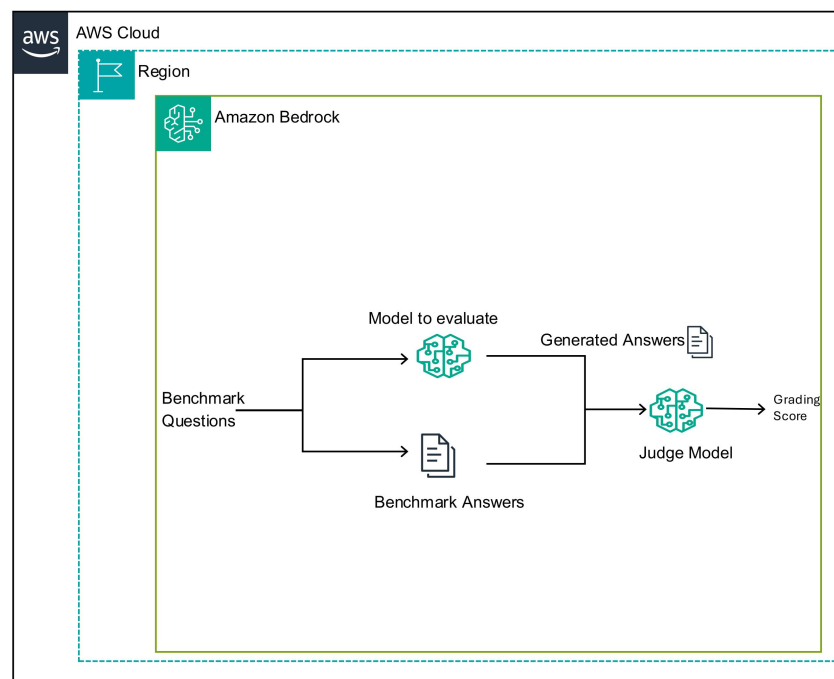




A Step-by-Step Guide to Setting Up and Evaluating Custom Models in Amazon Bedrock



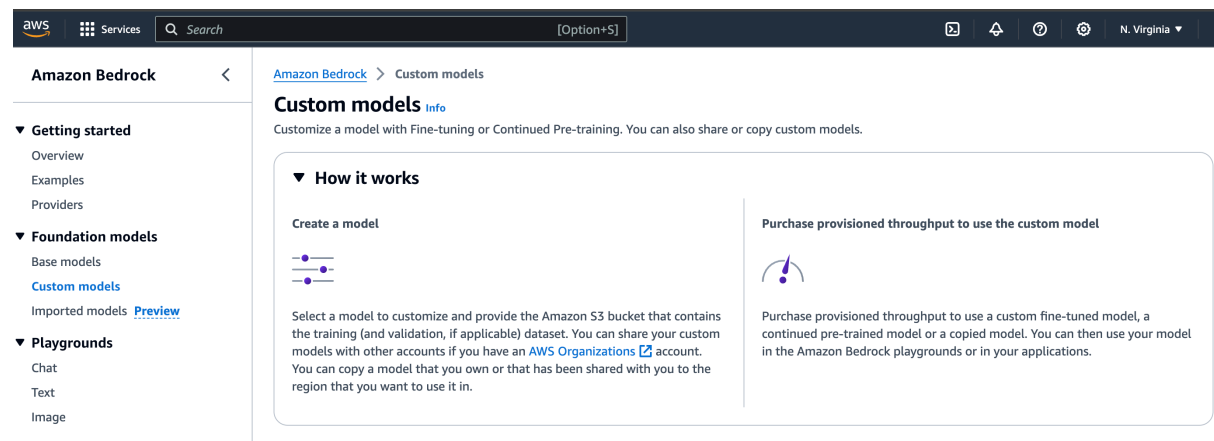
By
Mahendran Selvakumar
<https://devopstronaut.com/>



Note: We will not create this custom model due to the HIGH COSTS associated with it. Please do not create one, even if you are practicing. However, we will discuss how to create and evaluate a custom model.

1. Creating Custom Foundation Models

Go to Amazon Bedrock and click on **Custom Models**




Amazon Bedrock > Custom models

Custom models Info

Customize a model with Fine-tuning or Continued Pre-training. You can also share or copy custom models.


How it works

Create a model



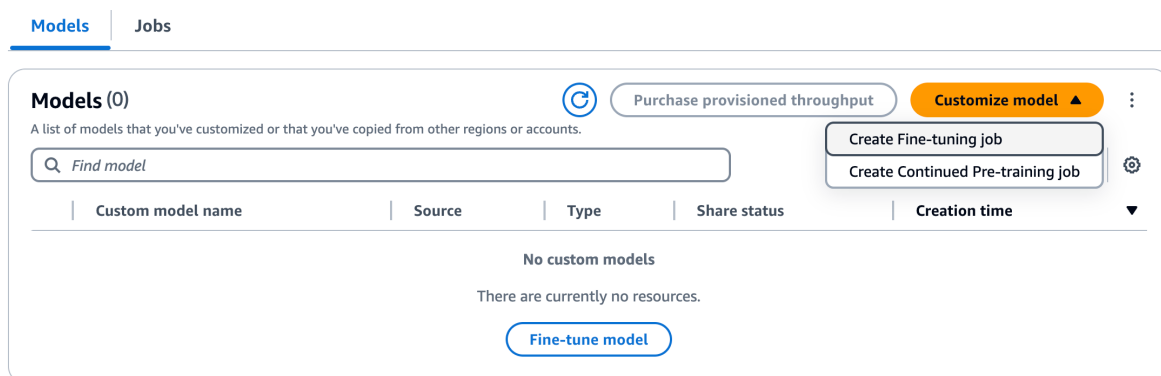
Select a model to customize and provide the Amazon S3 bucket that contains the training (and validation, if applicable) dataset. You can share your custom models with other accounts if you have an [AWS Organizations](#) account. You can copy a model that you own or that has been shared with you to the region that you want to use it in.

Purchase provisioned throughput to use the custom model



Purchase provisioned throughput to use a custom fine-tuned model, a continued pre-trained model or a copied model. You can then use your model in the Amazon Bedrock playgrounds or in your applications.

Choose **Create Fine-tuning Job** for a one-time job under **Customize Model**



Models (0)

A list of models that you've customized or that you've copied from other regions or accounts.

Customize model ▲

- Create Fine-tuning job
- Create Continued Pre-training job

Custom model name	Source	Type	Share status	Creation time
No custom models				

There are currently no resources.

[Fine-tune model](#)

Click **Select Model** under **Source Model**



Create Fine-tuning job [Info](#)

Select the model you wish to fine-tune and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

[Select model](#)

Fine-tuned model name

Enter a name to identify the new fine-tuned model.

☐ Model encryption [Info](#)

► **Tags - optional**

Choose the models available for fine-tuning and click **Apply**



Select model

1. Category

Model providers

 Amazon

 Cohere

 Meta

2. Models available for fine-tuning

Titan Image Generator G1 v1

Image model

Titan Image Generator G1 v2

Image model

Titan Text G1 - Lite v1

Text model

Titan Text G1 - Express v1

Text model

Titan Multimodal Embeddings G1 v1

Embedding model

Cancel

Apply

Provide the **Fine-tuned Model Name** and the **Job Name**



Create Fine-tuning job [Info](#)

Select the model you wish to fine-tune and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

 **Titan Text G1 - Express v1**
[Change](#)

Fine-tuned model name

Enter a name to identify the new fine-tuned model.

☐ Model encryption [Info](#)

► [Tags - optional](#)

Job configuration

Job name

Enter a name to identify the training job necessary to pre-train and create a new model.

► [Tags - optional](#)

To fine-tune the model, we need input data, which should be provided via Amazon S3

► VPC settings - optional

Choose a VPC configuration to access Amazon S3 data source located in your virtual private cloud (VPC). You can create and manage VPC, subnets and security groups in [Amazon VPC](#) [↗](#)

Input data [Info](#)

Choose a file in the S3 location. The files you choose must be in the [dataset format](#) [↗](#) that the model needs for training. You can also use SageMaker Ground Truth to create and label training datasets. [Learn more](#) [↗](#)

S3 location

[View](#) [↗](#)[Browse S3](#)

Validation dataset S3 location (optional)

[View](#) [↗](#)[Browse S3](#)

We can leave the hyperparameters as default because this is a machine learning training job



Hyperparameters [Info](#)

Epochs

The total number of iterations of all the training data in one cycle for training the model.

Enter an integer between 1 and 10.

Batch size

The number of samples processed before model parameters are updated.

Enter a value between 1 and 64.

Learning rate

The rate at which model parameters are updated after each batch of training data.

Enter a float value between 0 and 1.

Learning rate warmup steps

Number of iterations over which learning rate is gradually increased to the initial rate specified.

Enter an integer between 1 and 250.

Provide the S3 bucket location for the output data and also specify the service role name for the new role

Output data [Info](#)

Choose S3 location to store the model validation outputs.


S3 location

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

Service access [Info](#)



Bedrock model customization job requires permissions to write to S3 on your behalf.

 You need to create a new service role because there are no suitable roles currently available in the account.

Service role name

Maximum 64 characters. Use alphanumeric and '+,=,.,@,-,_' characters.

[View permission details](#)

Click **View Permission Details** to see the permissions



Permission details



Bedrock fine-tuning job requires the following permissions to write to S3 on your behalf.

Trusted entity

bedrock.amazonaws.com

Permissions

Permission policy

```
1 {  
2   "Version": "2012-10-17",  
3   "Statement": [  
4     ]  
}
```

Trust relationship for new service role

Trust relationship

```
{  
  "Version": "2012-10-17",  
  "Statement": [  
    {  
      "Effect": "Allow",  
      "Principal": {  
        "Service": "bedrock.amazonaws.com"  
      },  
      "Action": "sts:AssumeRole",  
      "Condition": {  
        "StringEquals": {  
          "aws:SourceAccount": "851725583489"  
        },  
        "ArnEquals": {  
          "aws:SourceArn": "arn:aws:bedrock:us-east-1:851725583489:model-customization-job/*"  
        }  
      }  
    }  
  ]  
}
```



Don't create this, as we need to purchase provisioned throughput to use the custom model, which incurs hourly-based costs that can be high

Service access [Info](#)

Bedrock model customization job requires permissions to write to S3 on your behalf.

You need to create a new service role because there are no suitable roles currently available in the account.

Service role name

bedrock-finetune

Maximum 64 characters. Use alphanumeric and '+=,.-_' characters.

[View permission details](#)

Purchase provisioned throughput to use fine-tuned model

After this custom model is created, you need to purchase provisioned throughput to be able to use this model.

[Learn more](#)

Cancel

Create Fine-tuning job

2.Foundation Model Evaluation

To evaluate the foundation models, click **Create Model Evaluation**

▼ Builder tools

Prompt management [Preview](#)

Knowledge bases

Agents

Prompt flows [Preview](#)

▼ Safeguards

Guardrails

Watermark detection

▼ Inference

Provisioned Throughput

Batch inference [New](#)

Cross-region inference [New](#)

▼ Assessment

[Model Evaluation](#)

User guide

Bedrock Service Terms

▼ Bedrock configurations

Model access [2 new](#)

Bedrock Studio [Preview](#)

Settings

[Amazon Bedrock](#) > Model evaluation jobs

Model evaluation [Info](#)

Create and view model evaluation jobs

Build an evaluation

Automatic

Evaluates a single model using recommended metrics. Provides results based on the parameters that you specify when you create the evaluation, such as accuracy, toxicity, and robustness. Choose from built-in task types, text summarization, question and answer, text classification, and open-ended text generation, and scores will be calculated automatically. Model scores are calculated using various statistical methods such as BERTScore, F1, and more. You can bring your own prompt dataset or use built-in curated prompt datasets.

Create automatic evaluation

Human: Bring your own work team

Evaluates up to 2 models using a work team of your choice to provide feedback. Provides results based on the parameters that you specify when you create the evaluation. You can use recommended task types and their associated metrics, or customize the task types and metrics that are important to your needs. You provide your own prompt dataset to ensure the evaluation is relevant to you. This is a good option if you want feedback on subjective or complex evaluation metrics.

Create human-based evaluation

Human: AWS Managed work team

Customize the number of models to evaluate using a work-team designated by AWS. Provides results based on the parameters that you specify when you create the evaluation. You provide your own prompt dataset, define the task types and metrics that are important to your evaluation, and engage with an AWS team directly. The AWS team will ensure that your evaluation meets your needs. This is a good option if you want feedback on subjective or complex evaluation metrics, and want an expert AWS team to manage the whole evaluation workflow within your guidelines.

Create AWS managed evaluation

Model Evaluation Jobs

Stop evaluation

Delete evaluations

Create model evaluation ▼

Model Evaluation Jobs you have created will appear here.



Choose **Automatic** in **Create Model Evaluation**

Build an evaluation



Automatic

Evaluates a single model using recommended metrics. Provides results based on the parameters that you specify when you create the evaluation, such as accuracy, toxicity, and robustness. Choose from built-in task types, text summarization, question and answer, text classification, and open-ended text generation, and scores will be calculated automatically. Model scores are calculated using various statistical methods such as BERTScore, F1, and more. You can bring your own prompt dataset or use built-in curated prompt datasets.

Create automatic evaluation



Human: Bring your own work team

Evaluates up to 2 models using a work team of your choice to provide feedback. Provides results based on the parameters that you specify when you create the evaluation. You can use recommended task types and their associated metrics, or customize the task types and metrics that are important to your needs. You provide your own prompt dataset to ensure the evaluation is relevant to you. This is a good option if you want feedback on subjective or complex evaluation metrics.

Create human-based evaluation



Human: AWS Managed work team

Customize the number of models to evaluate using a work-team designated by AWS. Provides results based on the parameters that you specify when you create the evaluation. You provide your own prompt dataset, define the task types and metrics that are important to your evaluation, and engage with an AWS team directly. The AWS team will ensure that your evaluation meets your needs. This is a good option if you want feedback on subjective or complex evaluation metrics, and want an expert AWS team to manage the whole evaluation workflow within your guidelines.

Create
Automatic
Human: Bring your own work team
Human: AWS Managed work team

Model Evaluation Jobs

Model Evaluation Jobs you have created will appear here.



Stop evaluation

Delete evaluations

Create model evaluation ▲

Provide the **Evaluation Name** and a description

[Amazon Bedrock](#) > [Model evaluation jobs](#) > Create automatic evaluation

Create automatic evaluation [Info](#)

Choose a model you want to evaluate, and select the task you want the model to perform. For each task, you can choose metrics to evaluate the model's performance.

Model evaluation details

Evaluation name

Name must be unique within your account's AWS region.

aievaluation

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

Description - optional

Test evaluation


The description can have up to 200 characters.



Select Base Model as Amazon

Model selector [Info](#)

Choose the model you want to evaluate. To change the hyperparameters of the model, choose **update**.
If you can't find the model you're looking for, check [model access](#).



Amazon

Select model

Base Models

- AI21 Labs
- Amazon
- Anthropic
- Cohere
- Meta
- Mistral AI

Task type [Info](#)

Choose a Model Evaluation task type to define the model evaluation criteria.

Task type


☐ General text generation
The model performs natural language processing and text generation tasks.

☐ Text summarization
The model summarizes text based on the prompts that you provide.

Select Model

Model selector [Info](#)

Choose the model you want to evaluate. To change the hyperparameters of the model, choose **update**.
If you can't find the model you're looking for, check [model access](#).



Amazon

Select model


Task type [Info](#)

Choose a Model Evaluation task type to define the model evaluation criteria.

Task type

☐ General text generation
The model performs natural language processing and text generation tasks.


☐ Text summarization
The model summarizes text based on the prompts that you provide.



Titan Text G1 - Lite

v1

Context size = 4k



Titan Text G1 - Express

v1

Context size = [object Object]

Choose Task Type as General Text Generation

Task type [Info](#)

Choose a Model Evaluation task type to define the model evaluation criteria.

Task type

☒ General text generation
The model performs natural language processing and text generation tasks.

☐ Text summarization
The model summarizes text based on the prompts that you provide.

☐ Question and answer
The answers that models provide are based on your prompts.

☐ Text classification
The model categorizes text into predefined classes based on the input dataset.

Choose Metrics and Prompt Dataset as Available Built-in Datasets



Metrics and datasets [info](#)
 Choose the metrics and datasets for evaluating the model's performance.

Metric

Toxicity
Gauges propensity to generate harmful, offensive, or inappropriate context.

Remove

Choose a prompt dataset
☒ Available built-in datasets
☐ Use your own prompt dataset
This is the S3 bucket where your prompt dataset is stored.

☒ **Real Toxicity**
RealToxicityPrompts is a dataset for measuring the degree to which racist, sexist, or otherwise toxic language presents in Pretrained neural language models (LMs).
☐ **BOLD**
Bias in Open-ended Language Generation Dataset (BOLD) is a dataset to evaluate fairness in open-ended language generation in English language. It consists of 23,679 different text generation prompts that allow fairness measurement across five domains: profession, gender, race, religious ideologies, and political ideologies.

Metric

Accuracy
Examines the model's ability to encode factual knowledge about the real world.

Remove

Choose a prompt dataset
☒ Available built-in datasets
☐ Use your own prompt dataset
This is the S3 bucket where your prompt dataset is stored.

☒ **TREX**
TREX is a Large Scale Alignment of Natural Language with Knowledge Base Triples for Relation Extraction and Natural Language Generation.

Metric

Robustness
Assesses the degree to which minor, semantic-preserving changes impact the model's output.

Remove


Choose a prompt dataset
☒ Available built-in datasets
☐ Use your own prompt dataset
This is the S3 bucket where your prompt dataset is stored.

☒ **BOLD**
Bias in Open-ended Language Generation Dataset (BOLD) is a dataset to evaluate fairness in open-ended language generation in English language. It consists of 23,679 different text generation prompts that allow fairness measurement across five domains: profession, gender, race, religious ideologies, and political ideologies.
☒ **TREX**
TREX is a Large Scale Alignment of Natural Language with Knowledge Base Triples for Relation Extraction and Natural Language Generation.
☒ **WikiText-2**
The WikiText language modeling dataset is a collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia.

Provide the S3 bucket location to store the evaluation results

Evaluation results [info](#)
Specify the S3 location
This is the folder in the S3 bucket where the results of the model evaluation job are stored.

S3 URI

View 

Browse S3

KMS key - Optional
By default your evaluation job data is encrypted with an AWS owned KMS key. If you want to use your own KMS key, choose the checkbox below. Then provide the ID or ARN of the key you would like to use.
☐ Customize encryption settings (advanced)
To use the default key, clear this option.

Choose **Create a New Role** that grants Amazon Bedrock the necessary permissions to access S3 buckets. If the IAM role doesn't have the required permissions

Amazon Bedrock IAM role - Permissions [info](#)
 Choose or create an IAM service role that grants Amazon Bedrock permission to the S3 buckets specified in your model evaluation job and the models you selected.

☒ Create a new role
☐ Use an existing role

Service role name

Amazon-Bedrock-IAM-Role-

The valid characters are a-z, A-Z, 0-9, _(underscore) and -(hyphen). The name can have up to 40 characters. Character count: 15

Specify S3 buckets



Provide the S3 bucket for output, and don't click Create, as it will incur high costs

Specific S3 buckets
Specify the S3 buckets to include in the new Bedrock IAM service role's permissions policy.

Output bucket
The S3 bucket corresponding to the S3 location selected for evaluation results will be added to the new Bedrock IAM service role.

Create role

Cancel Create

Keep Learning, Keep A!!!

Feel free to reach out to me, if you have any other queries or suggestions

Stay connected on LinkedIn <https://www.linkedin.com/in/mahendran-selvakumar-36444a77/>

Stay connected on Medium <https://devopstronaut.com/>