Amazon Bedrock

- Overview
- Features
- Pricing
- Model Providers
- FAQs
- Testimonials
- Resources

Meta Llama 3.1 on AWS | Build the future of AI with Meta's most advanced and capable Llama models in Amazon Bedrock »

Generative AI › Pricing

Amazon Bedrock pricing

Pricing overview

Amazon Bedrock is a fully managed servic rm foundation models (FMs) through a single bi

Hi, I can connect you with an AWS representative or answer questions you have on AWS.

1

You have a choice of two pricing plans for inference: 1. On-Demand and Batch: This mode allows you to use FMs on a pay-as-you-go basis without having to make any time-based term commitments. 2. Provisioned Throughput: This mode allows you to provision sufficient throughput to meet your application's performance requirements in exchange for a time-based term commitment.

Pricing models

On-Demand

With the On-Demand mode, you only pay for what you use, with no time-based term commitments. For text-generation models, you are charged for every input token processed and every output token generated. For embeddings models, you are charged for every input token processed. A token comprises a few characters and refers to the basic unit of text that a model learns to understand the user input and prompt. For image-generation models, you are charged for every image generated.

Batch

With Batch mode, you can provide a set of prompts as a single input file and receive responses as a single output file, allowing you to get simultaneous large-scale predictions. The responses are processed and stored in your Amazon S3 bucket so you can access them at a later time. Pricing for Batch mode is the same as pricing for On-Demand mode.

Provisioned Throughput

With the Provisioned Throughput mode, you can purchase model units for a specific base or custom model. The Provisioned Throughput mode is primarily designed for large hour, you have the flexibility to choose between 1-month or 6-month commitment terms.

Model customization

With Amazon Bedrock, you can customize FMs with your data to deliver tailored responses for specific tasks and your business context. You can fine-tune models with labeled data or using continued pretraining with unlabeled data. For customization of a text-generation model, you are charged for the model training based on the total number of tokens processed by the model (number of tokens in the training data corpus x the number of epochs) and for model storage charged per month per model. An epoch refers to one full pass through your training dataset during fine-tuning or continued pretraining. Inferences using customized models are charged under the Provisioned Throughput plan and requires you purchase Provisioned Throughput. One model unit is made available with no commitment term for inference on a customized model. You will be charged for the number of hours you use in the first model unit for custom model inference. If you want to increase your throughput beyond one model unit, then you must purchase a 1-month or 6-month commitment term.

Model evaluation

With model evaluation on Amazon Bedrock you pay for what you use, with no volume commitments on the number of prompts or responses. For automatic evaluation, you only pay for the inference from your choice of model in the evaluation. The automatically-generated algorithmic scores are provided at no extra charge. For human-based evaluation where you bring your own workteam, you are charged for the model inference in the evaluation, and a charge of $0.21 per completed human task. A human task is defined as an instance of a human worker submitting an evaluation of a single prompt and its associated inference responses in the human evaluation user interface. The price is the same whether you have one or two models in your evaluation job and also the same regardless of how many evaluation metrics and rating methods you include. The charges for the human tasks will appear under the Amazon SageMaker section in your AWS bill and are the same for all AWS Regions. There is no separate

Powerful tools to build at no extra charge

When using Agents for Amazon Bedrock and Knowledge Bases for Amazon Bedrock, you are only charged for the models and the vector databases you use with these capabilities.

Pricing breakdown

Pricing is dependent on the modality, provider, and model. Please select the model provider to see detailed pricing.

AI21 Labs

On-Demand and Batch pricing

AI21 Labs models	Price per 1,000 input tokens	Price per 1,000 output tokens
Jurassic-2 Mid	$0.0125	$0.0125

Jamba-Instruct	$0.0005	$0.0007

Amazon

Region

US East (N. Virginia)

On-Demand and Batch pricing for text models

Amazon Titan models	Price per 1,000 input tokens	Price per 1,000 output tokens
Amazon Titan Text Premier	$0.0005	$0.0015
Amazon Titan Text Lite	$0.00015	$0.0002
Amazon Titan Text Express	$0.0002	$0.0006
Amazon Titan Text Embeddings	$0.0001	n/a
Amazon Titan Text Embeddings V2	$0.00002	n/a

Amazon Titan models	Image resolution	Price per image generated for Standard quality	Price per image generated for Premium quality
Amazon Titan Image Generator	512 x 512	$0.008	$0.01
	1024 x 1024	$0.01	$0.012
Amazon Titan Image Generator (custom models)	512 x 512	$0.018	$0.02

Amazon Titan models	Price per 1,000 input tokens	Price per input image
Amazon Titan Multimodal Embeddings	$0.0008	$0.00006

Pricing for model customization (fine-tuning and continued pretraining)

Amazon Titan	Price to train 1,000 tokens*	Price to store each custom model per month	Price to infer for 1 model unit per hour**
Amazon Titan Text Lite	$0.0004	$1.95	$7.10
Amazon Titan	$0.008	$1.95	$20.50

Amazon Titan	Price per image seen	Price to store each custom model per month	Price to infer for 1 model unit per hour**
Amazon Titan Image Generator	$0.005	$1.95	$23.40

Amazon Titan Text Lite	$7.10	$6.40	$5.10
Amazon Titan Text Express	$20.50	$18.40	$14.80
Amazon Titan Embeddings	N/A	$6.40	$5.10
Amazon Titan Image Generator	N/A	$16.20	$13.00
Amazon Titan Image Generator (custom models)	$23.40	$21.00	$16.85

Anthropic

On-Demand and Batch pricing

Region: US East (N. Virginia) and US West (Oregon)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude 3.5 Sonnet**	$0.003	$0.015
Claude 3 Opus*	$0.015	$0.075
Claude 3 Haiku	$0.00025	$0.00125

Claude 2.1	$0.008	$0.024
Claude 2.0	$0.008	$0.024
Claude Instant	$0.0008	$0.0024
**Claude 3 Opus is currently available in the US West (Oregon) Region*		
***Claude 3.5 Sonnet is currently available in the US East (N. Virginia) Region*		

Region: EU (London)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude 3 Sonnet	$0.003	$0.015
Claude 3 Haiku	$0.00025	$0.00125

Region: South America (Sao Paolo)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude 3 Sonnet	$0.003	$0.015
Claude 3 Haiku	$0.00025	$0.00125

Region: Canada (Central)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens

Claude 3 Haiku	$0.00025	$0.00125

Region: Asia Pacific (Mumbai)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude 3 Sonnet	$0.003	$0.015
Claude 3 Haiku	$0.00025	$0.00125

Region: Asia Pacific (Sydney)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude 3 Sonnet	$0.003	$0.015
Claude 3 Haiku	$0.00025	$0.00125

Region: Asia Pacific (Tokyo)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude Instant	$0.0008	$0.0024
Claude 2.0/2.1	$0.008	$0.024

Region: Europe (Paris)

Claude 3 Haiku	$0.00025	$0.00125

Region: Europe (Frankfurt)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude Instant	$0.0008	$0.0024
Claude 2.0/2.1	$0.008	$0.024
Claude 3 Sonnet	$0.003	$0.015
Claude 3 Haiku	$0.00025	$0.00125

Provisioned Throughput pricing

Region: US East (N. Virginia) and US West (Oregon)

Anthropic models	Price per hour per model with no commitment	Price per hour per model unit for 1-month commitment	Price per hour per model unit for 6-month commitment
Claude Instant	$44.00	$39.60	$22.00
Claude 2.0/2.1	$70.00	$63.00	$35.00

	no commitment		
Claude Instant	$44.00	$39.60	$22.00
Claude 2.0/2.1	$70.00	$63.00	$35.00

Region: Asia Pacific (Tokyo)

Anthropic models	Price per hour per model unit for 1-month commitment	Price per hour per model unit for 6-month commitment
Claude Instant	$53.00	$29.00
Claude 2.0/2.1	$86.00	$48.00

Region: Europe (Frankfurt)

Anthropic models	Price per hour per model unit for 1-month commitment	Price per hour per model unit for 6-month commitment
Claude Instant	$49.00	$27.00
Claude 2.0/2.1	$79.00	$44.00

Please reach out to your AWS account team for more details on model units.

Cohere models	Price per 1,000 input tokens	Price per 1,000 output tokens
Command	$0.0015	$0.0020
Command-Light	$0.0003	$0.0006
Command R+	$0.0030	$0.0150
Command R	$0.0005	$0.0015
Embed - English	$0.0001	N/A
Embed - Multilingual	$0.0001	N/A

Pricing for customization (fine-tuning)

Cohere models	Price to train 1,000 tokens	Price to store each custom model per month	Price to infer from a custom model per model unit per hour (with no-commit Provisioned Throughput pricing)
Cohere Command	$0.004	$1.95	$49.50
Cohere Command-Light	$0.001	$1.95	$8.56

**Total tokens trained = number of tokens in training data corpus x number of epochs*

Provisioned Throughput pricing

	commitment		
Cohere Command	$49.50	$39.60	$23.77
Cohere Command - Light	$8.56	$6.85	$4.11
Embed - English	$7.12	$6.76	$6.41
Embed - Multilingual	$7.12	$6.76	$6.41

Please reach out to your AWS account or sales team for more details on model units.

Meta Llama

Llama 3.1

On-Demand and Batch pricing

Region

US West (Oregon)

Meta models	Price per 1,000 input tokens	Price per 1,000 output tokens
Llama 3.1 Instruct (8B)	$0.0003	$0.0006
Llama 3.1 Instruct (70B)	$0.00265	$0.0035
Llama 3.1 Instruct (405B)	$0.00532	$0.016

US East (N. Virginia)

Meta models	Price per 1,000 input tokens	Price per 1,000 output tokens
Llama 3 Instruct (8B)	$0.0003	$0.0006
Llama 3 Instruct (70B)	$0.00265	$0.0035

Llama 2

On-Demand and Batch pricing

Region: US East (N. Virginia) and US West (Oregon)

Meta models	Price per 1,000 input tokens	Price per 1,000 output tokens
Llama 2 Chat (13B)	$0.00075	$0.001
Llama 2 Chat (70B)	$0.00195	$0.00256

Pricing for model customization (fine-tuning)

Meta models	Price to train 1,000 tokens	Price to store each custom model* per month	Price to infer from a custom model for 1 model unit per hour (with no-commit Provisioned Throughput pricing)
Llama 2 Pretrained (13B)	$0.00149	$1.95	$23.50

**Custom model storage = $1.95*

Provisioned Throughput pricing

Meta models	Price per hour per model unit for 1-month commitment	Price per hour per model unit for 6-month commitment
Llama 2 Pretrained and Chat (13B)	$21.18	$13.08
Llama 2 Pretrained (70B)	$21.18	$13.08

**Llama 2 Pre-trained models are available only in provisioned throughput after customization.*

Please reach out to your AWS account or sales team for more details on model units.

Mistral AI

Region

US East (N. Virginia)

Mistral 7B	$0.00015	$0.0002
Mixtral 8*7B	$0.00045	$0.0007
Mistral Small (24.02)	$0.001	$0.003
Mistral Large (24.02)	$0.004	$0.012

Stability AI

On-Demand and Batch pricing

Image models offered by Stability AI are priced per image, depending on step count and image resolution

Stability AI model	Image resolution	Price per image generated for standard quality (<=50 steps)	Price per image generated for premium quality (>50 steps)
SDXL 0.8	512 x 512 or smaller	$0.018 per image	$0.036 per image
	Larger than 512 x 512	$0.036 per image	$0.072 per image
SDXL 1.0	Up to 1024 x 1024	$0.04	$0.08

Provisioned Throughput pricing

**Includes inference for base and custom models*

Please reach out to your AWS account or sales team for more details on model units.

Currently, model customization (fine-tuning) is not supported for Stability AI models on Amazon Bedrock.

Guardrails for Amazon Bedrock

On-Demand pricing

Guardrail policy*	Price per 1,000 text units**
Content filters	$0.75
Denied topics	$1
Contextual grounding check***	$0.1
Sensitive information filter (PII)	$0.1
Sensitive information filter (regular expression)	Free

** Each guardrail policy is optional and can be enabled based on your application requirements. Charges will be incurred based on the policy type used in the guardrail. For example, if a guardrail is configured with content filters and denied topics, charges will be incurred for these two policies, while there will be no charges associated with sensitive information filters.*

***A text unit can contain up to 1000 characters. If a text input is more than 1000 characters, it is processed as multiple text units, each containing 1000 characters or less. For example, if a text input contains 5600 characters, it will be charged for 6 text units.*

**** Contextual grounding check uses a reference source and a query to determine if the model response is grounded based on the source and relevant to the query. The total number of text units charged is calculated by combining all the characters in the source, query, and model response.*

Guardrails are not supported for images and embeddings.

Pricing examples

AI21 labs +

Amazon +

Cohere

Meta Llama

Mistral AI

Stability AI

Model evaluation

Guardrails for Amazon Bedrock

Learn About AWS

- What Is AWS?
- What Is Cloud Computing?
- AWS Accessibility
- AWS Inclusion, Diversity & Equity
- What Is DevOps?
- What Is a Container?
- What Is a Data Lake?
- What is Artificial Intelligence (AI)?
- What is Generative AI?

Resources for AWS

- Getting Started
- Training and Certification
- AWS Solutions Library
- Architecture Center
- Product and Technical FAQs
- Analyst Reports
- AWS Partners

Developers on AWS

- Developer Center
- SDKs & Tools
- .NET on AWS
- Python on AWS
- Java on AWS
- PHP on AWS
- JavaScript on AWS

Press Releases

Help

Contact Us

Get Expert Help

File a Support Ticket

AWS re:Post

Knowledge Center

AWS Support Overview

Legal

AWS Careers

Amazon is an Equal Opportunity Employer: *Minority / Women / Disability / Veteran / Gender Identity / Sexual Orientation / Age.*

Language

عربي |
Bahasa Indonesia |
Deutsch |
English |
Español |
Français |
Italiano |
Português |
Tiếng Việt |
Türkçe |
Русский |
ไทย |
日本語 |
한국어 |

Accessibility

|

Site Terms

|

Cookie Preferences

|