Amazon Bedrock

API Reference

Amazon Bedrock: API Reference

Table of Contents

Bedrock API Reference

This document provides detailed information about the Bedrock API actions and their parameters.

For information about the IAM access control permissions you need to use this API, see Identity-based policy examples for Amazon Bedrock.

You can use the following AWS SDKs to access Bedrock APIs.

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java
- AWS SDK for JavaScript
- AWS SDK for .NET
- AWS SDK for Python (Boto3)
- AWS SDK for Ruby

The SDKs automatically perform useful tasks for you, such as:

- Cryptographically sign your service requests
- Retry requests
- Handle error responses

The following resources provide additional information about the Amazon Bedrock API.

- *AWS General Reference*
  - Amazon Bedrock endpoints and quotas
- *AWS Command Line Interface*
  - Amazon Bedrock CLI commands
  - Amazon Bedrock Runtime CLI commands

Topics

Actions

The following actions are supported by Amazon Bedrock:

- GetCustomModel (p. 18)
- GetFoundationModel (p. 22)
- GetModelCustomizationJob (p. 24)
- GetModelInvocationLoggingConfiguration (p. 29)
- GetProvisionedModelThroughput (p. 31)
- ListCustomModels (p. 35)
- ListFoundationModels (p. 38)
- ListModelCustomizationJobs (p. 40)
- ListProvisionedModelThroughputs (p. 43)
- ListTagsForResource (p. 46)
- PutModelInvocationLoggingConfiguration (p. 48)
- StopModelCustomizationJob (p. 50)
- TagResource (p. 52)
- UntagResource (p. 54)
- UpdateProvisionedModelThroughput (p. 56)

The following actions are supported by Amazon Bedrock Runtime:

- InvokeModel (p. 59)
- InvokeModelWithResponseStream (p. 63)

Amazon Bedrock

The following actions are supported by Amazon Bedrock:

- CreateModelCustomizationJob (p. 3)
- CreateProvisionedModelThroughput (p. 8)
- DeleteCustomModel (p. 12)
- DeleteModelInvocationLoggingConfiguration (p. 14)
- DeleteProvisionedModelThroughput (p. 16)
- GetCustomModel (p. 18)
- GetFoundationModel (p. 22)
- GetModelCustomizationJob (p. 24)
- GetModelInvocationLoggingConfiguration (p. 29)
- GetProvisionedModelThroughput (p. 31)
- ListCustomModels (p. 35)
- ListFoundationModels (p. 38)
- ListModelCustomizationJobs (p. 40)
- ListProvisionedModelThroughputs (p. 43)
- ListTagsForResource (p. 46)
- PutModelInvocationLoggingConfiguration (p. 48)
- StopModelCustomizationJob (p. 50)
- TagResource (p. 52)
- UntagResource (p. 54)
- UpdateProvisionedModelThroughput (p. 56)

CreateModelCustomizationJob

Service: Amazon Bedrock

Creates a fine-tuning job to customize a base model.

You specify the base foundation model and the location of the training data. After the model-customization job completes successfully, your custom model resource will be ready to use. Training data contains input and output text for each record in a JSONL format. Optionally, you can specify validation data in the same format as the training data. Bedrock returns validation loss metrics and output generations after the job completes.

Model-customization jobs are asynchronous and the completion time depends on the base model and the training/validation data size. To monitor a job, use the `GetModelCustomizationJob` operation to retrieve the job status.

For more information, see Custom models in the Bedrock User Guide.

Request Syntax

```
POST /model-customization-jobs HTTP/1.1
Content-type: application/json

{
   "baseModelIdentifier": "string",
   "clientRequestToken": "string",
   "customModelKmsKeyId": "string",
   "customModelName": "string",
   "customModelTags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "hyperParameters": {
      "string" : "string"
   },
   "jobName": "string",
   "jobTags": [
      {
         "key": "string",
         "value": "string"
      }
   ],
   "outputDataConfig": {
      "s3Uri": "string"
   },
   "roleArn": "string",
   "trainingDataConfig": {
      "s3Uri": "string"
   },
   "validationDataConfig": {
      "validators": [
         {
            "s3Uri": "string"
         }
      ]
   },
   "vpcConfig": {
      "securityGroupIds": [ "string" ],
      "subnetIds": [ "string" ]
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

baseModelIdentifier (p. 3)

Name of the base model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)$`

Required: Yes

clientRequestToken (p. 3)

Unique token value that you can provide. The GetModelCustomizationJob response includes the same token value.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

customModelKmsKeyId (p. 3)

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:((key/[a-zA-Z0-9-]{36})|(alias/[a-zA-Z0-9-_/]+))$`

Required: No

customModelName (p. 3)

Enter a name for the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

Required: Yes

customModelTags (p. 3)

Assign tags to the custom model.

Type: Array of Tag (p. 84) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

hyperParameters (p. 3)

Parameters related to tuning the model.

Type: String to string map

Required: Yes

jobName (p. 3)

Enter a unique name for the fine-tuning job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*$`

Required: Yes

jobTags (p. 3)

Assign tags to the job.

Type: Array of Tag (p. 84) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

outputDataConfig (p. 3)

S3 location for the output data.

Type: OutputDataConfig (p. 79) object

Required: Yes

roleArn (p. 3)

The Amazon Resource Name (ARN) of an IAM role that Bedrock can assume to perform tasks on your behalf. For example, during model training, Bedrock needs your permission to read input data from an S3 bucket, write model artifacts to an S3 bucket. To pass this role to Bedrock, the caller of this API must have the `iam:PassRole` permission.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+$`

Required: Yes

trainingDataConfig (p. 3)

Information about the training dataset.

Type: TrainingDataConfig (p. 85) object

Required: Yes

validationDataConfig (p. 3)

Information about the validation dataset.

Type: ValidationDataConfig (p. 87) object

Required: No

vpcConfig (p. 3)

VPC configuration (optional). Configuration parameters for the private Virtual Private Cloud (VPC) that contains the resources you are using for this job.

Type: VpcConfig (p. 90) object

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "jobArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

jobArn (p. 6)

ARN of the fine tuning job

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateProvisionedModelThroughput

Service: Amazon Bedrock

Creates a provisioned throughput with dedicated capacity for a foundation model or a fine-tuned model.

For more information, see Provisioned throughput in the Bedrock User Guide.

Request Syntax

```
POST /provisioned-model-throughput HTTP/1.1
Content-type: application/json

{
   "clientRequestToken": "string",
   "commitmentDuration": "string",
   "modelId": "string",
   "modelUnits": number,
   "provisionedModelName": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

clientRequestToken (p. 8)

Unique token value that you can provide. If this token matches a previous request, Bedrock ignores the request, but does not return an error.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

Required: No

commitmentDuration (p. 8)

Commitment duration requested for the provisioned throughput.

Type: String

Valid Values: `OneMonth | SixMonths`

Required: No

modelId (p. 8)

Name or ARN of the model to associate with this provisioned throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)$`

Required: Yes

modelUnits (p. 8)

Number of model units to allocate.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

provisionedModelName (p. 8)

Unique name for this provisioned throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

Required: Yes

tags (p. 8)

Tags to associate with this provisioned throughput.

Type: Array of Tag (p. 84) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: No

Response Syntax

```
HTTP/1.1 201
Content-type: application/json

{
   "provisionedModelArn": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 201 response.

The following data is returned in JSON format by the service.

provisionedModelArn (p. 9)

The ARN for this provisioned throughput.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}$`

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3

- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteCustomModel

Service: Amazon Bedrock

Deletes a custom model that you created earlier. For more information, see Custom models in the Bedrock User Guide.

Request Syntax

```
DELETE /custom-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier (p. 12)

Name of the model to delete.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Delete the invocation logging.

Request Syntax

```
DELETE /logging/modelinvocations HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2

- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteProvisionedModelThroughput

Service: Amazon Bedrock

Deletes a provisioned throughput. For more information, see Provisioned throughput in the Bedrock User Guide.

Request Syntax

```
DELETE /provisioned-model-throughput/provisionedModelId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

provisionedModelId (p. 16)

The ARN or name of the provisioned throughput.

Pattern: `^((([0-9a-zA-Z][_-]?)+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetCustomModel

Service: Amazon Bedrock

Get the properties associated with a Bedrock custom model that you have created.For more information, see Custom models in the Bedrock User Guide.

Request Syntax

```
GET /custom-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier (p. 18)

Name or ARN of the custom model.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "baseModelArn": "string",
   "creationTime": "string",
   "hyperParameters": {
      "string" : "string"
   },
   "jobArn": "string",
   "jobName": "string",
   "modelArn": "string",
   "modelKmsKeyArn": "string",
   "modelName": "string",
   "outputDataConfig": {
      "s3Uri": "string"
   },
   "trainingDataConfig": {
      "s3Uri": "string"
   },
   "trainingMetrics": {
      "trainingLoss": number
   },
   "validationDataConfig": {
      "validators": [
```

```
        {
           "s3Uri": "string"
        }
     ]
  },
  "validationMetrics": [
     {
        "validationLoss": number
     }
  ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

baseModelArn (p. 18)

ARN of the base model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

creationTime (p. 18)

Creation time of the model.

Type: Timestamp

hyperParameters (p. 18)

Hyperparameter values associated with this model.

Type: String to string map

jobArn (p. 18)

Job ARN associated with this model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

jobName (p. 18)

Job name associated with this model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*$`

modelArn (p. 18)

ARN associated with this model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

modelKmsKeyArn (p. 18)

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

modelName (p. 18)

Model name associated with this model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

outputDataConfig (p. 18)

Output data configuration associated with this custom model.

Type: OutputDataConfig (p. 79) object

trainingDataConfig (p. 18)

Information about the training dataset.

Type: TrainingDataConfig (p. 85) object

trainingMetrics (p. 18)

The training metrics from the job creation.

Type: TrainingMetrics (p. 86) object

validationDataConfig (p. 18)

Array of up to 10 validators.

Type: ValidationDataConfig (p. 87) object

validationMetrics (p. 18)

The validation metrics from the job creation.

Type: Array of ValidatorMetric (p. 89) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetFoundationModel

Service: Amazon Bedrock

Get details about a Bedrock foundation model.

Request Syntax

```
GET /foundation-models/modelIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

modelIdentifier (p. 22)

The model identifier.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}((:[a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelDetails": {
      "customizationsSupported": [ "string" ],
      "inferenceTypesSupported": [ "string" ],
      "inputModalities": [ "string" ],
      "modelArn": "string",
      "modelId": "string",
      "modelName": "string",
      "outputModalities": [ "string" ],
      "providerName": "string",
      "responseStreamingSupported": boolean
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelDetails (p. 22)

Information about the foundation model.

Type: FoundationModelDetails (p. 72) object

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetModelCustomizationJob

Service: Amazon Bedrock

Retrieves the properties associated with a model-customization job, including the status of the job. For more information, see Custom models in the Bedrock User Guide.

Request Syntax

```
GET /model-customization-jobs/jobIdentifier HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier (p. 24)

Identifier for the customization job.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*)$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "baseModelArn": "string",
   "clientRequestToken": "string",
   "creationTime": "string",
   "endTime": "string",
   "failureMessage": "string",
   "hyperParameters": {
      "string" : "string"
   },
   "jobArn": "string",
   "jobName": "string",
   "lastModifiedTime": "string",
   "outputDataConfig": {
      "s3Uri": "string"
   },
   "outputModelArn": "string",
   "outputModelKmsKeyArn": "string",
   "outputModelName": "string",
   "roleArn": "string",
   "status": "string",
   "trainingDataConfig": {
      "s3Uri": "string"
   },
```

```
   "trainingMetrics": {
      "trainingLoss": number
   },
   "validationDataConfig": {
      "validators": [
         {
            "s3Uri": "string"
         }
      ]
   },
   "validationMetrics": [
      {
         "validationLoss": number
      }
   ],
   "vpcConfig": {
      "securityGroupIds": [ "string" ],
      "subnetIds": [ "string" ]
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

baseModelArn (p. 24)

ARN of the base model.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

clientRequestToken (p. 24)

The token that you specified in the CreateCustomizationJob request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9])*$`

creationTime (p. 24)

Time that the resource was created.

Type: Timestamp

endTime (p. 24)

Time that the resource transitioned to terminal state.

Type: Timestamp

failureMessage (p. 24)

Information about why the job failed.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

hyperParameters (p. 24)

The hyperparameter values for the job.

Type: String to string map

jobArn (p. 24)

The ARN of the customization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

jobName (p. 24)

The name of the customization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*$`

lastModifiedTime (p. 24)

Time that the resource was last modified.

Type: Timestamp

outputDataConfig (p. 24)

Output data configuration

Type: OutputDataConfig (p. 79) object

outputModelArn (p. 24)

The ARN of the output model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

outputModelKmsKeyArn (p. 24)

The custom model is encrypted at rest using this key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:kms:[a-zA-Z0-9-]*:[0-9]{12}:key/[a-zA-Z0-9-]{36}$`

outputModelName (p. 24)

The name of the output model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

roleArn (p. 24)

The ARN of the IAM role.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+$`

status (p. 24)

The status of the job. A successful job transitions from in-progress to completed when the output model is ready to use. If the job failed, the failure message contains information about why the job failed.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

trainingDataConfig (p. 24)

S3 Location of the training data.

Type: TrainingDataConfig (p. 85) object

trainingMetrics (p. 24)

Metrics associated with the custom job.

Type: TrainingMetrics (p. 86) object

validationDataConfig (p. 24)

Array of up to 10 validators.

Type: ValidationDataConfig (p. 87) object

validationMetrics (p. 24)

The loss metric for each validator that you provided in the createjob request.

Type: Array of ValidatorMetric (p. 89) objects

vpcConfig (p. 24)

VPC configuration for the custom model job.

Type: VpcConfig (p. 90) object

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Get the current configuration values for model invocation logging.

Request Syntax

```
GET /logging/modelinvocations HTTP/1.1
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "loggingConfig": {
      "cloudWatchConfig": {
         "largeDataDeliveryS3Config": {
            "bucketName": "string",
            "keyPrefix": "string"
         },
         "logGroupName": "string",
         "roleArn": "string"
      },
      "embeddingDataDeliveryEnabled": boolean,
      "imageDataDeliveryEnabled": boolean,
      "s3Config": {
         "bucketName": "string",
         "keyPrefix": "string"
      },
      "textDataDeliveryEnabled": boolean
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

loggingConfig (p. 29)

The current configuration values.

Type: LoggingConfig (p. 76) object

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetProvisionedModelThroughput

Service: Amazon Bedrock

Get details for a provisioned throughput. For more information, see Provisioned throughput in the Bedrock User Guide.

Request Syntax

```
GET /provisioned-model-throughput/provisionedModelId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

provisionedModelId (p. 31)

The ARN or name of the provisioned throughput.

Pattern: `^((([0-9a-zA-Z][_-]?)+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "commitmentDuration": "string",
   "commitmentExpirationTime": "string",
   "creationTime": "string",
   "desiredModelArn": "string",
   "desiredModelUnits": number,
   "failureMessage": "string",
   "foundationModelArn": "string",
   "lastModifiedTime": "string",
   "modelArn": "string",
   "modelUnits": number,
   "provisionedModelArn": "string",
   "provisionedModelName": "string",
   "status": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

commitmentDuration (p. 31)

Commitment duration of the provisioned throughput.

Type: String

Valid Values: `OneMonth | SixMonths`

commitmentExpirationTime (p. 31)

Commitment expiration time for the provisioned throughput.

Type: Timestamp

creationTime (p. 31)

The timestamp of the creation time for this provisioned throughput.

Type: Timestamp

desiredModelArn (p. 31)

The ARN of the new model to asssociate with this provisioned throughput.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

desiredModelUnits (p. 31)

The desired number of model units that was requested to be available for this provisioned throughput.

Type: Integer

Valid Range: Minimum value of 1.

failureMessage (p. 31)

Failure message for any issues that the create operation encounters.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

foundationModelArn (p. 31)

ARN of the foundation model.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

lastModifiedTime (p. 31)

The timestamp of the last modified time of this provisioned throughput.

Type: Timestamp

modelArn (p. 31)

The ARN or name of the model associated with this provisioned throughput.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

modelUnits (p. 31)

The current number of model units requested to be available for this provisioned throughput.

Type: Integer

Valid Range: Minimum value of 1.

provisionedModelArn (p. 31)

The ARN of the provisioned throughput.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}$`

provisionedModelName (p. 31)

The name of the provisioned throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

status (p. 31)

Status of the provisioned throughput.

Type: String

Valid Values: `Creating | InService | Updating | Failed`

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListCustomModels

Service: Amazon Bedrock

Returns a list of the custom models that you have created with the `CreateModelCustomizationJob` operation.

For more information, see Custom models in the Bedrock User Guide.

Request Syntax

```
GET /custom-models?
baseModelArnEquals=baseModelArnEquals&creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTi
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

baseModelArnEquals (p. 35)

Return custom models only if the base model ARN matches this parameter.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

creationTimeAfter (p. 35)

Return custom models created after the specified time.

creationTimeBefore (p. 35)

Return custom models created before the specified time.

foundationModelArnEquals (p. 35)

Return custom models only if the foundation model ARN matches this parameter.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

maxResults (p. 35)

Maximum number of results to return in the response.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains (p. 35)

Return custom models only if the job name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

nextToken (p. 35)

Continuation token from the previous response, for Bedrock to list the next set of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: ^\S*$

sortBy (p. 35)

The field to sort by in the returned list of models.

Valid Values: `CreationTime`

sortOrder (p. 35)

The sort order of the results.

Valid Values: `Ascending | Descending`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelSummaries": [
      {
         "baseModelArn": "string",
         "baseModelName": "string",
         "creationTime": "string",
         "modelArn": "string",
         "modelName": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelSummaries (p. 36)

Model summaries.

Type: Array of CustomModelSummary (p. 70) objects

nextToken (p. 36)

Continuation token for the next request to list the next set of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: ^\S*$

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListFoundationModels

Service: Amazon Bedrock

List of Bedrock foundation models that you can use. For more information, see Foundation models in the Bedrock User Guide.

Request Syntax

```
GET /foundation-models?
byCustomizationType=byCustomizationType&byInferenceType=byInferenceType&byOutputModality=byOutputModali
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

byCustomizationType (p. 38)

List by customization type.

Valid Values: `FINE_TUNING`

byInferenceType (p. 38)

List by inference type.

Valid Values: `ON_DEMAND | PROVISIONED`

byOutputModality (p. 38)

List by output modality type.

Valid Values: `TEXT | IMAGE | EMBEDDING`

byProvider (p. 38)

A Bedrock model provider.

Pattern: `^[a-z0-9-]{1,63}$`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelSummaries": [
      {
         "customizationsSupported": [ "string" ],
         "inferenceTypesSupported": [ "string" ],
         "inputModalities": [ "string" ],
         "modelArn": "string",
         "modelId": "string",
         "modelName": "string",
         "outputModalities": [ "string" ],
         "providerName": "string",
         "responseStreamingSupported": boolean
```

```
        }
    ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelSummaries (p. 38)

A list of bedrock foundation models.

Type: Array of FoundationModelSummary (p. 74) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListModelCustomizationJobs

Service: Amazon Bedrock

Returns a list of model customization jobs that you have submitted. You can filter the jobs to return based on one or more criteria.

For more information, see Custom models in the Bedrock User Guide.

Request Syntax

```
GET /model-customization-jobs?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults&nameCor
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

creationTimeAfter (p. 40)

Return customization jobs created after the specified time.

creationTimeBefore (p. 40)

Return customization jobs created before the specified time.

maxResults (p. 40)

Maximum number of results to return in the response.

Valid Range: Minimum value of 1. Maximum value of 1000.

nameContains (p. 40)

Return customization jobs only if the job name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*$`

nextToken (p. 40)

Continuation token from the previous response, for Bedrock to list the next set of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

sortBy (p. 40)

The field to sort by in the returned list of jobs.

Valid Values: `CreationTime`

sortOrder (p. 40)

The sort order of the results.

Valid Values: `Ascending | Descending`

statusEquals (p. 40)

Return customization jobs with the specified status.

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "modelCustomizationJobSummaries": [
      {
         "baseModelArn": "string",
         "creationTime": "string",
         "customModelArn": "string",
         "customModelName": "string",
         "endTime": "string",
         "jobArn": "string",
         "jobName": "string",
         "lastModifiedTime": "string",
         "status": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

modelCustomizationJobSummaries (p. 41)

Job summaries.

Type: Array of ModelCustomizationJobSummary (p. 77) objects

nextToken (p. 41)

Page continuation token to use in the next request.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: ^\S*$

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListProvisionedModelThroughputs

Service: Amazon Bedrock

List the provisioned capacities. For more information, see Provisioned throughput in the Bedrock User Guide.

Request Syntax

```
GET /provisioned-model-throughputs?
creationTimeAfter=creationTimeAfter&creationTimeBefore=creationTimeBefore&maxResults=maxResults&modelAr
 HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

creationTimeAfter (p. 43)

Return provisioned capacities created after the specified time.

creationTimeBefore (p. 43)

Return provisioned capacities created before the specified time.

maxResults (p. 43)

THe maximum number of results to return in the response.

Valid Range: Minimum value of 1. Maximum value of 1000.

modelArnEquals (p. 43)

Return the list of provisioned capacities where their model ARN is equal to this parameter.

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

nameContains (p. 43)

Return the list of provisioned capacities if their name contains these characters.

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

nextToken (p. 43)

Continuation token from the previous response, for Bedrock to list the next set of results.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

sortBy (p. 43)

The field to sort by in the returned list of provisioned capacities.

Valid Values: `CreationTime`

sortOrder (p. 43)

The sort order of the results.

Valid Values: `Ascending | Descending`

statusEquals (p. 43)

Return the list of provisioned capacities that match the specified status.

Valid Values: `Creating | InService | Updating | Failed`

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "nextToken": "string",
   "provisionedModelSummaries": [
      {
         "commitmentDuration": "string",
         "commitmentExpirationTime": "string",
         "creationTime": "string",
         "desiredModelArn": "string",
         "desiredModelUnits": number,
         "foundationModelArn": "string",
         "lastModifiedTime": "string",
         "modelArn": "string",
         "modelUnits": number,
         "provisionedModelArn": "string",
         "provisionedModelName": "string",
         "status": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken (p. 44)

Continuation token for the next request to list the next set of results.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^\S*$`

provisionedModelSummaries (p. 44)

List of summaries, one for each provisioned throughput in the response.

Type: Array of ProvisionedModelSummary (p. 80) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTagsForResource

Service: Amazon Bedrock

List the tags associated with the specified resource.

For more information, see Tagging resources in the Bedrock User Guide.

Request Syntax

```
POST /listTagsForResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string"
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN (p. 46)

The ARN of the resource.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `(^[a-zA-Z0-9][a-zA-Z0-9\-]*$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}|)((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12})$)|(:provisioned-model/[a-z0-9]{12}$)))`

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-type: application/json

{
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags (p. 46)

An array of the tags associated with this resource.

Type: Array of Tag (p. 84) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

PutModelInvocationLoggingConfiguration

Service: Amazon Bedrock

Set the configuration values for model invocation logging.

Request Syntax

```
PUT /logging/modelinvocations HTTP/1.1
Content-type: application/json

{
   "loggingConfig": {
      "cloudWatchConfig": {
         "largeDataDeliveryS3Config": {
            "bucketName": "string",
            "keyPrefix": "string"
         },
         "logGroupName": "string",
         "roleArn": "string"
      },
      "embeddingDataDeliveryEnabled": boolean,
      "imageDataDeliveryEnabled": boolean,
      "s3Config": {
         "bucketName": "string",
         "keyPrefix": "string"
      },
      "textDataDeliveryEnabled": boolean
   }
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

loggingConfig (p. 48)

The logging configuration values to set.

Type: LoggingConfig (p. 76) object

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

StopModelCustomizationJob

Service: Amazon Bedrock

Stops an active model customization job. For more information, see Custom models in the Bedrock User Guide.

Request Syntax

```
POST /model-customization-jobs/jobIdentifier/stop HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

jobIdentifier (p. 50)

Job identifier of the job to stop.

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12})|([a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*)$`

Required: Yes

Request Body

The request does not have a request body.

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

ConflictException

Error occurred because of a conflict while performing an operation.

HTTP Status Code: 400

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

TagResource

Service: Amazon Bedrock

Associate tags with a resource. For more information, see Tagging resources in the Bedrock User Guide.

Request Syntax

```
POST /tagResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string",
   "tags": [
      {
         "key": "string",
         "value": "string"
      }
   ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN (p. 52)

The ARN of the resource to tag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `(^[a-zA-Z0-9][a-zA-Z0-9\-]*$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}|)((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12})$)|(:provisioned-model/[a-z0-9]{12}$)))`

Required: Yes

tags (p. 52)

Tags to associate with the resource.

Type: Array of Tag (p. 84) objects

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

TooManyTagsException

The request contains more tags than can be associated with a resource (50 tags per resource). The maximum number of tags includes both existing tags and those included in your current request.

HTTP Status Code: 400

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UntagResource

Service: Amazon Bedrock

Remove one or more tags from a resource. For more information, see Tagging resources in the Bedrock User Guide.

Request Syntax

```
POST /untagResource HTTP/1.1
Content-type: application/json

{
   "resourceARN": "string",
   "tagKeys": [ "string" ]
}
```

URI Request Parameters

The request does not use any URI parameters.

Request Body

The request accepts the following data in JSON format.

resourceARN (p. 54)

The ARN of the resource to untag.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `(^[a-zA-Z0-9][a-zA-Z0-9\-]*$)|(^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:([0-9]{12}|)((:(fine-tuning-job|model-customization-job|custom-model)/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12})$)|(:provisioned-model/[a-z0-9]{12}$)))`

Required: Yes

tagKeys (p. 54)

Tag keys of the tags to remove from the resource.

Type: Array of strings

Array Members: Minimum number of 0 items. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Required: Yes

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateProvisionedModelThroughput

Service: Amazon Bedrock

Update a provisioned throughput. For more information, see Provisioned throughput in the Bedrock User Guide.

Request Syntax

```
PATCH /provisioned-model-throughput/provisionedModelId HTTP/1.1
Content-type: application/json

{
   "desiredModelId": "string",
   "desiredProvisionedModelName": "string"
}
```

URI Request Parameters

The request uses the following URI parameters.

provisionedModelId (p. 56)

The ARN or name of the provisioned throughput to update.

Pattern: `^((([0-9a-zA-Z][_-]?)+)|(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}))$`

Required: Yes

Request Body

The request accepts the following data in JSON format.

desiredModelId (p. 56)

The ARN of the new model to associate with this provisioned throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}(([:][a-z0-9-]{1,63}){0,2})?/[a-z0-9]{12})|(:foundation-model/([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})))|(([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2}))|(([0-9a-zA-Z][_-]?)+)$`

Required: No

desiredProvisionedModelName (p. 56)

The new name for this provisioned throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

Required: No

Response Syntax

```
HTTP/1.1 200
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Amazon Bedrock Runtime

The following actions are supported by Amazon Bedrock Runtime:

InvokeModel

Service: Amazon Bedrock Runtime

Invokes the specified Bedrock model to run inference using the input provided in the request body. You use InvokeModel to run inference for text models, image models, and embedding models.

For more information, see Run inference in the Bedrock User Guide.

For example requests, see Examples (after the Errors section).

Request Syntax

```
POST /model/modelId/invoke HTTP/1.1
Accept: accept
Content-Type: contentType

body
```

URI Request Parameters

The request uses the following URI parameters.

accept (p. 59)

The desired MIME type of the inference body in the response. The default value is `application/json`.

contentType (p. 59)

The MIME type of the input data in the request. The default value is `application/json`.

modelId (p. 59)

Identifier of the model.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|([0-9]{12}:provisioned-model/[a-z0-9]{12})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)$`

Required: Yes

Request Body

The request accepts the following binary data.

body (p. 59)

Input data in the format specified in the content-type request header. To see the format and content of this field for different models, refer to Inference parameters and to documentation from the model provider.

- Anthropic Claude – https://docs.anthropic.com/claude/reference/complete_post
- AI21 Labs Jurassic-2 – https://docs.ai21.com/reference/j2-complete-ref
- Cohere Command – https://docs.cohere.com/reference/generate

- Stability.ai Diffusion – https://docs.cohere.com/reference/generate

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: Yes

Response Syntax

```
HTTP/1.1 200
Content-Type: contentType

body
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType (p. 60)

The MIME type of the inference result.

The response returns the following as the HTTP body.

body (p. 60)

Inference response from the model in the format specified in the content-type header field. To see the format and content of this field for different models, refer to Inference parameters and to documentation from the model provider.

- Anthropic Claude – https://docs.anthropic.com/claude/reference/complete_post
- AI21 Labs Jurassic-2 – https://docs.ai21.com/reference/j2-complete-ref
- Cohere Command – https://docs.cohere.com/reference/generate
- Stability.ai Diffusion – https://docs.cohere.com/reference/generate

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests.

HTTP Status Code: 429

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Run inference on a text model

Send an invoke request to run inference on a Titan Text G1 - Express model. We set the `accept` parameter to accept any content type in the response.

```
POST https://bedrock.us-east-1.amazonaws.com/model/amazon.titan-text-express-v1/invoke

-H accept: */*
-H content-type: application/json

Payload
{"inputText": "Hello world"}
```

Example response

Response for the above request.

```
-H content-type: application/json

Payload
<the  model response>
```

Run inference on an image model

In the following example, the request sets the `accept` parameter to `image/png`.

```
POST https://bedrock.us-east-1.amazonaws.com/model/stability.stable-diffusion-xl-v0/invoke

-H accept: image/png
-H content-type: application/json

Payload
{"inputText": "Picture of a bird"}
```

Example response

Response for the above example.

```
-H content-type: image/png

Payload
<image bytes>
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

InvokeModelWithResponseStream

Service: Amazon Bedrock Runtime

Invoke the specified Bedrock model to run inference using the input provided. Return the response in a stream.

To find out if a model supports streaming, call GetFoundationModel and check the `responseStreamingSupported` field in the response.

For more information, see Run inference in the Bedrock User Guide.

For an example request and response, see Examples (after the Errors section).

Request Syntax

```
POST /model/modelId/invoke-with-response-stream HTTP/1.1
X-Amzn-Bedrock-Accept: accept
Content-Type: contentType

body
```

URI Request Parameters

The request uses the following URI parameters.

accept (p. 63)

The desired MIME type of the inference body in the response. The default value is `application/json`.

contentType (p. 63)

The MIME type of the input data in the request. The default value is `application/json`.

modelId (p. 63)

Id of the model to invoke using the streaming request.

Length Constraints: Minimum length of 1. Maximum length of 2048.

Pattern: `^(arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|([0-9]{12}:provisioned-model/[a-z0-9]{12})))|([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([.]?[a-z0-9-]{1,63})([:][a-z0-9-]{1,63}){0,2})|(([0-9a-zA-Z][_-]?)+)$`

Required: Yes

Request Body

The request accepts the following binary data.

body (p. 63)

Inference input in the format specified by the content-type. To see the format and content of this field for different models, refer to Inference parameters and to documentation from the model provider.

- Anthropic Claude – https://docs.anthropic.com/claude/reference/complete_post
- AI21 Labs Jurassic-2 – https://docs.ai21.com/reference/j2-complete-ref
- Cohere Command – https://docs.cohere.com/reference/generate
- Stability.ai Diffusion – https://docs.cohere.com/reference/generate

Length Constraints: Minimum length of 0. Maximum length of 25000000.

Required: Yes

Response Syntax

```
HTTP/1.1 200
X-Amzn-Bedrock-Content-Type: contentType
Content-type: application/json

{
   "chunk": {
      "bytes": blob
   },
   "internalServerException": {
   },
   "modelStreamErrorException": {
   },
   "throttlingException": {
   },
   "validationException": {
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The response returns the following HTTP headers.

contentType (p. 64)

The MIME type of the inference result.

The following data is returned in JSON format by the service.

chunk (p. 64)

Content included in the response.

Type: PayloadPart (p. 91) object

internalServerException (p. 64)

An internal server error occurred. Retry your request.

Type: Exception
HTTP Status Code: 500

modelStreamErrorException (p. 64)

An error occurred while streaming the response.

Type: Exception

HTTP Status Code: 424

throttlingException (p. 64)

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception
HTTP Status Code: 429

validationException (p. 64)

Input validation failed. Check your request parameters and retry the request.

Type: Exception
HTTP Status Code: 400

Errors

For information about the errors that are common to all actions, see Common Errors (p. 94).

AccessDeniedException

The request is denied because of missing access permissions.

HTTP Status Code: 403

InternalServerException

An internal server error occurred. Retry your request.

HTTP Status Code: 500

ModelErrorException

The request failed due to an error while processing the model.

HTTP Status Code: 424

ModelNotReadyException

The model specified in the request is not ready to serve inference requests.

HTTP Status Code: 429

ModelStreamErrorException

An error occurred while streaming the response.

HTTP Status Code: 424

ModelTimeoutException

The request took too long to process. Processing time exceeded the model timeout length.

HTTP Status Code: 408

ResourceNotFoundException

The specified resource ARN was not found. Check the ARN and try your request again.

HTTP Status Code: 404

ServiceQuotaExceededException

The number of requests exceeds the service quota. Resubmit your request later.

HTTP Status Code: 400

ThrottlingException

The number of requests exceeds the limit. Resubmit your request later.

HTTP Status Code: 429

ValidationException

Input validation failed. Check your request parameters and retry the request.

HTTP Status Code: 400

Examples

Run inference with streaming on a text model

For streaming, you can set `x-amzn-bedrock-accept-type` in the header to contain the desired content type of the response. In this example, we set it to accept any content type. The default value is `application/json`.

```
POST https://bedrock.us-east-1.amazonaws.com/model/amazon.titan-text-express-v1/invoke-
with-response-stream

-H accept: application/vnd.amazon.eventstream
-H content-type: application/json
-H x-amzn-bedrock-accept: */*

Payload
{"inputText": "Hello world"}
```

Example response

For streaming, the content type in the response is always set to `application/vnd.amazon.eventstream`. The response includes an additional header (x-amzn-bedrock-content-type), which contains the actual content type of the response.

```
-H content-type: application/vnd.amazon.eventstream
-H x-amzn-bedrock-content-type: application/json

Payload (stream events)
<response chunk>
```

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Data Types

The following data types are supported by Amazon Bedrock:

The following data types are supported by Amazon Bedrock Runtime:

Amazon Bedrock

The following data types are supported by Amazon Bedrock:

CloudWatchConfig

Service: Amazon Bedrock

CloudWatch logging configuration.

Contents

logGroupName

The log group name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 512.

Required: Yes

roleArn

The role ARN.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 2048.

Pattern: `^arn:aws(-[^:]+)?:iam::([0-9]{12})?:role/.+$`

Required: Yes

largeDataDeliveryS3Config

S3 configuration for delivering a large amount of data.

Type: S3Config (p. 83) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomModelSummary

Service: Amazon Bedrock

Summary information for a custom model.

Contents

baseModelArn

The base model ARN.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

Required: Yes

baseModelName

The base model name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63})$`

Required: Yes

creationTime

Creation time of the model.

Type: Timestamp

Required: Yes

modelArn

The ARN of the custom model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

Required: Yes

modelName

The name of the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FoundationModelDetails

Service: Amazon Bedrock

Information about a foundation model.

Contents

modelArn

The model ARN.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

Required: Yes

modelId

The model identifier.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)$`

Required: Yes

customizationsSupported

The customization that the model supports.

Type: Array of strings

Valid Values: `FINE_TUNING`

Required: No

inferenceTypesSupported

The inference types that the model supports.

Type: Array of strings

Valid Values: `ON_DEMAND | PROVISIONED`

Required: No

inputModalities

The input modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

modelName

The model name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^.*$`

Required: No

outputModalities

The output modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

providerName

he model's provider name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^.*$`

Required: No

responseStreamingSupported

Indicates whether the model supports streaming.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

FoundationModelSummary

Service: Amazon Bedrock

Summary information for a foundation model.

Contents

modelArn

The ARN of the foundation model.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

Required: Yes

modelId

The model Id of the foundation model.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 140.

Pattern: `^[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}(/[a-z0-9]{12}|)$`

Required: Yes

customizationsSupported

Whether the model supports fine-tuning or continual pre-training.

Type: Array of strings

Valid Values: `FINE_TUNING`

Required: No

inferenceTypesSupported

The inference types that the model supports.

Type: Array of strings

Valid Values: `ON_DEMAND | PROVISIONED`

Required: No

inputModalities

The input modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

modelName

The name of the model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^.*$`

Required: No

outputModalities

The output modalities that the model supports.

Type: Array of strings

Valid Values: `TEXT | IMAGE | EMBEDDING`

Required: No

providerName

The model's provider name.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 20.

Pattern: `^.*$`

Required: No

responseStreamingSupported

Indicates whether the model supports streaming.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

LoggingConfig

Service: Amazon Bedrock

Configuration fields for invokation logging.

Contents

cloudWatchConfig

CloudWatch logging configuration.

Type: CloudWatchConfig (p. 69) object

Required: No

embeddingDataDeliveryEnabled

Set to include embeddings data in the log delivery.

Type: Boolean

Required: No

imageDataDeliveryEnabled

Set to include image data in the log delivery.

Type: Boolean

Required: No

s3Config

S3 configuration for storing log data.

Type: S3Config (p. 83) object

Required: No

textDataDeliveryEnabled

Set to include text data in the log delivery.

Type: Boolean

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ModelCustomizationJobSummary

Service: Amazon Bedrock

Information about one customization job

Contents

baseModelArn

ARN of the base model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

Required: Yes

creationTime

Creation time of the custom model.

Type: Timestamp

Required: Yes

jobArn

ARN of the customization job.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:model-customization-job/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

Required: Yes

jobName

Name of the customization job.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^[a-zA-Z0-9](-*[a-zA-Z0-9\+\-\.])*$`

Required: Yes

status

Status of the customization job.

Type: String

Valid Values: `InProgress | Completed | Failed | Stopping | Stopped`

Required: Yes

customModelArn

ARN of the custom model.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}/[a-z0-9]{12}$`

Required: No

customModelName

Name of the custom model.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

Required: No

endTime

Time that the customization job ended.

Type: Timestamp

Required: No

lastModifiedTime

Time that the customization job was last modified.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

OutputDataConfig

Service: Amazon Bedrock

S3 Location of the output data.

Contents

s3Uri

The S3 URI where the output data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ProvisionedModelSummary

Service: Amazon Bedrock

Set of fields associated with a provisioned throughput.

Contents

creationTime

The time that this provisioned throughput was created.

Type: Timestamp

Required: Yes

desiredModelArn

Desired model ARN.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

Required: Yes

desiredModelUnits

Desired model units.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

foundationModelArn

Foundation model ARN.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}::foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}$`

Required: Yes

lastModifiedTime

The time that this provisioned throughput was last modified.

Type: Timestamp

Required: Yes

modelArn

The ARN of the model associated with this provisioned throughput.

Type: String

Length Constraints: Minimum length of 20. Maximum length of 1011.

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:(([0-9]{12}:custom-model/[a-z0-9-]{1,63}[.]{1}[a-z0-9-]{1,63}/[a-z0-9]{12})|(:foundation-model/[a-z0-9-]{1,63}[.]{1}([a-z0-9-]{1,63}[.]){0,2}[a-z0-9-]{1,63}([:][a-z0-9-]{1,63}){0,2}))$`

Required: Yes

modelUnits

The number of model units allocated.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

provisionedModelArn

The ARN of the provisioned throughput.

Type: String

Pattern: `^arn:aws(-[^:]+)?:bedrock:[a-z0-9-]{1,20}:[0-9]{12}:provisioned-model/[a-z0-9]{12}$`

Required: Yes

provisionedModelName

The name of the provisioned throughput.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 63.

Pattern: `^([0-9a-zA-Z][_-]?)+$`

Required: Yes

status

Status of the provisioned throughput.

Type: String

Valid Values: `Creating | InService | Updating | Failed`

Required: Yes

commitmentDuration

Commitment duration for the provisioned throughput.

Type: String

Valid Values: `OneMonth | SixMonths`

Required: No

commitmentExpirationTime

Commitment expiration time for the provisioned throughput.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

S3Config

Service: Amazon Bedrock

S3 configuration for storing log data.

Contents

bucketName

S3 bucket name.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 63.

Required: Yes

keyPrefix

S3 prefix.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 1024.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Tag

Service: Amazon Bedrock

Definition of the key/value pair for a tag.

Contents

key

Key for the tag.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Required: Yes

value

Value for the tag.

Type: String

Length Constraints: Minimum length of 0. Maximum length of 256.

Pattern: `^[a-zA-Z0-9\s._:/=+@-]*$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingDataConfig

Service: Amazon Bedrock

S3 Location of the training data.

Contents

s3Uri

The S3 URI where the training data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TrainingMetrics

Service: Amazon Bedrock

Metrics associated with the custom job.

Contents

trainingLoss

Loss metric associated with the custom job.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidationDataConfig

Service: Amazon Bedrock

Array of up to 10 validators.

Contents

validators

Information about the validators.

Type: Array of Validator (p. 88) objects

Array Members: Minimum number of 0 items. Maximum number of 10 items.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Validator

Service: Amazon Bedrock

Information about a validator.

Contents

s3Uri

The S3 URI where the validation data is stored.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1024.

Pattern: `^s3://[a-z0-9][\.\-a-z0-9]{1,61}[a-z0-9](/.*)?$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidatorMetric

Service: Amazon Bedrock

The metric for the validator.

Contents

validationLoss

The validation loss associated with this validator.

Type: Float

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

VpcConfig

Service: Amazon Bedrock

VPC configuration.

Contents

securityGroupIds

VPC configuration security group Ids.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 5 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `^[-0-9a-zA-Z]+$`

Required: Yes

subnetIds

VPC configuration subnets.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 16 items.

Length Constraints: Minimum length of 0. Maximum length of 32.

Pattern: `^[-0-9a-zA-Z]+$`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Amazon Bedrock Runtime

The following data types are supported by Amazon Bedrock Runtime:

- PayloadPart (p. 91)
- ResponseStream (p. 92)

PayloadPart

Service: Amazon Bedrock Runtime

Payload content included in the response.

Contents

bytes

Base64-encoded bytes of payload data.

Type: Base64-encoded binary data object

Length Constraints: Minimum length of 0. Maximum length of 1000000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ResponseStream

Service: Amazon Bedrock Runtime

Definition of content in the response stream.

Contents

chunk

Content included in the response.

Type: PayloadPart (p. 91) object

Required: No

internalServerException

An internal server error occurred. Retry your request.

Type: Exception
HTTP Status Code: 500

Required: No

modelStreamErrorException

An error occurred while streaming the response.

Type: Exception
HTTP Status Code: 424

Required: No

throttlingException

The number of requests exceeds the limit. Resubmit your request later.

Type: Exception
HTTP Status Code: 429

Required: No

validationException

Input validation failed. Check your request parameters and retry the request.

Type: Exception
HTTP Status Code: 400

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Common Parameters

The following list contains the parameters that all actions use for signing Signature Version 4 requests with a query string. Any action-specific parameters are listed in the topic for that action. For more information about Signature Version 4, see Signing AWS API requests in the *IAM User Guide*.

Action

The action to be performed.

Type: string

Required: Yes

Version

The API version that the request is written for, expressed in the format YYYY-MM-DD.

Type: string

Required: Yes

X-Amz-Algorithm

The hash algorithm that you used to create the request signature.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Valid Values: `AWS4-HMAC-SHA256`

Required: Conditional

X-Amz-Credential

The credential scope value, which is a string that includes your access key, the date, the region you are targeting, the service you are requesting, and a termination string ("aws4_request"). The value is expressed in the following format: *access_key*/*YYYYMMDD*/*region*/*service*/aws4_request.

For more information, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-Date

The date that is used to create the signature. The format must be ISO 8601 basic format (YYYYMMDD'T'HHMMSS'Z'). For example, the following date time is a valid X-Amz-Date value: `20120325T120000Z`.

Condition: X-Amz-Date is optional for all requests; it can be used to override the date used for signing requests. If the Date header is specified in the ISO 8601 basic format, X-Amz-Date is not required. When X-Amz-Date is used, it always overrides the value of the Date header. For more information, see Elements of an AWS API request signature in the *IAM User Guide*.

Type: string

Required: Conditional

X-Amz-Security-Token

The temporary security token that was obtained through a call to AWS Security Token Service (AWS STS). For a list of services that support temporary security credentials from AWS STS, see AWS services that work with IAM in the *IAM User Guide*.

Condition: If you're using temporary security credentials from AWS STS, you must include the security token.

Type: string

Required: Conditional

X-Amz-Signature

Specifies the hex-encoded signature that was calculated from the string to sign and the derived signing key.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-SignedHeaders

Specifies all the HTTP headers that were included as part of the canonical request. For more information about specifying signed headers, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

Common Errors

This section lists the errors common to the API actions of all AWS services. For errors specific to an API action for this service, see the topic for that API action.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 400

IncompleteSignature

The request signature does not conform to AWS standards.

HTTP Status Code: 400

InternalFailure

The request processing has failed because of an unknown error, exception or failure.

HTTP Status Code: 500

InvalidAction

The action or operation requested is invalid. Verify that the action is typed correctly.

HTTP Status Code: 400

InvalidClientTokenId

The X.509 certificate or AWS access key ID provided does not exist in our records.

HTTP Status Code: 403

NotAuthorized

You do not have permission to perform this action.

HTTP Status Code: 400

OptInRequired

The AWS access key ID needs a subscription for the service.

HTTP Status Code: 403

RequestExpired

The request reached the service more than 15 minutes after the date stamp on the request or more than 15 minutes after the request expiration date (such as for pre-signed URLs), or the date stamp on the request is more than 15 minutes in the future.

HTTP Status Code: 400

ServiceUnavailable

The request has failed due to a temporary failure of the server.

HTTP Status Code: 503

ThrottlingException

The request was denied due to request throttling.

HTTP Status Code: 400

ValidationError

The input fails to satisfy the constraints specified by an AWS service.

HTTP Status Code: 400