Quotas for Amazon Bedrock

PDF (/pdfs/bedrock/latest/userguide/bedrock-ug.pdf#quotas) | RSS (bedrock-ug.rss)

Your AWS account has default quotas, formerly referred to as limits, for Amazon Bedrock. To view service quotas for Amazon Bedrock, follow the steps at Viewing service quotas (https://docs.aws.amazon.com/servicequotas/latest/userguide/gs-request-quota.html) and select **Amazon Bedrock** as the service. Some quotas differ by model. Unless specified otherwise, a quota applies to all versions of a model.

To maintain the performance of the service and to ensure appropriate usage of Amazon Bedrock, the default quotas assigned to an account might be updated depending on regional factors, payment history, fraudulent usage, and/or approval of a quota increase request.

You can request a quota increase for your account by following the steps below:

- If a quota is marked as **Yes** in the **Adjustable through Service Quotas** column in the following tables, you can adjust it by following the steps at Requesting a Quota Increase (https://docs.aws.amazon.com/servicequotas/latest/userguide/request-quota-increase.html) in the *Service Quotas User Guide* in the Service Quotas User Guide (https://docs.aws.amazon.com/servicequotas/latest/userguide/) .
- If a quota is marked as **No** in the **Adjustable through Service Quotas** column in the following tables, you might be able to request a quota increase in one of the following ways:
  - To request a quota increase for a Runtime quota (#quotas-runtime) , contact your AWS account manager. If you don't have an AWS account manager, you can't increase your quota at this time.
  - To request other quota increases, submit a request through the limit increase form (https://console.aws.amazon.com/support/home#/case/create?issueType=service-limit-increase) to be considered for an increase.

> **Note**
>
> Due to overwhelming demand, priority will be given to customers who generate traffic that consumes their existing quota allocation. Your request might be denied if you don't meet this condition.

Select a topic to learn more about the default global quotas for it. All global and Regional quotas are the same unless otherwise specified.

Runtime quotas

The following quotas apply when you carry out model inference. These quotas consider the combined sum for Converse (https://docs.aws.amazon.com/bedrock/latest/APIReference/API_runtime_Converse.html) , ConverseStream (https://docs.aws.amazon.com/bedrock/latest/APIReference/API_runtime_ConverseStream.html) , InvokeModel (https://docs.aws.amazon.com/bedrock/latest/APIReference/API_runtime_InvokeModel.html) , and InvokeModelWithResponseStream (https://docs.aws.amazon.com/bedrock/latest/APIReference/API_runtime_InvokeModelWithResponseStream.html) requests. Inference latency differs by model and is directly proportional to the number of input and output tokens and the total number of ongoing on-demand requests by all customers at the time. For guaranteed throughput, we encourage you to try Provisioned Throughput (./prov-throughput.html) .

Model	Requests processed per minute	Tokens processed per minute	Regions	Adjustable through Service Quotas
AI21 Labs Jurassic-2 Mid	400	300,000	us-east-1	No

Model	Requests processed per minute	Tokens processed per minute	Regions	Adjustable through Service Quotas
AI21 Labs Jurassic-2 Ultra	100	300,000	All	No
AI21 Jamba-Instruct	100	300,000	All	No
Amazon Titan Embeddings G1 - Text	2,000	300,000	All	No
Amazon Titan Text Embeddings V2	2,000	300,000	All	No
Amazon Titan Image Generator G1 V1	60	N/A	All	No
Amazon Titan Image Generator G1 V2	60	N/A	All	No
Amazon Titan Multimodal Embeddings G1	2,000	300,000	All	No
Amazon Titan Text G1 - Express	400	300,000	All	No
Amazon Titan Text G1 - Lite	800	300,000	All	No
Amazon Titan Text Premier	100	300,000	All	No
Anthropic Claude Instant	1,000	1,000,000	us-east-1 us-west-2	No
	400	300,000	Other regions	
Anthropic Claude 2.x	500	500,000	us-east-1 us-west-2	No
	100	200,000	Other regions	
Anthropic Claude 3 Sonnet	500	1,000,000	us-east-1 us-west-2	No
	100	200,000	Other regions	
Anthropic Claude 3 Haiku	1,000	2,000,000	us-east-1 us-west-2	No
	200	200,000	ap-northeast-1 ap-southeast-1	
	400	300,000	Other regions	

Model	Requests processed per minute	Tokens processed per minute	Regions	Adjustable through Service Quotas
Anthropic Claude 3.5 Sonnet	250	2,000,000	us-west-2	No
	20	200,000	ap-northeast-1 ap-southeast-1 eu-central-1	No
	50	400,000	Other regions	No
Anthropic Claude 3 Opus	50	400,000	All	No
Cohere Command R	400	300,000	All	No
Cohere Command R+	400	300,000	All	No
Cohere Command	400	300,000	All	No
Cohere Command Light	800	300,000	All	No
Cohere Embed (English)	2,000	300,000	All	No
Cohere Embed (Multilingual)	2,000	300,000	All	No
Meta Llama 2 13B	800	300,000	All	No
Meta Llama 2 70B	400	300,000	All	No
Meta Llama 3 8B Instruct	800	300,000	All	No
Meta Llama 3 70B Instruct	400	300,000	All	No
Meta Llama 3.1 8B Instruct	800	300,000	us-west-2	No
Meta Llama 3.1 70B Instruct	400	300,000	us-west-2	No
Meta Llama 3.1 405B Instruct	50	400,000	us-west-2	No
Mistral AI Mistral 7B Instruct	800	300,000	All	No
Mistral AI Mixtral 8X7B Instruct	400	300,000	All	No
Mistral AI Mistral Large	400	300,000	All	No
Mistral AI Mistral Large 2 (24.07)	400	300,000	us-west-2	No
Mistral AI Mistral Small	400	300,000	All	No
Stable Diffusion XL	60	N/A	All	No

▼ API requests per second

The following table shows the maximum number of API requests that are allowed per second for different API operations in Amazon Bedrock:

Feature	API operation	Maximum requests per second
N/A	Converse	200
	ConverseStream	200
	DeleteModelInvocationLoggingConfiguration	1
	GetFoundationModel	10
	GetModelInvocationLoggingConfiguration	10
	InvokeModel	200
	InvokeModelWithResponseStream	200
	ListFoundationModels	10
	ListTagsForResource	20
	PutModelInvocationLoggingConfiguration	1
	TagResource	20
	UntagResource	20
Agents	AssociateAgentKnowledgeBase	6
	CreateAgent	6
	CreateAgentActionGroup	12
	CreateAgentAlias	2
	DeleteAgent	2
	DeleteAgentActionGroup	2
	DeleteAgentAlias	2
	DeleteAgentVersion	2
	DisassociateAgentKnowledgeBase	4
	GetAgent	15
	GetAgentActionGroup	20
	GetAgentAlias	10
	GetAgentKnowledgeBase	15
	GetAgentVersion	10
	ListAgents	10
	ListAgentActionGroups	10
	ListAgentAliases	10
	ListAgentKnowledgeBases	10
	ListAgentVersions	10
	PrepareAgent	2
	UpdateAgent	4
	UpdateAgentActionGroup	6
	UpdateAgentAlias	2
	UpdateAgentKnowledgeBase	4
Custom models	CreateModelCustomizationJob	1
	DeleteCustomModel	10
	GetCustomModel	10
	GetModelCustomizationJob	10
	ListModelCustomizationJobs	10
	StopModelCustomizationJob	10

Feature	API operation	Maximum requests per second
Guardrails	CreateGuardrail	1
	CreateGuardrailVersion	1
	DeleteGuardrail	1
	GetGuardrail	10
	ListGuardrails	10
	UpdateGuardrail	1
Knowledge bases	CreateDataSource	2
	CreateKnowledgeBase	2
	DeleteDataSource	2
	DeleteKnowledgeBase	2
	GetDataSource	10
	GetIngestionJob	10
	GetKnowledgeBase	10
	ListDataSources	10
	ListIngestionJobs	10
	ListKnowledgeBases	10
	Retrieve	5
	RetrieveAndGenerate	5
	StartIngestionJob	0.1
	UpdateDataSource	2
	UpdateKnowledgeBase	2
Model evaluation	CreateEvaluationJob	5
	GetEvaluationJob	10
	ListEvaluationJobs	10
	StopEvaluationJob	5
Provisioned Throughput	CreateProvisionedModelThroughput	1
	DeleteProvisionedModelThroughput	1
	GetProvisionedModelThroughput	10
	ListProvisionedModelThroughputs	10
	UpdateProvisionedModelThroughput	1

▼ Model inference prompt quotas

Select a tab to see model-specific quotas for prompts.

Amazon Titan Text models | Amazon Titan Image Generator G1 V1 | Amazon Titan Embeddings G1 - Text | An

Description	Value	Adjustable through Service Quotas
Text prompt length, in characters	42,000	No

Batch inference quotas

The following quotas apply when you run batch inference. The quotas depend on the modality of the input and output data.

Modality	Minimum file size	Maximum file size	Adjustable through Service Quotas
Text to embeddings	75 MB	500 MB	No
Text to text	20 MB	150 MB	No
Text/image to image	1 MB	50 MB	No

Guardrails quotas

The following quotas are enforced when you use guardrails.

Quota	Description	Value
Guardrails per account	The maximum number of guardrails in an account.	100
Versions per guardrail	The maximum number of versions that a guardrail can have.	20
Topics per topic guardrail	The maximum number of topics that can be defined across guardrail topic policies.	30
Example phrases per topic	The maximum number of topic examples that can be included in a topic.	5
Regex expressions in the Sensitive information filter	The maximum number of guardrail filter regexes that can be included in a Sensitive information policy	10
Regex length in characters	The maximum length, in characters, of a guardrail filter regex.	500
Words per Word policy	The maximum number of words that can be included in a blocked word list.	10,000
Word length in characters	The maximum length of a word, in characters, in a blocked word list.	100
On-demand ApplyGuardrail requests per second	The maximum number of ApplyGuardrail API calls allowed per second.	25
On-demand ApplyGuardrail Denied topic policy text units per second.	The maximum number of text units that can be processed for Denied topic policies per second.	25
On-demand ApplyGuardrail Content filter policy text units per second	The maximum number of text units that can be processed for Content filter policies per second.	25

Quota	Description	Value
On-demand ApplyGuardrail Word filter policy text units per second	The maximum number of text units that can be processed for Word filter policies per second.	25
On-demand ApplyGuardrail Sensitive information filter policy text units per second	The maximum number of text units that can be processed for Sensitive information filter policies per second.	25

Note

A text unit can be up to 1,000 characters

Knowledge base quotas

The following quotas apply to Knowledge bases for Amazon Bedrock.

Description	Maximum	Adjustable through Service Quotas	Description
Knowledge bases per account	100	No	The maximum number of knowledge bases per account.
Data sources per knowledge base	5	No	The maximum number of data sources per knowledge base.
Data source chunk size (Titan Text G1 - Embeddings)	8,192	No	The maximum size (in KB) of a data source using Titan Embeddings G1 - Text.
Data source chunk size (Cohere Embed English)	512	No	The maximum size (in KB) of a data source using Cohere Embed English.
Data source chunk size (Cohere Embed Multilingual)	512	No	The maximum size (in KB) of a data source using Cohere Embed Multilingual.
Data source total metadata fields/attributes per chunk.	250	No	The maximum number of document metadata fields/attributes per chunk.
Data source total crawled content items for Web Crawler	25,000	No	The maximum number of web page content items (50 MB max per content item) that can be crawled.
Data source total crawled files	2.5 million	No	The maximum number of data source files or content items (50 MB max per file/content item) that can be crawled.
Advanced parsing total data size	100 MB	No	The maximum combined size (in MB) of data that can be parsed using advanced parsing.
Advanced parsing total files	100	No	The maximum number of files that can be parsed using advanced parsing.
Files to add or update per ingestion job	5,000,000	No	The maximum number of new and updated files that can be ingested per ingestion job.
Files to delete per ingestion job	5,000,000	No	The maximum number of files that can be deleted per ingestion job.
Ingestion job file size (source document)	50 MB	No	The maximum size (in MB) of a source document file in an ingestion job.

Description	Maximum	Adjustable through Service Quotas	Description
Ingestion job file size (metadata file)	10 KB	No	The maximum size (in KB) of a metadata file in an ingestion job.
Ingestion job size	100 GB	No	The maximum size (in GB) of the ingestion job.
Concurrent ingestion jobs per data source	1	No	The maximum number of ingestion jobs that can take place at the same time for a data source.
Concurrent ingestion jobs per knowledge base	1	No	The maximum number of ingestion jobs that can take place at the same time for a knowledge base.
Concurrent ingestion jobs per account	5	No	The maximum number of ingestion jobs that can take place at the same time in an account.
User query size	1,000	No	The maximum size (in characters) of a user query.

▼ Agent quotas

The following quotas apply to Agents for Amazon Bedrock.

Quota	Maximum	Adjustable through Service Quotas	Description
Agents per account	50	Yes	The maximum number of Agents in one account.
Associated aliases per agent	10	No	The maximum number of aliases that you can associate with an agent.
Characters in agent instructions	4,000	Yes	The maximum number of characters in the instructions for an agent.
Action groups per agent	20	Yes	The maximum number of action groups that you can add to an agent.
Enabled action groups per agent	11	Yes	The maximum number of action groups that can be enabled in an agent.
APIs or Functions per Agent	11	Yes	The maximum number of APIs that you can add to an Agent.
Parameters per Function	5	Yes	The maximum number of parameters that you can add to a function for an action group.
Lambda response payload size	25 KB	No	The maximum size of the payload in an action group Lambda response.
Associated knowledge bases per Agent	2	Yes	The maximum number of knowledge bases that you can associate with an Agent.

▼ Prompt management quotas

The following quotas apply to Prompt management.

Quota	Maximum	Adjustable through Service Quotas	Description
Prompts per account	50	No	The maximum number of prompts in Prompt management that you can have in an account.

Quota	Maximum	Adjustable through Service Quotas	Description
Versions per prompt	10	No	The maximum number of versions that a prompt in Prompt management can have.

Prompt flows quotas

The following quotas apply to Prompt flows.

Quota	Maximum	Adjustable through Service Quotas	Description
Prompt flows per account	10	No	The maximum number of prompt flows that you can have in an account.
Nodes per prompt flow	20	No	The maximum number of nodes that you can have in a prompt flow.
Versions per prompt flow	10	No	The maximum number of versions that a prompt flow can have.
Aliases per prompt flow	10	No	The maximum number of aliases that you can associate with a prompt flow.
Prompt flows per account	10	No	The maximum number of prompt flows that you can have in an account.
Prompt flows per account	10	No	The maximum number of prompt flows that you can have in an account.
Flow input nodes per prompt flow	1	No	The maximum number of flow input nodes that you can add to a prompt flow.
Flow output nodes per prompt flow	5	No	The maximum number of flow output nodes that you can add to a prompt flow.
Condition nodes per prompt flow	5	No	The maximum number of condition nodes that you can add to a prompt flow.
Iterator nodes per prompt flow	1	No	The maximum number of iterator nodes that you can add to a prompt flow.
Collector nodes per prompt flow	1	No	The maximum number of collector nodes that you can add to a prompt flow.
Prompt nodes per prompt flow	5	No	The maximum number of prompt nodes that you can add to a prompt flow.
Lambda nodes per prompt flow	5	No	The maximum number of Lambda nodes that you can add to a prompt flow.
Lex nodes per prompt flow	5	No	The maximum number of Lex nodes that you can add to a prompt flow.
Nodes per node type per prompt flow	5	No	The maximum number of nodes you can add for each type in a prompt flow.
Conditions per condition node	5	No	The maximum number of conditions that you can add to a condition node in a prompt flow.

Model customization quotas

The following quotas apply to model customization.

Description	Maximum	Adjustable through Service Quotas
The maximum number of imported models in an account.	0	Yes
The maximum number of scheduled customization jobs.	2	No
The maximum number of custom models in an account.	100	Yes

To see hyperparameter quotas, see Custom model hyperparameters (./custom-models-hp.html) .

Select a tab to see model-specific quotas that apply to training and validation datasets used for customizing different foundation models.

Amazon Titan Text Premier | Amazon Titan Text G1 - Express | Amazon Titan Text G1 - Lite | Amazon Titan Im

Description	Maximum (Continued Pre-training) Not available	Maximum (Fine-tuning) Preview only	Adjustable through Service Quotas
Sum of input and output tokens when batch size is 1	N/A	4,096	No
Sum of input and output tokens when batch size is 2, 3, or 4	N/A	N/A	No
Character quota per sample in dataset	N/A	Token quota x 6	No
Sum of training and validation records	N/A	20,000	Yes
Training dataset file size	N/A	1 GB	No
Validation dataset file size	N/A	100 MB	No

Provisioned Throughput quotas

The following quotas apply to Provisioned Throughput.

Note

If a quota is marked as not adjustable through Service Quotas, you can submit a request through the limit increase form (https://console.aws.amazon.com/support/home#/case/create?issueType=service-limit-increase) to be considered for an increase.

Description	Default	Adjustable through Service Quotas
Model units that can be distributed across no-commitment Provisioned Throughputs	2	No
Model units that can be distributed across Provisioned Throughputs with commitment	0	No

Model evaluation job quotas

The following quotas apply to model evaluation jobs,

Job type	Description	Default	Adjustable
Automated	The maximum number of datasets that you can specify in an automated model evaluation job. This includes both custom and built-in prompt datasets.	5	No
Automated	The maximum number of metrics that you can specify per dataset in an automated model evaluation job. This includes both custom and built-in metrics.	3	No
Human	The maximum number of custom metrics that you can specify in a model evaluation job that uses human workers.	10	No
Automated	The maximum number of models that you can specify in an automated model evaluation job.	1	No
Human	The maximum number of models that you can specify in a model evaluation job that uses human workers.	2	No
Automated	The maximum number of automatic model evaluation jobs that you can specify at one time in this account in the current Region.	20	No
Human	The maximum number of model evaluation jobs that use human workers you can specify at one time in this account in the current Region.	10	No
Both	The maximum number of model evaluation jobs that you can create in this account in the current Region.	500	No
Human	The maximum number of custom prompt datasets that you can specify in a human-based model evaluation job in this account in the current Region.	1	No
Both	The maximum number of prompts a custom prompt dataset can contains.	1,000	No
Both	The maximum size (in KB) of an individual prompt is a custom prompt dataset.	4 KB	No
Human	The maximum length (in days) of time that a worker can have to complete tasks.	30	No