdentifier=customization_job_name)[ "status"]
print(fine_tune_job)

while fine_tune_job == "InProgress":
    time.sleep(10)
```

The screenshot shows a browser window for Amazon WorkSpaces with the URL: https://d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb. The page title is "Amazon SageMaker Studio Classic". The top navigation bar includes File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. A notification bar indicates "default-20240221t112169 / Personal Studio". The left sidebar has icons for Home, Files, Folders, Recent, and Help. The main content area displays a Jupyter notebook tab labeled "02_fine-tuning_llama2.ipynb". The notebook contains the following code:

```
[10]: import time
fine_tune_job = bedrock.get_model_customization_job(jobIdentifier=customization_job_name)[ "status"]
print(fine_tune_job)

while fine_tune_job == "InProgress":
    time.sleep(60)
    fine_tune_job = bedrock.get_model_customization_job(jobIdentifier=customization_job_name)[ "status"]
    print (fine_tune_job)
```

The output of the code shows the job status "InProgress" repeated 25 times.

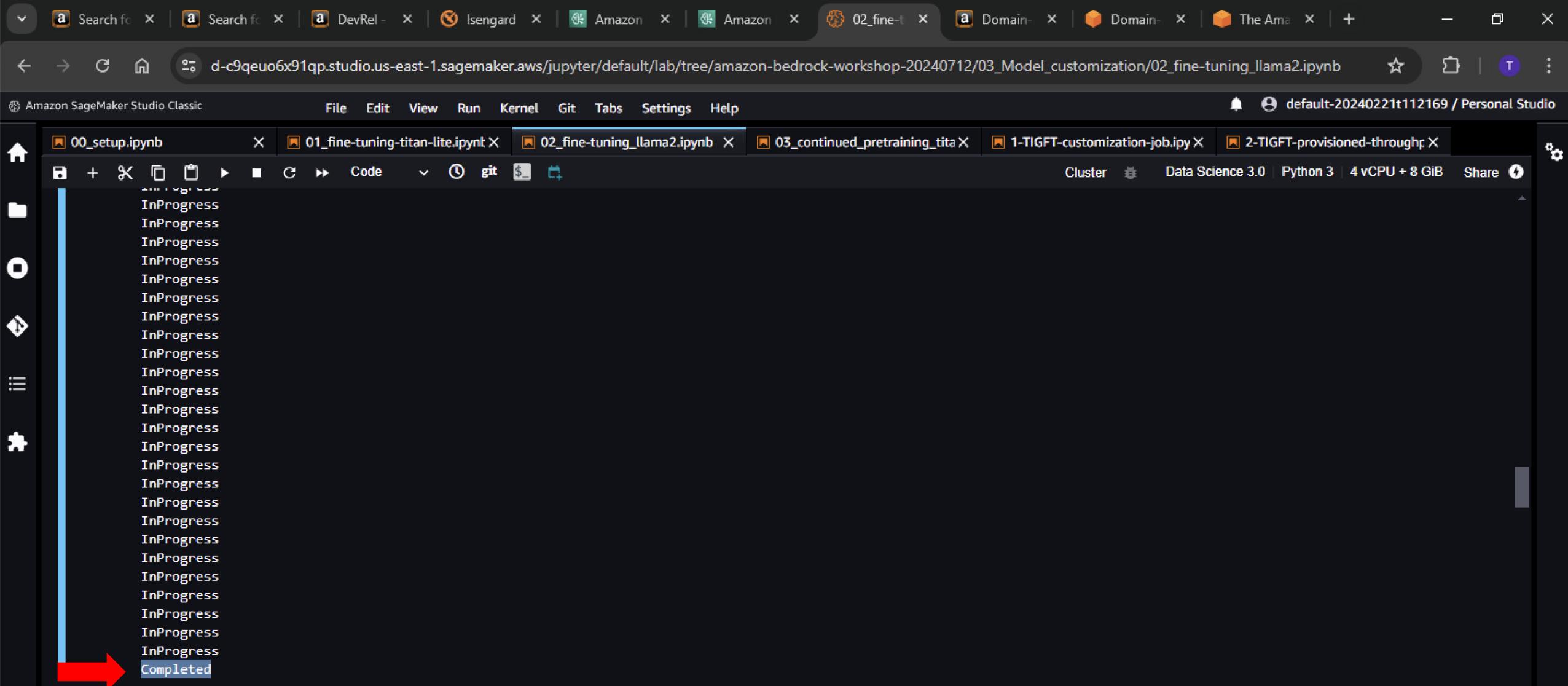
Screenshot of the Amazon Bedrock console showing the Custom models page.

The URL in the browser is us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models?tab=jobs.

The sidebar on the left includes sections for Getting started, Foundation models (Base models, Custom models selected), Playgrounds (Chat, Text, Image), Safeguards (Guardrails, Watermark detection), Builder tools (Knowledge bases, Agents, Prompt management Preview, Prompt flows Preview), and Assessment & deployment.

The main content area shows the "Custom models" section with the sub-section "How it works". It includes three cards: "Create a model" (with a slider icon), "Test a custom model" (with a play button icon), and "Use a custom model" (with a cube icon). Below this is a "Jobs" tab section showing 10 training jobs. A red arrow points to the "Status" column for the first job, which is labeled "Training".

Job name	Status	Source	Job type	Custom model name	Creation time
llama2-finetune-sm-te...	Training	Llama 2 13B	Fine-tuning	llama2-finetune-20...	July 12, 2024, 13:...



A screenshot of the Amazon SageMaker Studio interface. The top navigation bar shows multiple tabs, including "02_fine-tun". The main workspace displays a list of fine-tuning jobs. Most jobs are listed as "InProgress", but one job at the bottom is listed as "Completed". A large red arrow points from the bottom left towards the "Completed" job. The top right corner shows the cluster configuration: "default-20240221t112169 / Personal Studio", "Data Science 3.0", "Python 3", and "4 vCPU + 8 GiB".

Job Status
InProgress
Completed

Retrieve Custom Model

Once the customization job is finished, you can check your existing custom model(s) and retrieve the `modelArn` of your fine-tuned Llama2 model.

```
[11]: # You can list your custom models using the command below  
bedrock.list_custom_models()
```

AI/ML and Generative AI Enable | Isengard | Amazon Bedrock | us-east-1 | SageMaker Studio | 02_fine-tuni... (4) - JupyterLab | Paused

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models?tab=jobs

Services Search [Alt+S]

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

▼ How it works

Create a model

— — —

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Test a custom model

! Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

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

Models [Jobs](#)

Jobs (10)

Training jobs are initiated when you begin customizing a model.

Find job

C Stop job Customize model ▾

< 1 > ⌂

Job name	Status	Source	Job type	Custom model name	Creation...
llama2-finetune-sm-test-mod...	Complete	Llama 2 13B	Fine-tuning	llama2-finetune-20...	July 12, 2...

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughp X

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

Retrieve Custom Model

Once the customization job is finished, you can check your existing custom model(s) and retrieve the modelArn of your fine-tuned Llama2 model.

You can list your custom models using the command below

```
bedrock.list_custom_models()
```

[11]: {
 'ResponseMetadata': {'RequestId': 'c13432a9-498c-4300-8b6b-5297b2664c86',
 'HTTPStatusCode': 200,
 'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 06:26:07 GMT',
 'content-type': 'application/json',
 'content-length': '3719',
 'connection': 'keep-alive',
 'x-amzn-requestid': 'c13432a9-498c-4300-8b6b-5297b2664c86'},
 'RetryAttempts': 0},
 'modelSummaries': [{}{'modelArn': 'arn:aws:bedrock:us-east-1:094784590684:custom-model/meta.llama2-13b-v1:0:4k/vrme87cn8271',
 'modelName': 'llama2-finetune-2024-07-12-05-39-59',
 'creationTime': datetime.datetime(2024, 7, 12, 5, 39, 59, 672000, tzinfo=tzlocal()),
 'baseModelArn': 'arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-13b-v1:0:4k',
 'baseModelName': '',
 'customizationType': 'FINE_TUNING'},
 {}{'modelArn': 'arn:aws:bedrock:us-east-1:094784590684:custom-model/cohere.command-light-text-v14:7:4k/jglb0h06zthe',
 'modelName': 'cnn-daily-ft-20240630',
 'creationTime': datetime.datetime(2024, 6, 30, 2, 24, 3, 210000, tzinfo=tzlocal()),
 'baseModelArn': 'arn:aws:bedrock:us-east-1::foundation-model/cohere.command-light-text-v14:7:4k',
 'baseModelName': '',
 'customizationType': 'FINE_TUNING'},
 {}{'modelArn': 'arn:aws:bedrock:us-east-1:094784590684:custom-model/meta.llama2-13b-v1:0:4k/34kn0fkqp2sb',
 'modelName': 'llama2-finetune-2024-06-26-14-13-22',
 'creationTime': datetime.datetime(2024, 6, 26, 14, 13, 22, 359000, tzinfo=tzlocal()),
 'baseModelArn': 'arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-13b-v1:0:4k',
 'baseModelName': '',
 'customizationType': 'FINE_TUNING'},
 {}
] }

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

[12]: # retrieve the modelArn of the fine-tuned model
fine_tune_job = bedrock.get_custom_model(modelIdentifier=custom_model_name)
custom_model_id = fine_tune_job['modelArn']

[13]: output_job_name = "model-customization-job-"+fine_tune_job['jobArn'].split('/')[-1]
output_job_name

[13]: 'model-customization-job-hj6q5t4jcg91' 

Visualize Training and Validation Loss

Now that we have completed fine-tuning job, lets visualize our results to see if our job is not underfitting or overfitting.

Download model customization job metrics from S3 and plot the learning curves.

[14]: output_metrics_path = f"fine-tuning-datasets/{output_job_name}"

[15]: !mkdir \$output_metrics_path

[16]: train_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/training_artifacts/step_wise_training_metrics.csv'
validation_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/validation_artifacts/post_fine_tuning_validation/validation/validation_metrics.csv'
train_metrics_name='train_metrics.csv'
validation_metrics_name='validation_metrics.csv'
train_file_name_local=output_metrics_path+'/'+train_metrics_name
validation_file_name_local=output_metrics_path+'/'+validation_metrics_name

[17]: s3_client.download_file(bucket_name, train_metrics_s3_prefix, train_file_name_local)
s3_client.download_file(bucket_name, validation_metrics_s3_prefix, validation_file_name_local)

[18]: train_data = pd.read_csv(train_file_name_local)
'''The training loss is at an iteration level. To calculate loss at the epoch level,

A screenshot of an Amazon SageMaker Studio Classic interface showing a Jupyter notebook titled "02_fine-tuning_llama2.ipynb". The notebook contains code for visualizing training and validation loss.

The top navigation bar includes tabs for "File", "Edit", "View", "Run", "Kernel", "Git", "Tabs", "Settings", and "Help". The right side of the header shows the current workspace: "default-20240221t112169 / Personal Studio".

The toolbar below the header includes icons for file operations (New, Open, Save, etc.), a "Code" dropdown, a "git" icon, and a "Cluster" dropdown showing "Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB".

Visualize Training and Validation Loss

Now that we have completed fine-tuning job, lets visualize our results to see if our job is not underfitting or overfitting.

Download model customization job metrics from S3 and plot the learning curves.

```
[14]: output_metrics_path = f"fine-tuning-datasets/{output_job_name}"  
[15]: !mkdir $output_metrics_path  
[16]: train_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/training_artifacts/step_wise_training_metrics.csv'  
validation_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/validation_artifacts/post_fine_tuning_validation/validation/validation_metrics.csv'  
train_metrics_name='train_metrics.csv'  
validation_metrics_name='validation_metrics.csv'  
train_file_name_local=output_metrics_path+'/'+train_metrics_name  
validation_file_name_local=output_metrics_path+'/'+validation_metrics_name  
[17]: s3_client.download_file(bucket_name, train_metrics_s3_prefix, train_file_name_local)  
s3_client.download_file(bucket_name, validation_metrics_s3_prefix, validation_file_name_local)  
[18]: train_data = pd.read_csv(train_file_name_local)  
'''The training loss is at an iteration level. To calculate loss at the epoch level,  
average the iteration-level loss for each epoch'''  
train_metrics_epoch=train_data.groupby('epoch_number').mean()  
validation_metrics_epoch=pd.read_csv(validation_file_name_local)  
[19]: plt.plot(validation_metrics_epoch.epoch_number, validation_metrics_epoch.validation_loss,label='validation')  
plt.plot(train_metrics_epoch.index, train_metrics_epoch.training_loss,label='training')  
plt.title('Training vs Validation Loss')  
plt.ylabel('Loss')  
plt.xlabel('Epoch')
```

Search fo X | Search fo X | DevRel - X | Isengard X | Amazon X | Amazon X | 02_fine-t X | Domain- X | Domain- X | The Ama X | +

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughput.ipynb X

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

```
plt.xlabel('Epoch')
plt.legend()
plt.show()
```

Training vs Validation Loss

Epoch	Validation Loss	Training Loss
1.0	~2.85	~1.52
2.0	~0.95	~0.95

Create Provisioned Throughput

Provisioned Throughput for A X +

docs.aws.amazon.com/bedrock/latest/userguide/prov-throughput.html

Star | Print | T | :

aws Search in this guide Contact Us English ▾ Return to the Console

AWS > Documentation > Amazon Bedrock > User Guide Feedback Preferences

Amazon Bedrock User Guide

Recently added to this guide [Preview](#)

Delete a project from an Amazon Bedrock Studio workspace
July 23, 2024

Update a workspace for Prompt management and Prompt flows
July 23, 2024

Contextual grounding check
July 13, 2024

[View all](#)

► What is Amazon Bedrock?

► Getting started

Manage model access

► Foundation model information

Console overview

► Run model inference

► Provisioned throughput

Provisioned Throughput for Amazon Bedrock

[PDF](#) | [RSS](#)

Throughput refers to the number and rate of inputs and outputs that a model processes and returns. You can purchase **Provisioned Throughput** to provision a higher level of throughput for a model at a fixed cost. If you customized a model, you must purchase Provisioned Throughput to be able to use it.

You're billed hourly for a Provisioned Throughput that you purchase. For detailed information about pricing, see [Amazon Bedrock Pricing](#). The price per hour depends on the following factors:

1. The model that you choose (for custom models, pricing is the same as the base model that it was customized from).
2. The number of Model Units (MUs) that you specify for the Provisioned Throughput. An MU delivers a specific throughput level for the specified model. The throughput level of an MU specifies the following:
 - The number of input tokens that an MU can process across all requests within a span of one minute.
 - The number of output tokens that an MU can generate across all requests within a span of one minute.

Note

For more information about what an MU specifies, contact your AWS account manager.

3. The duration of time you commit to keeping the Provisioned Throughput. The longer the commitment duration, the more discounted the hourly price becomes. You can choose between the following levels of commitment:

Like | Dislike

00 setup.ipynb X 01 fine-tuning-titan-lite.ipynb X 02 fine-tuning_llama2.ipynb X 03 continued pretraining titan.ipynb X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughput.ipynb X

A set of small, light-gray navigation icons typically found in presentation software like Beamer. They include symbols for back, forward, search, and other document-related functions.

Note: Creating provisioned throughput will take around 20-30mins to complete.

You will need to create provisioned throughput to be able to evaluate the model performance. You can do so through the [console] (<https://docs.aws.amazon.com/bedrock/latest/userguide/prov-cap-console.html>) or use the following API call.

```
# Create the provision throughput job and retrieve the provisioned model id
provisioned_model_id = bedrock.create_provisioned_model_throughput(
    modelUnits=1,
    # create a name for your provisioned throughput model
    provisionedModelName='test-model-v1-001',
    modelId=custom_model_id
)[['provisionedModelArn']]
```

```
[21]: # check provisioned throughput job status
import time
status_provisioning = bedrock.get_provisioned_model_throughput(provisionedModelId = provisioned_model_id)['status']
while status_provisioning == 'Creating':
    time.sleep(60)
    status_provisioning = bedrock.get_provisioned_model_throughput(provisionedModelId=provisioned_model_id)['status']
    print(status_provisioning)
```

A | AI/ML and Generative AI Enable x | Isengard x | Amazon Bedrock | us-east-1 x | SageMaker Studio x | 02_fine-tuni... (4) - JupyterLab x | +

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/provisioned-throughput

Services Search [Alt+S]

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

Paused

Amazon Bedrock > Provisioned Throughput

Provisioned Throughput Info

Provisioned Throughput allows you to have dedicated capacity to deploy your models

Overview

Active	Failed	Updating	Inactive
1	0	0	0

Provisioned Throughput (1)

Name	Status	Model	Model units	Creation time	Commitment term	End
test-model-v1-001	<input checked="" type="checkbox"/> In service	llama2-finetune-2024-07-12-05-39-59	1	July 12, 2024, 14:26 (UTC+08:00)	No commitment	-

[Find Provisioned Throughput](#)

[Actions ▾](#) [Purchase Provisioned Throughput](#)

< 1 > [@](#)

Amazon SageMaker Studio Classic default-20240221t112169 / Personal Studio

File Edit View Run Kernel Git Tabs Settings Help

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ip X 2-TIGFT-provisioned-throughput.ip X

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

Invoke the Provisioned Custom Model

Invoke the provisioned custom model. You can replace the following prompt_txt with the prompts that are more similar to your fine-tuning dataset, this helps to check whether the fine-tuned model is performing as you expected.

Note: Please make sure your provisioned throughput job status becomes InService before proceeding.

```
[22]: # Provide the prompt text
test_file_path = f'{data_folder}/{test_file_name}'
with open(test_file_path) as f:
    lines = f.read().splitlines()
```

```
[23]: test_prompt = json.loads(lines[0])['prompt']
reference_summary = json.loads(lines[0])['completion']
print(test_prompt)
print()
print(reference_summary)
```

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

Pep Guardiola was left purring at Bayern Munich's first-half performance against Porto as the club put a turbulent week behind them to storm in to the Champions League semi-finals. A 3-1 first leg defeat in Portugal last week followed by the dramatic resignation of long-serving club doctor Hans-Wilhelm Muller-Wohlfahrt had turned the spotlight on Guardiola heading in to the tie. But the Bayern boss reminded the watching world why he is considered by many to be the best manager in the business as his side raced in to a 5-0 lead before half-time on Tuesday night. Bayern Munich players celebrate in front of their fans after the 6-1 demolition of Porto on Tuesday night. Pep Guardiola looks animated as he issues instructions from the touchline at the Allianz Arena. 'It's easy to love my players now,' said Guardiola. 'I am the coach of extraordinary player

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_titaX 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughp X

A screenshot of a Jupyter Notebook interface showing several tabs at the top. The active tab is '02_fine-tuning_llama2.ipynb'. Below the tabs is a toolbar with icons for file operations like new, open, save, and run. The main area contains code in a Python cell, starting with reading a file from a specific path.

```
test_file_path = r'{data_root}/{test_file_name}'  
with open(test_file_path) as f:  
    lines = f.read().splitlines()
```

```
[23]: test_prompt = json.loads(lines[0])['prompt']  
reference_summary = json.loads(lines[0])['completion']  
print(test_prompt)  
print()  
print(reference_summary)
```

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

Pep Guardiola was left purring at Bayern Munich's first-half performance against Porto as the club put a turbulent week behind them to storm in to the Champions League semi-finals. A 3-1 first leg defeat in Portugal last week followed by the dramatic resignation of long-serving club doctor Hans-Wilhelm Muller-Wohlfahrt had turned the spotlight on Guardiola heading in to the tie. But the Bayern boss reminded the watching world why he is considered by many to be the best manager in the business as his side raced in to a 5-0 lead before half-time on Tuesday night. Bayern Munich players celebrate in front of their fans after the 6-1 demolition of Porto on Tuesday night. Pep Guardiola looks animated as he issues instructions from the touchline at the Allianz Arena. 'It's easy to love my players now,' said Guardiola. 'I am the coach of extraordinary players. I didn't expect such a first half.' Typical of the man though, Guardiola found faults in Bayern's second half performance as Porto pulled back to 5-1 before Xabi Alonso added a late sixth. 'We can still play better,' he added. 'We lost the ball here and there and we can improve on that. We had a few problems in the second half. 'They had only a few chances but you should never relax in the Champions League.' Guardiola talks to captain Philipp Lahm (third right) as Mario Gotze (far left), Thiago and Thomas Müller listen in. It is believed that Muller-Wohlfahrt and his team resigned as they felt they were blamed for the first-leg defeat, though neither the doctor, the club or Guardiola have said as much in public. Bayern find out their semi-final opponents on Thursday with Guardiola's former club Barcelona already through to the last four and the winners of Real Madrid v Atletico Madrid and Monaco v Juventus to join them in the draw. Guardiola celebrates as Robert Lewandowski scores Bayern's fifth goal in the first half.

.

response:

Bayern Munich beat Porto 6-1 at the Allianz Arena on Tuesday night.
Result gave them a 7-4 aggregate victory in Champions League last eight.
Bayern manager Pep Guardiola hailed his players after the match.
READ: Luis Enrique 'happy' to see Pep Guardiola prove doubters wrong.

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughput.ipynb X

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



Construct model input following the format needed by Llama2 model following instructions [here](#). Please pay attention to the "Model invocation request body field" section

```
[24]: body = json.dumps({
    "prompt": test_prompt,
    # specify the parameters as needed
    "max_gen_len": 200,
    "temperature": 0.4,
    "top_p": 0.3,
})

# provide the modelId of the provisioned custom model
modelId = provisioned_model_id
accept = 'application/json'
contentType = 'application/json'

# invoke the provisioned custom model
response = bedrock_runtime.invoke_model(body=body, modelId=modelId, accept=accept, contentType=contentType)

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

{'generation': " response:\n\nBayern Munich put a turbulent week behind them to storm in to the Champions League semi-finals .\nPep Guardiola's side raced in to a 5-0 lead before half-time on Tuesday night .\nGuardiola reminded the watching world why he is considered by many to be the best manager in the business .", 'prompt_token_count': 531, 'generation_token_count': 85, 'stop_reason': 'stop'}

Clean up

Warning: Please make sure to delete provisioned throughput with the following code as there will be cost incurred if its left in running state, even if you are not using it.

```
[25]: # delete the provisioned throughput
bedrock.delete_provisioned_model_throughput(provisionedModelId=provisioned_model_id)
```

```
[25]: {'ResponseMetadata': {'RequestId': '2d0ec641-3779-4b8b-8812-bdb231931047',
```

AWS Sagemaker Studio Classic Jupyter Notebook - 02_fine-tuning_llama2.ipynb

File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb X 01_fine-tuning-titan-lite.ipynb X 02_fine-tuning_llama2.ipynb X 03_continued_pretraining_tita X 1-TIGFT-customization-job.ipynb X 2-TIGFT-provisioned-throughput.ipynb X

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

Clean up

Warning: Please make sure to delete provisioned throughput with the following code as there will be cost incurred if its left in running state, even if you are not using it.

```
# delete the provisioned throughput
bedrock.delete_provisioned_model_throughput(provisionedModelId=provisioned_model_id)
```

[25]: { 'ResponseMetadata': { 'RequestId': '2d0ec641-3779-4b8b-8812-bdb231931047', 'HTTPStatusCode': 200, 'HTTPHeaders': { 'date': 'Fri, 12 Jul 2024 06:38:00 GMT', 'content-type': 'application/json', 'content-length': '2', 'connection': 'keep-alive', 'x-amzn-requestid': '2d0ec641-3779-4b8b-8812-bdb231931047'}, 'RetryAttempts': 0}}

Note: Please finish up the cleaning process by running 04_cleanup.ipynb to clean up the other resources.

AI/ML and Generative AI Enable x | Isengard x | Amazon Bedrock | us-east-1 x | SageMaker Studio x | 02_fine-tuni... (4) - JupyterLab x | + - □ X

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/provisioned-throughput

Services Search [Alt+S]

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

Amazon Bedrock > Provisioned Throughput

Provisioned Throughput [Info](#)

Provisioned Throughput allows you to have dedicated capacity to deploy your models

Overview

Active	Failed	Updating	Inactive
0	0	0	0

Provisioned Throughput (0)

[Find Provisioned Throughput](#)

Name	Status	Model	Model units	Creation time	Commitment term	End time
No Provisioned Throughput						

There are currently no resources.

[Purchase Provisioned Throughput](#)

Actions ▾

Purchase Provisioned Throughput

N. Virginia ▾ Admin/linmicht-Isengard @ 0947-8459-0684 ▾

Labs

- Introducing Claude Chat App
- Testing user scenarios with Bedrock APIs
- Running Bedrock Knowledge Base
- Validating Bedrock agent
- Exploring Bedrock fine-tune
- Q & A



Thank you!

Michael Lin

linmicht@amazon.com