



API Reference

SageMaker Training Session



API Version 2025-02-10

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

SageMaker Training Session: API Reference

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Services or capabilities described in AWS documentation might vary by Region. To see the differences applicable to the AWS European Sovereign Cloud Region, see the [AWS European Sovereign Cloud User Guide](#).

Table of Contents

Welcome	1
Actions	2
ExecuteSamplingOperation	3
Request Syntax	3
URI Request Parameters	3
Request Body	3
Response Syntax	4
Response Elements	4
Errors	4
See Also	5
ExecuteTrainingOperation	7
Request Syntax	7
URI Request Parameters	7
Request Body	7
Response Syntax	8
Response Elements	8
Errors	9
See Also	10
GetSamplingOperationResult	11
Request Syntax	11
URI Request Parameters	11
Request Body	11
Response Syntax	11
Response Elements	12
Errors	12
See Also	13
GetTrainingOperationResult	15
Request Syntax	15
URI Request Parameters	15
Request Body	15
Response Syntax	15
Response Elements	16
Errors	17
See Also	17

SampleTrainingSession	19
Request Syntax	19
URI Request Parameters	19
Request Body	19
Response Syntax	20
Response Elements	20
Errors	20
See Also	21
SampleTrainingSessionWithResponseStream	22
Request Syntax	22
URI Request Parameters	22
Request Body	22
Response Syntax	23
Response Elements	23
Errors	23
See Also	24
Data Types	25
AdamParamsConfig	27
Contents	27
See Also	28
ComputeLogprobsConfig	29
Contents	29
See Also	29
ComputeLogprobsResult	30
Contents	30
See Also	30
Datum	31
Contents	31
See Also	31
EncodedTextChunk	32
Contents	32
See Also	32
ForwardBackwardConfig	33
Contents	33
See Also	33
ForwardBackwardResult	35

Contents	35
See Also	35
ForwardConfig	37
Contents	37
See Also	37
ForwardResult	39
Contents	39
See Also	39
LoadStateConfig	41
Contents	41
See Also	41
LoadStateResult	42
Contents	42
See Also	42
LoadWeightsForSamplingConfig	43
Contents	43
See Also	43
LoadWeightsForSamplingResult	44
Contents	44
See Also	44
ModelInput	45
Contents	45
See Also	45
OperationConfig	46
Contents	46
See Also	47
OperationResult	48
Contents	48
See Also	49
OptimizerStepConfig	50
Contents	50
See Also	50
OptimizerStepResult	51
Contents	51
See Also	51
SampleConfig	52

Contents	52
See Also	53
SampledSequence	54
Contents	54
See Also	54
SampleOperationConfig	56
Contents	56
See Also	56
SampleOperationResult	58
Contents	58
See Also	58
SampleResponse	60
Contents	60
See Also	60
SampleTokenUsage	61
Contents	61
See Also	61
SamplingParams	62
Contents	62
See Also	63
SaveStateConfig	64
Contents	64
See Also	64
SaveStateResult	65
Contents	65
See Also	65
SaveWeightsConfig	66
Contents	66
See Also	66
SaveWeightsResult	67
Contents	67
See Also	67
TensorData	68
Contents	68
See Also	68
TokenLogprob	69

Contents	69
See Also	69
TrainingTokenUsage	70
Contents	70
See Also	70
Common Parameters	71
Common Error Types	74

Welcome

Amazon SageMaker Training Session Runtime provides APIs for executing training operations and sampling operations against a running training session. Use these APIs to submit asynchronous training steps (such as forward-backward passes, optimizer steps, and checkpoint saves) and to generate text samples from the model being trained.

This document was last published on October 16, 2026.

Actions

The following actions are supported:

- [ExecuteSamplingOperation](#)
- [ExecuteTrainingOperation](#)
- [GetSamplingOperationResult](#)
- [GetTrainingOperationResult](#)
- [SampleTrainingSession](#)
- [SampleTrainingSessionWithResponseStream](#)

ExecuteSamplingOperation

Submits an asynchronous sampling operation to generate text or compute log-probabilities from the model associated with the specified training session. The session must be in Running or Blocked state. Use `GetSamplingOperationResult` to poll for the result.

Request Syntax

```
POST /v1/training-sessions/trainingSessionName/sampling-operations HTTP/1.1
Content-type: application/json

{
  "operationConfig": { ... }
}
```

URI Request Parameters

The request uses the following URI parameters.

trainingSessionName

The name of the training session to sample from.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9-]+`

Required: Yes

Request Body

The request accepts the following data in JSON format.

operationConfig

The configuration for the sample operation. Specify either a sample configuration to generate text or a `computeLogprobs` configuration to compute token log-probabilities.

Type: [SampleOperationConfig](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

Response Syntax

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

{
  "operationId": "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.

[operationId](#)

The unique identifier for the submitted sample operation. Use this identifier with `GetSamplingOperationResult` to retrieve the operation status and result.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9-]+`

Errors

For information about the errors that are common to all actions, see [Common Error Types](#).

ConflictException

The training session is in a state that cannot accept new operations. The session must be in Running or Blocked state to accept requests.

HTTP Status Code: 409

InternalServerErrorException

An internal service error occurred. Retry the request using exponential backoff.

HTTP Status Code: 500

ResourceNotFoundException

The specified training session or operation does not exist, or the operation does not belong to the specified training session.

HTTP Status Code: 404

ThrottlingException

The request was throttled. Retry the request after a brief wait using exponential backoff.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters. Examples include invalid sequence numbers, malformed client tokens, missing required fields, or invalid operation configuration.

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 V2](#)
- [AWS SDK for .NET V4](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for Kotlin](#)
- [AWS SDK for PHP V3](#)

- [AWS SDK for Python \(Boto3\)](#)
- [AWS SDK for Ruby V3](#)

ExecuteTrainingOperation

Submits an asynchronous training operation to the specified training session. The service executes operations in sequence number order. The session must be in Running or Blocked state. Use `GetTrainingOperationResult` to poll for the result. Supported operation types: forward, forwardBackward, optimizerStep, saveWeights, saveState, and loadState.

Request Syntax

```
POST /v1/training-sessions/trainingSessionName/operations HTTP/1.1
Content-type: application/json

{
  "clientToken": "string",
  "operationConfig": { ... },
  "sequenceNumber": number
}
```

URI Request Parameters

The request uses the following URI parameters.

trainingSessionName

The name of the training session to submit the operation to.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: [a-zA-Z0-9-]+

Required: Yes

Request Body

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure that the operation completes no more than one time. If you retry a request with the same client token and the same parameters, the service ignores the request but does not return an error.

Type: String

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

Pattern: [a-zA-Z0-9-]+

Required: No

operationConfig

The configuration that specifies the type and parameters of the training operation. Exactly one member of the union must be set.

Type: [OperationConfig](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

Required: Yes

sequenceNumber

The monotonically increasing sequence number that determines execution order. The training session processes operations strictly in sequence number order. If a gap exists, the session blocks until the missing operation arrives or times out. Must be at least 1.

Type: Integer

Valid Range: Minimum value of 1.

Required: Yes

Response Syntax

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

{
  "operationId": "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.

operationId

The unique identifier for the submitted operation. Use this identifier with `GetTrainingOperationResult` to retrieve the operation status and result.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: [a-zA-Z0-9-]+

Errors

For information about the errors that are common to all actions, see [Common Error Types](#).

ConflictException

The training session is in a state that cannot accept new operations. The session must be in Running or Blocked state to accept requests.

HTTP Status Code: 409

InternalServerError

An internal service error occurred. Retry the request using exponential backoff.

HTTP Status Code: 500

ResourceNotFoundException

The specified training session or operation does not exist, or the operation does not belong to the specified training session.

HTTP Status Code: 404

ThrottlingException

The request was throttled. Retry the request after a brief wait using exponential backoff.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters. Examples include invalid sequence numbers, malformed client tokens, missing required fields, or invalid operation configuration.

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 V2](#)
- [AWS SDK for .NET V4](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for Kotlin](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python \(Boto3\)](#)
- [AWS SDK for Ruby V3](#)

GetSamplingOperationResult

Retrieves the current status and result of a previously submitted sampling operation. Poll this operation until the status reaches a terminal state (Completed or Failed).

Request Syntax

```
GET /v1/training-sessions/trainingSessionName/sampling-operations/operationId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

operationId

The unique identifier of the sample operation to retrieve, returned by `ExecuteSamplingOperation`.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9-]+`

Required: Yes

trainingSessionName

The name of the training session that owns the sample operation.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[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

{
  "errorDetails": "string",
  "operationResult": { ... },
  "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.

[errorDetails](#)

A message describing why the sample operation failed. The service populates this field only when the status is Failed.

Type: String

[operationResult](#)

The result payload of the sample operation. The service populates this field only when the status is Completed.

Type: [SampleOperationResult](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

[status](#)

The current status of the sample operation. Possible values are Pending, InProgress, Completed, or Failed.

Type: String

Valid Values: Pending | InProgress | Completed | Failed | Skipped

Errors

For information about the errors that are common to all actions, see [Common Error Types](#).

InternalServerErrorException

An internal service error occurred. Retry the request using exponential backoff.

HTTP Status Code: 500

ResourceNotFoundException

The specified training session or operation does not exist, or the operation does not belong to the specified training session.

HTTP Status Code: 404

ThrottlingException

The request was throttled. Retry the request after a brief wait using exponential backoff.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters. Examples include invalid sequence numbers, malformed client tokens, missing required fields, or invalid operation configuration.

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 V2](#)
- [AWS SDK for .NET V4](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for Kotlin](#)
- [AWS SDK for PHP V3](#)

- [AWS SDK for Python \(Boto3\)](#)
- [AWS SDK for Ruby V3](#)

GetTrainingOperationResult

Retrieves the current status and result of a previously submitted training operation. Poll this operation until the status reaches a terminal state (Completed, Failed, or Skipped).

Request Syntax

```
GET /v1/training-sessions/trainingSessionName/operations/operationId HTTP/1.1
```

URI Request Parameters

The request uses the following URI parameters.

operationId

The unique identifier of the operation to retrieve, returned by `ExecuteTrainingOperation`.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[a-zA-Z0-9-]+`

Required: Yes

trainingSessionName

The name of the training session that owns the operation.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: `[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
```

```
{
  "errorDetails": "string",
  "operationResult": { ... },
  "status": "string",
  "tokenUsage": {
    "trainingTokens": 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.

errorDetails

A message describing why the operation failed. The service populates this field only when the status is Failed.

Type: String

operationResult

The result payload of the operation. The service populates this field only when the status is Completed. The populated member of the union corresponds to the operation type that was submitted.

Type: [OperationResult](#) object

Note: This object is a Union. Only one member of this object can be specified or returned.

status

The current status of the operation. Possible values are Pending, InProgress, Completed, Failed, or Skipped.

Type: String

Valid Values: Pending | InProgress | Completed | Failed | Skipped

tokenUsage

The token usage statistics for metering purposes. The service populates this field for forward and forwardBackward operations upon completion.

Type: [TrainingTokenUsage](#) object

Errors

For information about the errors that are common to all actions, see [Common Error Types](#).

InternalServerError

An internal service error occurred. Retry the request using exponential backoff.

HTTP Status Code: 500

ResourceNotFoundException

The specified training session or operation does not exist, or the operation does not belong to the specified training session.

HTTP Status Code: 404

ThrottlingException

The request was throttled. Retry the request after a brief wait using exponential backoff.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters. Examples include invalid sequence numbers, malformed client tokens, missing required fields, or invalid operation configuration.

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 V2](#)
- [AWS SDK for .NET V4](#)
- [AWS SDK for C++](#)

- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for Kotlin](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python \(Boto3\)](#)
- [AWS SDK for Ruby V3](#)

SampleTrainingSession

Sends an inference request to the model associated with the specified training session and returns the response synchronously. The service passes the request and response bodies to and from the model without modification.

Request Syntax

```
POST /training-sessions/trainingSessionName/sample HTTP/1.1
Content-Type: contentType
Accept: accept
X-Amzn-SageMaker-Model-Package: modelPackageArn

body
```

URI Request Parameters

The request uses the following URI parameters.

[accept](#)

The desired MIME type of the inference body in the response.

[contentType](#)

The MIME type of the input data in the request body.

[modelPackageArn](#)

The ARN of the model package whose checkpoint to invoke.

[trainingSessionName](#)

The name of the training session whose model to invoke.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: [a-zA-Z0-9-]+

Required: Yes

Request Body

The request accepts the following binary data.

body

The raw inference request body.

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

The MIME type of the inference result in the response body.

The response returns the following as the HTTP body.

body

The raw inference response body from the model.

Errors

For information about the errors that are common to all actions, see [Common Error Types](#).

ConflictException

The training session is in a state that cannot accept new operations. The session must be in Running or Blocked state to accept requests.

HTTP Status Code: 409

InternalServerErrorException

An internal service error occurred. Retry the request using exponential backoff.

HTTP Status Code: 500

ResourceNotFoundException

The specified training session or operation does not exist, or the operation does not belong to the specified training session.

HTTP Status Code: 404

ThrottlingException

The request was throttled. Retry the request after a brief wait using exponential backoff.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters. Examples include invalid sequence numbers, malformed client tokens, missing required fields, or invalid operation configuration.

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 V2](#)
- [AWS SDK for .NET V4](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for Kotlin](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python \(Boto3\)](#)
- [AWS SDK for Ruby V3](#)

SampleTrainingSessionWithResponseStream

Streams inference responses from the model associated with the specified training session. Use this operation for real-time token-by-token generation. The service passes the request body to the model without modification.

Request Syntax

```
POST /training-sessions/trainingSessionName/sample-with-response-stream HTTP/1.1
Content-Type: contentType
Accept: accept
X-Amzn-SageMaker-Model-Package: modelPackageArn

body
```

URI Request Parameters

The request uses the following URI parameters.

accept

The desired MIME type of the inference body in the response.

contentType

The MIME type of the input data in the request body.

modelPackageArn

The ARN of the model package whose checkpoint to invoke.

trainingSessionName

The name of the training session whose model to invoke.

Length Constraints: Minimum length of 1. Maximum length of 64.

Pattern: [a-zA-Z0-9-]+

Required: Yes

Request Body

The request accepts the following binary data.

body

The raw inference request body.

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

The MIME type of the streaming inference result.

The response returns the following as the HTTP body.

body

The streaming inference response from the model, delivered as a series of chunks.

Errors

For information about the errors that are common to all actions, see [Common Error Types](#).

ConflictException

The training session is in a state that cannot accept new operations. The session must be in Running or Blocked state to accept requests.

HTTP Status Code: 409

InternalServerErrorException

An internal service error occurred. Retry the request using exponential backoff.

HTTP Status Code: 500

ResourceNotFoundException

The specified training session or operation does not exist, or the operation does not belong to the specified training session.

HTTP Status Code: 404

ThrottlingException

The request was throttled. Retry the request after a brief wait using exponential backoff.

HTTP Status Code: 429

ValidationException

The request is not valid. Check the request syntax and parameters. Examples include invalid sequence numbers, malformed client tokens, missing required fields, or invalid operation configuration.

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 V2](#)
- [AWS SDK for .NET V4](#)
- [AWS SDK for C++](#)
- [AWS SDK for Go v2](#)
- [AWS SDK for Java V2](#)
- [AWS SDK for JavaScript V3](#)
- [AWS SDK for Kotlin](#)
- [AWS SDK for PHP V3](#)
- [AWS SDK for Python \(Boto3\)](#)
- [AWS SDK for Ruby V3](#)

Data Types

The Amazon SageMaker Training Session Runtime API contains several data types that various actions use. This section describes each data type in detail.

Note

The order of each element in a data type structure is not guaranteed. Applications should not assume a particular order.

The following data types are supported:

- [AdamParamsConfig](#)
- [ComputeLogprobsConfig](#)
- [ComputeLogprobsResult](#)
- [Datum](#)
- [EncodedTextChunk](#)
- [ForwardBackwardConfig](#)
- [ForwardBackwardResult](#)
- [ForwardConfig](#)
- [ForwardResult](#)
- [LoadStateConfig](#)
- [LoadStateResult](#)
- [LoadWeightsForSamplingConfig](#)
- [LoadWeightsForSamplingResult](#)
- [ModelInput](#)
- [OperationConfig](#)
- [OperationResult](#)
- [OptimizerStepConfig](#)
- [OptimizerStepResult](#)
- [SampleConfig](#)
- [SampledSequence](#)

- [SampleOperationConfig](#)
- [SampleOperationResult](#)
- [SampleResponse](#)
- [SampleTokenUsage](#)
- [SamplingParams](#)
- [SaveStateConfig](#)
- [SaveStateResult](#)
- [SaveWeightsConfig](#)
- [SaveWeightsResult](#)
- [TensorData](#)
- [TokenLogprob](#)
- [TrainingTokenUsage](#)

AdamParamsConfig

The Adam optimizer hyperparameters used during an optimizer step.

Contents

learningRate

The learning rate for the optimizer. Must be greater than 0.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 0.01.

Required: Yes

beta1

The exponential decay rate for the first moment estimates.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 0.999999.

Required: No

beta2

The exponential decay rate for the second moment estimates.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 0.999999.

Required: No

eps

A small constant added to the denominator for numerical stability.

Type: Double

Valid Range: Minimum value of 1.0e-16. Maximum value of 0.01.

Required: No

gradClipNorm

The maximum gradient norm for gradient clipping. Gradients are scaled down if their norm exceeds this value.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 100.

Required: No

loraPlusLrRatio

The learning rate multiplier for LoRA+ B matrices relative to A matrices. When set, B matrices use `learningRate * loraPlusLrRatio` while A matrices use the base learning rate.

Type: Double

Valid Range: Minimum value of 1.0. Maximum value of 100.0.

Required: No

weightDecay

The weight decay (L2 regularization) coefficient.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

ComputeLogprobsConfig

The configuration for computing log-probabilities of prompt tokens without generating new text.

Contents

prompt

The input prompt whose token log-probabilities will be computed.

Type: [ModelInput](#) object

Required: Yes

modelPackageArn

The ARN of the model package whose checkpoint this operation should run against. Must be a valid SageMaker model package ARN in the format `arn:aws:sagemaker:REGION:ACCOUNT:model-package/PACKAGE_NAME/VERSION`. When absent, the operation runs against the session's latest checkpoint.

Type: String

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

ComputeLogprobsResult

The result of a compute-logprobs operation.

Contents

logprobs

The per-token log-probabilities for the input prompt.

Type: Array of doubles

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

Datum

A single training datum containing model input and optional loss function inputs.

Contents

modelInput

The tokenized input to feed to the model.

Type: [ModelInput](#) object

Required: Yes

lossFnInputs

The tensor inputs for the loss function, keyed by name. Unsupported keys are rejected. Supported keys: target_tokens, weights, advantages, logprobs, ref_log_prob, temperature. Required keys depend on the loss function type: SFT (cross_entropy) requires target_tokens and weights; RL (ppo, cispo, importance_sampling) requires target_tokens, weights, advantages, and logprobs.

Type: String to [TensorData](#) object map

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

EncodedTextChunk

A chunk of pre-tokenized text represented as a list of token IDs.

Contents

tokens

The token IDs representing the encoded text.

Type: Array of integers

Required: Yes

type

The type of the chunk encoding.

Type: String

Valid Values: encoded_text

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

ForwardBackwardConfig

The configuration for a forward and backward pass through the model. This computes both loss values and gradients for use in subsequent optimizer steps.

Contents

datumList

The list of input data to feed through the model. Must contain at least one datum, and each datum must have a non-empty modelInput with at least one chunk.

Type: Array of [Datum](#) objects

Required: Yes

lossFnName

The name of the loss function to evaluate during the forward-backward pass.

Type: String

Valid Values: `cross_entropy` | `importance_sampling` | `ppo` | `cispo`

Required: Yes

lossFnConfig

Additional loss function configuration passed to the engine. Supported keys are `clip_low_threshold` (valid range: 0 to 1) and `clip_high_threshold` (valid range: 1 to 10). Unsupported keys are rejected.

Type: JSON value

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

ForwardBackwardResult

The result of a forward and backward pass through the model, including computed gradients.

Contents

lossFnOutputs

The loss function output tensors, keyed by name.

Type: Array of string to [TensorData](#) object maps

Required: No

lossFnOutputType

The data type of the loss function output tensors.

Type: String

Required: No

metrics

The training metrics emitted by the engine, such as loss values.

Type: String to double map

Required: No

tokenUsage

The token usage statistics for metering.

Type: [TrainingTokenUsage](#) 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 Java V2](#)
- [AWS SDK for Ruby V3](#)

ForwardConfig

The configuration for a forward-only pass through the model. This computes loss values without computing gradients.

Contents

datumList

The list of input data to feed through the model. Must contain at least one datum, and each datum must have a non-empty modelInput with at least one chunk.

Type: Array of [Datum](#) objects

Required: Yes

lossFnName

The name of the loss function to evaluate during the forward pass.

Type: String

Valid Values: `cross_entropy` | `importance_sampling` | `ppo` | `cispo`

Required: Yes

lossFnConfig

Additional loss function configuration passed to the engine. Supported keys are `clip_low_threshold` (valid range: 0 to 1) and `clip_high_threshold` (valid range: 1 to 10). Unsupported keys are rejected.

Type: JSON value

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

ForwardResult

The result of a forward-only pass through the model.

Contents

lossFnOutputs

The loss function output tensors, keyed by name.

Type: Array of string to [TensorData](#) object maps

Required: No

lossFnOutputType

The data type of the loss function output tensors.

Type: String

Required: No

metrics

The training metrics emitted by the engine, such as loss values.

Type: String to double map

Required: No

tokenUsage

The token usage statistics for metering.

Type: [TrainingTokenUsage](#) 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 Java V2](#)
- [AWS SDK for Ruby V3](#)

LoadStateConfig

The configuration for loading training state from a previously saved checkpoint. This restores model weights and optionally optimizer state.

Contents

loadOptimizer

Specifies whether to also restore the optimizer state alongside model weights. When true, optimizer state is restored to resume training from where it left off.

Type: Boolean

Required: Yes

modelPackageArn

The ARN of the model package containing the checkpoint to load. Must be a valid SageMaker model package ARN in the format `arn:aws:sagemaker:REGION:ACCOUNT:model-package/PACKAGE_NAME/VERSION`. Model package group ARNs are not accepted.

Type: String

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

LoadStateResult

The result of loading training state. An empty response indicates success.

Contents

The members of this exception structure are context-dependent.

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

LoadWeightsForSamplingConfig

The configuration for loading model weights for inference and sampling without restoring optimizer state.

Contents

modelPackageArn

The ARN of the model package containing the weights to load for sampling. Must be a valid SageMaker model package ARN in the format `arn:aws:sagemaker:REGION:ACCOUNT:model-package/PACKAGE_NAME/VERSION`. The model package must belong to the same AWS account as the training session.

Type: String

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

LoadWeightsForSamplingResult

The result of loading weights for sampling. An empty response indicates success.

Contents

The members of this exception structure are context-dependent.

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

ModelInput

The tokenized model input composed of ordered encoded text chunks.

Contents

chunks

The ordered list of encoded text chunks forming the input sequence.

Type: Array of [EncodedTextChunk](#) objects

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

OperationConfig

The configuration for a training operation. Exactly one member must be set to specify the operation type and its parameters.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

forward

The configuration for a forward-only pass through the model without gradient computation.

Type: [ForwardConfig](#) object

Required: No

forwardBackward

The configuration for a forward and backward pass that computes gradients.

Type: [ForwardBackwardConfig](#) object

Required: No

loadState

The configuration for loading training state from a previously saved checkpoint.

Type: [LoadStateConfig](#) object

Required: No

optimizerStep

The configuration for an optimizer step that updates model weights using accumulated gradients.

Type: [OptimizerStepConfig](#) object

Required: No

saveState

The configuration for saving full training state, including model weights and optimizer state.

Type: [SaveStateConfig](#) object

Required: No

saveWeights

The configuration for saving model weights as a checkpoint.

Type: [SaveWeightsConfig](#) 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 Java V2](#)
- [AWS SDK for Ruby V3](#)

OperationResult

The result of a completed training operation. Exactly one member is populated, corresponding to the operation type that was submitted.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

forward

The result from a forward-only pass.

Type: [ForwardResult](#) object

Required: No

forwardBackward

The result from a forward and backward pass.

Type: [ForwardBackwardResult](#) object

Required: No

loadState

The result from loading training state.

Type: [LoadStateResult](#) object

Required: No

optimizerStep

The result from an optimizer step.

Type: [OptimizerStepResult](#) object

Required: No

saveState

The result from saving full training state.

Type: [SaveStateResult](#) object

Required: No

saveWeights

The result from saving model weights.

Type: [SaveWeightsResult](#) 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 Java V2](#)
- [AWS SDK for Ruby V3](#)

OptimizerStepConfig

The configuration for an optimizer step that updates model weights using gradients accumulated from previous forward-backward passes.

Contents

adamParamsConfig

The Adam optimizer hyperparameters for this step.

Type: [AdamParamsConfig](#) object

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

OptimizerStepResult

The result of an optimizer step.

Contents

metrics

The training metrics emitted by the engine after the optimizer step.

Type: String to double map

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SampleConfig

The configuration for generating text samples from the model.

Contents

numSamples

The number of independent samples to generate for the prompt.

Type: Integer

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

Required: Yes

prompt

The input prompt to send to the model for text generation.

Type: [ModelInput](#) object

Required: Yes

samplingParams

The parameters controlling the sampling strategy.

Type: [SamplingParams](#) object

Required: Yes

includePromptLogprobs

Specifies whether to include per-token log-probabilities for the input prompt in the response.

Type: Boolean

Required: No

modelPackageArn

The ARN of the model package whose checkpoint this sample should run against. Must be a valid SageMaker model package ARN in the format

`arn:aws:sagemaker:REGION:ACCOUNT:model-package/PACKAGE_NAME/VERSION`. When absent, the sample runs against the session's latest checkpoint.

Type: String

Required: No

topkPromptLogprobs

The number of top-k log-probabilities to return per prompt token.

Type: Integer

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SampledSequence

A single generated sequence from the model.

Contents

stopReason

The reason why generation stopped. A value of `stop` indicates a stop sequence was encountered. A value of `length` indicates the maximum token limit was reached.

Type: String

Valid Values: `stop` | `length`

Required: Yes

tokens

The token IDs of the generated sequence.

Type: Array of integers

Required: Yes

logprobs

The per-token log-probabilities for the generated tokens.

Type: Array of doubles

Required: No

tokenUsage

The token usage statistics for this individual sample.

Type: [SampleTokenUsage](#) 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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SampleOperationConfig

The configuration for a sampling operation. Exactly one member must be set to specify the operation type.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

computeLogprobs

The configuration for computing log-probabilities of prompt tokens without generating new text.

Type: [ComputeLogprobsConfig](#) object

Required: No

loadWeightsForSampling

The configuration for loading model weights for inference and sampling without restoring optimizer state.

Type: [LoadWeightsForSamplingConfig](#) object

Required: No

sample

The configuration for generating text samples from the model.

Type: [SampleConfig](#) 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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SampleOperationResult

The result of a completed sampling operation. Exactly one member is populated, corresponding to the operation type that was submitted.

Contents

Important

This data type is a UNION, so only one of the following members can be specified when used or returned.

computeLogprobs

The computed log-probabilities for the prompt tokens.

Type: [ComputeLogprobsResult](#) object

Required: No

loadWeightsForSampling

The result from loading weights for sampling.

Type: [LoadWeightsForSamplingResult](#) object

Required: No

sample

The generated text samples from the model.

Type: [SampleResponse](#) 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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SampleResponse

The response containing generated samples and optional log-probabilities.

Contents

sequences

The list of generated sequences.

Type: Array of [SampledSequence](#) objects

Required: Yes

promptLogprobs

The per-token log-probabilities for the input prompt. Populated only when `includePromptLogprobs` is true.

Type: Array of doubles

Required: No

topkPromptLogprobs

The top-k log-probabilities per prompt token, as a list of lists. Populated only when `topkPromptLogprobs` is greater than 0.

Type: Array of arrays of [TokenLogprob](#) objects

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SampleTokenUsage

The token usage statistics for a sample generation.

Contents

decode

The number of tokens generated (decode phase).

Type: Integer

Required: Yes

prefill

The number of tokens in the input prompt (prefill phase).

Type: Integer

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SamplingParams

The parameters controlling the text generation sampling strategy.

Contents

maxTokens

The maximum number of tokens to generate per sample.

Type: Integer

Valid Range: Minimum value of 512. Maximum value of 131072.

Required: Yes

seed

The random seed for reproducible sampling. When set, the same seed with the same input produces deterministic outputs.

Type: Integer

Required: No

stop

The list of stop sequences that terminate generation when encountered in the output.

Type: Array of strings

Required: No

temperature

The sampling temperature controlling randomness. Higher values produce more diverse outputs.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 2.

Required: No

topK

The number of highest-probability vocabulary tokens to keep for top-k filtering. Set to -1 to disable.

Type: Integer

Required: No

topP

The cumulative probability threshold for nucleus (top-p) sampling. Only tokens with cumulative probability below this threshold are considered.

Type: Double

Valid Range: Minimum value of 0. Maximum value of 1.

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SaveStateConfig

The configuration for saving full training state, including model weights and optimizer state. Use this to create resumable checkpoints.

Contents

description

A human-readable description for the saved state checkpoint.

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SaveStateResult

The result of saving full training state.

Contents

modelPackageArn

The ARN of the model package created for the saved state checkpoint.

Type: String

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SaveWeightsConfig

The configuration for saving model weights as a checkpoint.

Contents

description

A human-readable description for the saved checkpoint.

Type: String

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

Required: No

mergeWeights

Specifies whether to merge LoRA adapter weights into the base model before saving. When true, the saved checkpoint contains the full merged model weights.

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

SaveWeightsResult

The result of saving model weights.

Contents

modelPackageArn

The ARN of the model package created for the saved weights.

Type: String

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

TensorData

A tensor with data, data type, and shape metadata.

Contents

data

The raw tensor data as JSON. The format is determined by the `dtype` field: integer arrays for `int64`, floating-point arrays for `float32`.

Type: JSON value

Required: Yes

dtype

The data type of the tensor elements, such as `int64` or `float32`.

Type: String

Required: Yes

shape

The dimensions of the tensor, such as `[batch_size, seq_length]`.

Type: Array of integers

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

TokenLogprob

The log-probability for a single token.

Contents

logprob

The log-probability of the token.

Type: Double

Required: Yes

tokenId

The vocabulary token ID.

Type: Integer

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 Java V2](#)
- [AWS SDK for Ruby V3](#)

TrainingTokenUsage

The token usage statistics for training operations.

Contents

trainingTokens

The number of training tokens consumed by the operation.

Type: Integer

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 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*.

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 Error Types

This section lists common error types that this AWS service may return. Not all services return all error types listed here. For errors specific to an API action for this service, see the topic for that API action.

AccessDeniedException

You don't have permission to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 403

ExpiredTokenException

The security token included in the request has expired. Request a new security token and try again.

HTTP Status Code: 403

IncompleteSignature

The request signature doesn't conform to AWS standards. Verify that you're using valid AWS credentials and that your request is properly formatted. If you're using an SDK, ensure it's up to date.

HTTP Status Code: 403

InternalFailure

The request can't be processed right now because of an internal server issue. Try again later. If the problem persists, contact AWS Support.

HTTP Status Code: 500

MalformedHttpRequestException

The request body can't be processed. This typically happens when the request body can't be decompressed using the specified content encoding algorithm. Verify that the content encoding header matches the compression format used.

HTTP Status Code: 400

NotAuthorized

You don't have permissions to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 401

OptInRequired

Your AWS account needs a subscription for this service. Verify that you've enabled the service in your account.

HTTP Status Code: 403

RequestAbortedException

The request was aborted before a response could be returned. This typically happens when the client closes the connection.

HTTP Status Code: 400

RequestEntityTooLargeException

The request entity is too large. Reduce the size of the request body and try again.

HTTP Status Code: 413

RequestTimeoutException

The request timed out. The server didn't receive the complete request within the expected time frame. Try again.

HTTP Status Code: 408

ServiceUnavailable

The service is temporarily unavailable. Try again later.

HTTP Status Code: 503

ThrottlingException

Your request rate is too high. The AWS SDKs automatically retry requests that receive this exception. Reduce the frequency of requests.

HTTP Status Code: 400

UnknownOperationException

The action or operation isn't recognized. Verify that the action name is spelled correctly and that it's supported by the API version you're using.

HTTP Status Code: 404

UnrecognizedClientException

The X.509 certificate or AWS access key ID you provided doesn't exist in our records. Verify that you're using valid credentials and that they haven't expired.

HTTP Status Code: 403

ValidationError

The input doesn't meet the required format or constraints. Check that all required parameters are included and that values are valid.

HTTP Status Code: 400