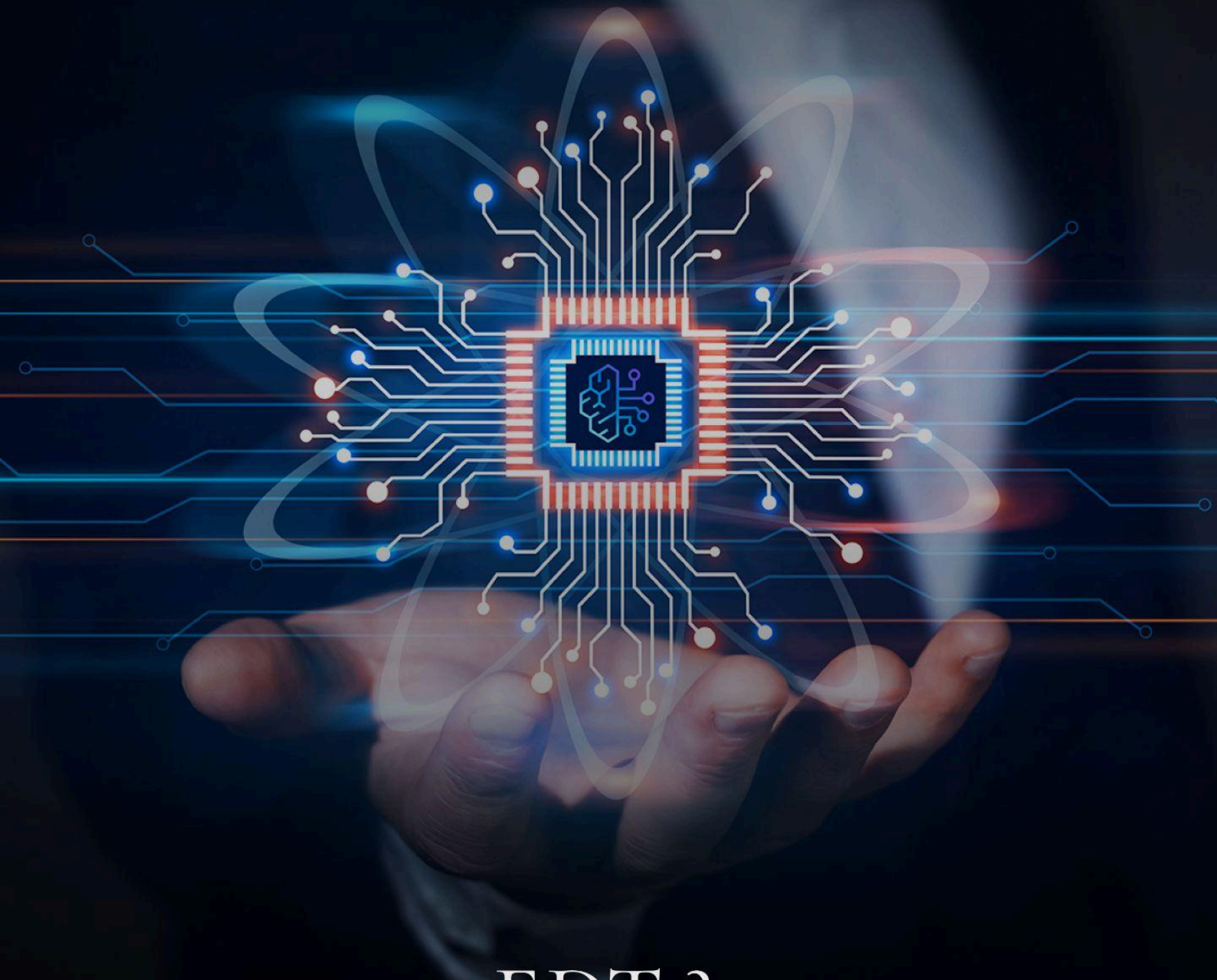
A guide on Bedrock pricing

January 2025



EDT &
PARTNERS

EDT&Partners *is a global, mission-driven consulting firm 100% dedicated to the education space. EDT&Partners is your Technology & Business of Education partner dedicated to IMAGINING, INSPIRING and IMPROVING education.*

We bring about fast, impactful change in education through our insights, network, global perspective, reach and innovative viewpoints. Our team is a unique blend of education, business and technology experts, boasting a global footprint and embracing a heterogeneous, multicultural view of the world of education, with a presence in 19 countries.

We help publishers, EdTechs, universities, nonprofits, school networks and public entities in the education space with go-to-market, product and cloud technology, strategy, optimization and digital transformation.

Pricing overview

Amazon Bedrock is a fully managed service that offers a choice of high-performing foundation models (FMs) through a single API, along with a broad set of capabilities you need to build generative AI applications with security, privacy, and responsible AI.

- With Amazon Bedrock, you will be charged for:
    - Model inference
    - Model customization
- You have a choice of two pricing plans for inference:
- **1. On-Demand and Batch:**
    - Pay-as-you-go pricing without time-based commitments.
- **2. Provisioned Throughput:**
    - Provision throughput to meet performance requirements in exchange for a time-based commitment.

Pricing Models

On Demand and Batch

- **On-Demand Mode:** Pay only for what you use with no term commitments.
    - **Text-Generation Models:** Charged per input and output token.
    - **Embeddings Models:** Charged per input token.
    - **Image-Generation Models:** Charged per image generated.
    - **Cross-Region Inference:** Supports using compute across different AWS Regions to manage traffic bursts, with no extra charge.
- **Batch Mode:** Submit prompts as a single input file and receive responses in an output file.
    - Responses are stored in an Amazon S3 bucket for future access.
    - Batch inference pricing is 50% lower than on-demand pricing for select models from providers like Anthropic, Meta, Mistral AI, and Amazon.

Provisioned Throughput

- **Provisioned Throughput Mode:** Purchase model units for a specific base or custom model.

- Designed for large, consistent inference workloads needing guaranteed throughput.
- Custom models are only available with this mode.
- **Model Unit:** Provides a defined throughput (tokens processed per minute).
- **Pricing:** Charged by the hour with a choice of 1-month or 6-month commitment terms.

Custom Model Import

- **Custom Model Import:** Import your customized models into Amazon Bedrock to use them like other hosted models.
  - **No Charge:** Importing a custom model to Bedrock is free.
  - On-Demand Serving: Imported models are available on-demand with no control plane actions required.
  - **Inference Pricing:** Charged based on the number of model copies needed for inference and their active duration (billed in 5-minute increments).
  - **Model Copy Cost:** Pricing depends on factors like architecture, context length, AWS Region, compute version, and model size tier.

Customization and optimization

Model customization

- **Model Customization:** Customize foundation models (FMs) with your data for tailored responses.
  - Fine-tune with labeled data or continue pre-training with unlabeled data.
  - **Pricing:** Charged for model training based on the total tokens processed and for model storage per month.
  - **Epoch:** One complete pass through the training dataset during customization.
  - **Inference with Customized Models:** Requires Provisioned Throughput.
    - One model unit available with no commitment term, charged by the hour.
    - To increase throughput beyond one model unit, a 1-month or 6-month commitment is required.

Model Distillation

Amazon Bedrock Model Distillation Pricing

- **Synthetic Data Generation:** Charged at **on-demand** pricing of the selected teacher model.
- **Fine-Tuning of the Student Model:** Charged at **model customization rates**.
- **Inference Costs:** Since a **distilled model** is considered a customized model, inferences are charged under the **Provisioned Throughput plan**.
- **Provisioned Throughput Requirement:** Customers are required to **purchase Provisioned Throughput** to use the distilled model for inferences.

Prompt Caching

- **Prompt Caching on Amazon Bedrock**
    - **Cost and Latency Reduction:** Cache **repeated context** across API calls to **reduce costs** and **improve response latencies**.
    - **Common Use Cases:** Useful for prompts with **common context or prefixes**, such as **long, multi-turn conversations, many-shot examples**, and **detailed instructions**.
    - **Caching Configuration:** Use **Amazon Bedrock APIs** to **cache prompt prefixes** for **five minutes** in an **AWS account-specific cache**.
    - **Discount and Latency Improvement:** Requests with **matching prefixes** receive **up to 90% discount** on cached tokens and **up to 85% latency improvement** during the caching period.
    - **Pricing and Performance Variability: Prices and performance improvements vary** based on the model and prompt length.
    - **Account Isolation:** All caches are **isolated to your AWS account**.

Advanced Tools

Agents

Amazon Bedrock Agents offer you the ability to build and configure autonomous agents within your application. These agents securely connect to your company's data sources and augment user requests with the right information to generate accurate responses. You can create an agent in Amazon Bedrock with just a few quick steps, accelerating the time it takes to build generative AI applications. These agents support code interpretation to dynamically generate

and execute code as well as return of control, which allows you to define an action schema and get the control back whenever the agent invokes the action. Additionally, Amazon Bedrock Agents can retain memory across interactions, offering more personalized and seamless user experiences.

Guardrails

Amazon Bedrock Guardrails helps you to implement customized safeguards and responsible AI policies for your generative AI applications. It provides additional customizable safety protections on top of the native protections offered by FMs. It is the only responsible AI capability offered by a major cloud provider that helps enable customers to build and customize safety, privacy, and truthfulness protections for their generative AI applications in a single solution, and it works with all FMs in Amazon Bedrock, as well as fine-tuned models. Bedrock Guardrails can also be integrated with Amazon Bedrock Agents and Amazon Bedrock Knowledge Bases to build generative AI applications aligned with your responsible AI policies. Additionally, it offers an ApplyGuardrail API to help evaluate user inputs and model responses generated by any custom or third-party FM outside of Bedrock.

Flows

Amazon Bedrock Flows is a workflow authoring and execution feature of Bedrock for generative AI applications. It accelerates the creation, testing, and deployment of user-defined generative AI workflows through an intuitive visual builder and a set of APIs. It allows you to seamlessly link the latest foundation models, Prompts, Agents, Knowledge Base, Guardrails, and AWS services (such as Amazon Lex, AWS Lambda, Amazon S3) along with business logic to build generative AI workflows. You can easily test and version your workflows, and run it in a secure serverless environment through a visual interface or API without having to stand up your own infrastructure.

Knowledge Bases

Amazon Bedrock Knowledge Bases provides a fully managed end-to-end retrieval-augmented generation (RAG) workflow, enabling FMs and agents to access contextual information from your company's private data sources. This allows them to deliver more relevant, accurate, and customized responses. You can securely connect FMs and agents to multiple data sources such as Amazon S3, Confluence, Salesforce, and SharePoint. If you don't have an existing vector database, Amazon Bedrock creates an Amazon OpenSearch Serverless vector store for you. Alternatively, you can specify an existing vector store in supported databases like Amazon

OpenSearch Serverless, Pinecone, Redis Enterprise Cloud, Amazon Aurora, and MongoDB. You can also fine-tune retrieval and ingestion to achieve better accuracy across use-cases with advanced parsing options for unstructured data, data chunking options like custom chunking, or built-in chunking strategies including default, fixed size, no chunking, hierarchical chunking, or semantic chunking.

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 charge for the workforce, as the workforce is supplied by you. For an evaluation managed by AWS, pricing is customized for your evaluation needs in a private engagement while working with the AWS expert evaluations team.

Pricing Details

Pricing is dependent on the modality, provider, and model.

AI21 Labs

On-Demand pricing

AI21 Labs models	Price per 1,000 input tokens	Price per 1,000 output tokens
Jamba 1.5 Large	$0.002	$0.008
Jamba 1.5 Mini	$0.0002	$0.0004
Jurassic-2 Mid	$0.0125	$0.0125
Jurassic-2 Ultra	$0.0188	$0.0188
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	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Amazon Titan Text Premier	$0.0005	$0.0015	N/A	N/A
Amazon Titan Text Lite	$0.00015	$0.0002	N/A	N/A
Amazon Titan Text Express	$0.0002	$0.0006	N/A	N/A
Amazon Titan Text Embeddings	$0.0001	N/A	N/A	N/A
Amazon Titan Text Embeddings V2	$0.00002	N/A	$0.00001	N/A

On-Demand and Batch pricing for multi-modal models

Amazon Titan models	Image resolution	Price per image generated for Standard quality	Price per image generated for Premium quality
Amazon Titan Image Generator v1	Smaller than 512 x 512	$0.008	$0.01
Amazon Titan Image Generator v1	Larger than 512 x 512	$0.01	$0.012
Amazon Titan Image Generator v2	Smaller than 512 x 512	$0.008	$0.01
Amazon Titan Image Generator v2	Larger than 1024 x 1024	$0.01	$0.012

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 Text Express	$0.008	$1.95	$20.50

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

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 Multimodal Embeddings	$0.0002	$1.95	$9.38

Provisioned Throughput pricing

Amazon Titan models	Price per hour per model with no commitment*	Price per hour per model unit for 1-month commitment**	Price per hour for 6-month commitment**
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 v1	N/A	$16.20	$13.00
Amazon Titan Image Generator v1 (custom models)	$23.40	$21.00	$16.85
Amazon Titan Image Generator v2	$23.40	$16.20	$13.00
Amazon Titan Image Generator v2 (custom models)	$23.40	$21.00	$16.85
Amazon Titan Multimodal Embeddings	$9.38	$8.45	$6.75

**Custom model inference is limited to 1 model unit in the no-commit option*

***Includes inference for base and custom models. You can buy 2+ model units*

Amazon Nova

Region: US East (N. Virginia)

On-Demand and Batch pricing for text generation models

Amazon Nova models	Price per 1,000 input tokens	Price per 1,000 input tokens (cache read)	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Amazon Nova Micro	$0.000035	$0.00000875	$0.00014	$0.0000175	$0.00007
Amazon Nova Lite	$0.00006	$0.000015	$0.00024	$0.00003	$0.00012
Amazon Nova Pro	$0.0008	$0.0002	$0.0032	$0.0004	$0.0016

Pricing for model customization (fine-tuning)

Amazon Nova Models	Price to train 1,000 tokens*	Price to store each custom model per month	Price to infer for 1 model unit per hour
Amazon Nova Micro	$0.001	$1.95	$108.15
Amazon Nova Lite	$0.002	$1.95	$108.15
Amazon Nova Pro	$0.008	$1.95	$108.15

Pricing for Creative Content Generation models

On-Demand pricing for Image Generator Model

Amazon Nova models	Image resolution	Price per image generated for Standard quality	Price per image generated for Premium quality
Amazon Nova Canvas	up to 1024 x 1024	$0.04	$0.06
Amazon Nova Canvas	up to 2048 x 2048	$0.06	$0.08

On-Demand pricing for Video Generator Model

Amazon Nova models	Video resolution	Price per second of video generated
Amazon Nova Reel	720p, 24 fps	$0.08

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	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3.5 Sonnet**	$0.003	$0.015	$0.0015	$0.0075
Claude 3.5 Haiku	$0.001	$0.005	$0.0005	$0.0025
Claude 3 Opus*	$0.015	$0.075	$0.0075	$0.0375
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 2.1	$0.008	$0.024	N/A	N/A
Claude 2.0	$0.008	$0.024	N/A	N/A
Claude Instant	$0.0008	$0.0024	N/A	N/A

Region: Europe (London)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625

Region: Europe (Zurich)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3.5 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625

Region: South America (Sao Paolo)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625

Region: Canada (Central)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625

Region: Asia Pacific (Mumbai)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625

Region: Asia Pacific (Sydney)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625

Region: Asia Pacific (Tokyo)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude Instant	$0.0008	$0.0024	N/A	N/A
Claude 2.0/2.1	$0.008	$0.024	N/A	N/A
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625
Claude 3.5 Sonnet	$0.003	$0.015	$0.0015	$0.0075

Region: Asia Pacific (Singapore)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude Instant	$0.0008	$0.0024	$0.0004	$0.0012
Claude 2.0/2.1	$0.008	$0.024	$0.004	$0.012
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625
Claude 3.5 Sonnet	$0.003	$0.015	N/A	N/A

Region: Europe (Paris)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075

Region: Europe (Frankfurt)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude Instant	$0.0008	$0.0024	N/A	N/A
Claude 2.0/2.1	$0.008	$0.024	N/A	N/A
Claude 3 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3.5 Sonnet	$0.003	$0.015	$0.0015	$0.0075
Claude 3 Haiku	$0.00025	$0.00125	$0.000125	$0.000625

Region: Asia Pacific (Seoul)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3.5 Sonnet	$0.003	$0.015	N/A	N/A
Claude 3 Haiku	$0.00025	$0.00125	N/A	N/A

Region: US East (Ohio)

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Claude 3.5 Sonnet	$0.003	$0.015	N/A	N/A
Claude 3 Haiku	$0.00025	$0.00125	N/A	N/A

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

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

Cohere

On-Demand pricing

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

Cohere 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
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

Meta Llama

Llama 3.2

On-Demand and Batch pricing

Meta models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Llama 3.2 Instruct (1B)	$0.0001	$0.0001	N/A	N/A
Llama 3.2 Instruct (3B)	$0.00015	$0.00015	N/A	N/A
Llama 3.2 Instruct (11B)	$0.00016	$0.00016	N/A	N/A
Llama 3.2 Instruct (90B)	$0.00072	$0.00072	N/A	N/A

Llama 3.1

On-Demand and Batch pricing

Meta models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Llama 3.1 Instruct (8B)	$0.00022	$0.00022	N/A	N/A
Llama 3.1 Instruct (70B)	$0.00099	$0.00099	N/A	N/A

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 3.1 Instruct (8B)	$0.00149	$1.95	$24.00
Llama 3.1 Instruct (70B)	$0.00799	$1.95	$24.00

Provisioned Throughput pricing

Meta models	Price per hour per model unit for no commitment	Price per hour per model unit for 1-month commitment	Price per hour per model unit for 6-month commitment
Llama 3.1 Instruct (8B)	$24.00	$21.18	$13.08
Llama 3.1 Instruct (70B)	$24.00	$21.18	$13.08

Llama 3

On-Demand pricing

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 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
Llama 2 Pretrained (70B)	$0.00799	$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

Mistral AI

Mistral models	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per 1,000 input tokens (batch)	Price per 1,000 output tokens (batch)
Mistral 7B	$0.00015	$0.0002	N/A	N/A
Mixtral 8*7B	$0.00045	$0.0007	N/A	N/A
Mistral Small (24.02)	$0.001	$0.003	$0.0005	$0.0015
Mistral Large (24.02)	$0.004	$0.012	N/A	N/A

Stability AI

On-Demand pricing

Stability AI Model	Price per generated image
Stable Image Core	$0.04
SD3 Large	$0.08
Stable Image Ultra	$0.14

Previous generation of 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 1.0	Up to 1024 x 1024	$0.04	$0.08

Provisioned Throughput pricing

Stability AI model	Price per hour per model unit for 1-month commitment*	Price per hour per model unit for 6-month commitment*
SDXL 1.0	$49.86	$46.18

Custom Model Import

Llama

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

Custom Model Unit version	v1.0
Price per Custom Model Unit per min*	$0.0785
Monthly storage cost per Custom Model Unit	$1.95

The Custom Model Units needed to host a model depend on a variety of factors - notably the model architecture, model parameter count, and context length. The exact number of Custom Model Units needed will be determined at the time of import. For reference, Llama 3.1 8B 128K model requires 2 Custom Model Units, a Llama 3.1 70B 128k model requires 8 Custom Model Units.

**Billed in 5 minute windows*

Multimodal Llama

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

Custom Model Unit version	**v1.0**
Price per Custom Model Unit per min*	$0.0785
Monthly storage cost per Custom Model Unit	$1.95

The Custom Model Units needed to host a model depend on a variety of factors - notably the model architecture, model parameter count, and context length. The exact number of Custom Model Units needed will be determined at the time of import. For reference, Llama 3.2 11B 128K model requires 4 Custom Model Units.

**Billed in 5 minute windows*

Mistral

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

Custom Model Unit version	**v1.0**
Price per Custom Model Unit per min*	$0.0785
Monthly storage cost per Custom Model Unit	$1.95

The Custom Model Units needed to host a model depend on a variety of factors - notably the model architecture, model parameter count, and context length. The exact number of Custom Model Units needed will be determined at the time of import. For reference, Mistral 7B 32K model requires 1 Custom Model Unit.

**Billed in 5 minute windows*

Mixtral

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

Custom Model Unit version	v1.0
Price per Custom Model Unit per min*	$0.0785
Monthly storage cost per Custom Model Unit	$1.95

The Custom Model Units needed to host a model depend on a variety of factors - notably the model architecture, model parameter count, and context length. The exact number of Custom Model Units needed will be determined at the time of import. For reference, Mixtral 8×7B 32K model requires 4 Custom Model Units.

**Billed in 5 minute windows*

Flan

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

Custom Model Unit version	v1.0
Price per Custom Model Unit per min*	$0.0785
Monthly storage cost per Custom Model Unit	$1.95

The Custom Model Units needed to host a model depend on a variety of factors - notably the model architecture, model parameter count, and context length. The exact number of Custom Model Units needed will be determined at the time of import. For reference, Flan-T5 XL 512 model requires 1 Custom Model Unit.

**Billed in 5 minute windows*

Pricing Advanced Tools (Details)

Amazon Bedrock Flows

You are charged based on the number of node transitions required to execute your application. Bedrock Flows counts a node transition each time a node in your workflow is executed. You are charged for the total number of node transitions across all your flows.

All charges are metered daily and billed monthly starting February 1st, 2025.

Price per 1,000 node transitions

$0.035

Additional Charges

You may incur additional charges if the execution of your application workflow utilizes other AWS services or transfers data. For example, if your workflow invokes an Amazon Bedrock Guardrail policy, you will be billed for the number of text units processed by the policy.

Amazon Bedrock Guardrails

Guardrail policy*	**Price per 1,000 text units****
Content filters	$0.15
Denied topics	$0.15
Contextual grounding check***	$0.1
Sensitive information filter (PII)	$0.1
Sensitive information filter (regular expression)	Free
Word filters	Free

On-Demand pricing

**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.

Model Evaluation

Model evaluation is charged for the inference from your choice of model. Automatically-generated algorithmic scores are provided at no extra charge. For human-based evaluation where you bring your own workstream, you are charged for the model inference in the evaluation, and a charge of $0.21 per completed human task.

Model	Price per 1,000 input tokens	Price per 1,000 output tokens	Price per human task
Model selected for evaluation	Based on model selected	Based on model selected	$0.21

Knowledgebase

The Knowledgebase feature does not have its own dedicated pricing model. Instead, the cost is determined by the underlying components it leverages. Specifically, it involves the cost of using the foundation models, which are charged based on inference and customization, combined with the cost of storing and retrieving data in the OpenSearch vector store. The pricing depends on both the compute resources used for model inference and the storage and indexing operations performed in OpenSearch, providing flexibility in usage but requiring awareness of the combined service costs.

Pricing examples

AI21 labs

- An application developer makes the following API calls to Amazon Bedrock: a request to AI21's Jurassic-2 Mid model to summarize an input of 10K tokens of input text to an output of 2K tokens.

Total cost incurred = 10K tokens/1000 * $0.0125 + 2K tokens/1000 * $0.0125 = $0.15

Amazon

On-Demand pricing

An application developer makes the following API calls to Amazon Bedrock on an hourly basis: a request to Amazon Titan Text Lite model to summarize an input of 2K tokens of input text to an output of 1K tokens.

Total hourly cost incurred is = 2K tokens/1000 * $0.0003 + 1K tokens/1000 * $0.0004 = $0.001.

An application developer makes the following API calls to Amazon Bedrock: a request to the Amazon Titan Image Generator base model to generate 1000 images of 1024 x 1024 in size of standard quality.

Total cost incurred = 1000 images * $0.01 per image = $10

Customization (fine-tuning and continued pretraining) pricing

An application developer customizes an Amazon Titan Image Generator model using 1000 image-text pairs. After training, the developer uses custom model provisioned throughput for 1 hour to evaluate the performance of the model. The fine-tuned model is stored for 1 month. After evaluation, the developer uses provisioned throughput (1-month commitment term) to host the customized model.

Monthly cost incurred for fine-tuning = fine-tuning training ($.005 * 500 * 64), where $0.005 is the price per image seen, 500 is the number of steps, and 64 is the batch size, + custom model storage per month ($1.95) + 1 hour of custom model inference ($21) = $160 + $1.95 + 21 = $182.95

Provisioned Throughput pricing

An application developer buys two model units of Amazon Titan Text Express with a 1-month commitment for their text summarization use case.

Total monthly cost incurred = 2 model units * $18.40/hour * 24 hours * 31 days = $27,379.20

An application developer buys one model unit of the base Amazon Titan Image Generator model with a 1-month commitment.

Total cost incurred = 1 model unit * $16.20 * 24 hours * 31 days = $12,052.80

Anthropic

On-Demand pricing

An application developer makes the following API calls to Amazon Bedrock in the US West (Oregon) Region: a request to Anthropic's Claude model to summarize an input of 11K tokens of input text to an output of 4K tokens.

Total cost incurred = 11K tokens/1000 * $0.008 + 4K tokens/1000 * $0.024 = $0.088 + $0.096 = $0.184

Provisioned Throughput pricing

An application developer buys one model unit of Anthropic Claude Instant in the US West (Oregon) Region:

Total monthly cost incurred = 1 model unit * $39.60 * 24 hours * 31 days = $29,462.40

Cohere

On-Demand pricing

An application developer makes the following API calls to Amazon Bedrock: a request to Cohere's Command model to summarize an input of 6K tokens of input text to an output of 2K tokens.

Total cost incurred = 6K tokens/1,000 * $0.0015 + 2K tokens/1,000 * $0.0020 = $0.013

An application developer makes the following API calls to Amazon Bedrock: A request to Cohere's Command - Light model to summarize an input of 6K tokens of input text to an output of 2K tokens.

Total cost incurred = 6K tokens/1000 * $0.0003 + 2K tokens/1000 * $0.0006 = $0.003

An application developer makes the following API calls to Amazon Bedrock: A request to either Cohere's Embed English or Embed Multilingual model to generate embeddings for 10K tokens of input.

Total cost incurred = 10K tokens/1000 * $0.0001 = $.001

Customization (fine-tuning) pricing

An application developer customizes a Cohere Command model using 1000 tokens of data. After training, uses custom model provisioned throughput for 1 hour to evaluate the performance of the model. The fine-tuned model is stored for 1 month. After evaluation, the developer uses provisioned throughput (1mo commit) to host the customized model.

Monthly cost incurred for fine-tuning = Fine-tuning training ($0.004 * 1000) + custom model storage per month ($1.95) + 1 hour of custom model inference ($49.50) = $55.45

Monthly cost incurred for provisioned throughput (1-month commitment) of custom model = $39.60

Provisioned Throughput pricing

An application developer, buys one model unit of Cohere Command with a 1-month commitment for their text summarization use case.

Total monthly cost incurred = 1 model unit * $39.60 * 24 hours * 31 days = $29,462.40

Meta Llama

On-Demand pricing

An application developer makes the following API calls to Amazon Bedrock: a request to Meta's Llama 2 Chat (13B) model to summarize an input of 2K tokens of input text to an output of 500 tokens.

Total cost incurred = 2K tokens/1000 * $0.00075 + 500 tokens/1000 * $0.001 = $0.002

Customization (fine-tuning) pricing

An application developer customizes the Llama 2 Pretrained (70B) model using 1000 tokens of data. After training, uses custom model provisioned throughput for 1 hour to evaluate the performance of the model. The fine-tuned model is stored for 1 month. After evaluation, the developer uses provisioned throughput (1mo commit) to host the customized model.

Monthly cost incurred for fine-tuning = Fine tuning training ($0.00799 * 1000) + custom model storage per month ($1.95) + 1 hour of custom model inference ($23.50) = $33.44

Monthly cost incurred for provisioned throughput (a 1-month commit) of custom model = $21.18

Provisioned Throughput pricing

An application developer buys one model unit of Meta Llama 2 with a 1-month commitment for their text summarization use case.

Total monthly cost incurred = 1 model unit * $21.18 * 24 hours * 31 days = $15,757.92

Mistral AI

On-Demand pricing

An application developer makes the following API calls to Amazon Bedrock on an hourly basis: a request to Mistral 7B model to summarize an input of 2K tokens of input text to an output of 1K tokens.

Total hourly cost incurred = 2K tokens/1000 * $0.00015 + 1K tokens/1000 * $0.0002 = $0.0005

An application developer makes the following API calls to Amazon Bedrock on an hourly basis: a request to Mixtral 8×7B model to summarize an input of 2K tokens of input text to an output of 1K tokens.

Total hourly cost incurred = 2K tokens/1000 * $0.00045 + 1K tokens/1000 * $0.0007 = $0.0016

An application developer makes the following API calls to Amazon Bedrock on an hourly basis: a request to Mistral Large model to summarize an input of 2K tokens of input text to an output of 1K tokens.

Total hourly cost incurred = 2K tokens/1000 * $0.008 + 1K tokens/1000 * $0.024 = $0.04

Stability AI

On-Demand pricing

An application developer makes the following API calls to Amazon Bedrock: a request to the SDXL model to generate a 512 x 512 image with a step size of 70 (premium quality).

Total cost incurred = 1 image * $0.036 per image = $0.036

An application developer makes the following API calls to Amazon Bedrock: A request to the SDXL 1.0 model to generate a 1024 x 1024 image with a step size of 70 (premium quality).

Total cost incurred = 1 image * $0.08 per image = $0.08

Provisioned Throughput pricing

An application developer buys one model unit of SDXL 1.0 with a 1-month commitment.

Total cost incurred = 1 * $49.86 * 24 hours * 31 days = $37,095.84

Model Evaluation

Model evaluation example 1:

On-demand pricing

An application developer submits a dataset for human-based model evaluation using Anthropic Claude 2.1 and Anthropic Claude Instant in the US East (N. Virginia) AWS Region.

The dataset contains 50 prompts, and the developer requires one worker to rate each prompt-response set (configurable in the evaluation job creation as “workers per prompt” parameter).

There will be 50 tasks in this evaluation job (one task for each prompt-response set per each worker). The 50 prompts combine to 5000 input tokens, and the associated responses combine to 15,000 tokens for Anthropic Claude Instant and 20,000 tokens for Anthropic Claude 2.1.

The following charges are incurred for this model evaluation job:

Item	Number of input tokens	Price per 1000 input tokens	Cost of input	Number of output tokens	Price per 1000 output tokens	Cost of output	Number of human tasks	Price per human task	Cost of human tasks	Total
Claude Instant Inference	5000	$0.0008	$0.004	15000	$0.0024	$0.036				$0.04
Claude 2.1 Inference	5000	$0.008	$0.04	20000	$0.024	$0.48				$0.52
Human Tasks							50	$0.21	$10.50	$10.50
Total										$11.06

Model evaluation example 2:

On-demand pricing

An application developer submits a dataset for human-based model evaluation using Anthropic Claude 2.1 and Anthropic Claude Instant in the US East (N. Virginia) AWS Region.

The dataset contains 50 prompts, and the developer requires two workers to rate each prompt-response set (configurable in the evaluation job creation as “workers per prompt” parameter). There will be 100 tasks in this evaluation job (1 task for each prompt-response set per each worker: 2 workers x 50 prompt-response sets = 100 human tasks).

The 50 prompts combine to 5000 input tokens, and the associated responses combine to 15000 tokens for Anthropic Claude Instant and 20000 tokens for Anthropic Claude 2.1.

The following charges are incurred for this model evaluation job:

Item	Number of input tokens	Price per 1000 input tokens	Cost of input	Number of output tokens	Price per 1000 output tokens	Cost of output	Number of human tasks	Price per human task	Cost of human tasks	Total
Claude Instant Inference	5000	$0.0008	$0.0040	15000	$0.0024	$0.036				$0.04
Claude 2.1 Inference	5000	$0.008	$0.0400	20000	$0.024	$0.48				$0.52
Human Tasks							100	$0.21	$21.00	$21.00
Total										$21.56

Amazon Bedrock Guardrails

Example 1: Customer support chatbot

An application developer creates a customer support chatbot and uses content filters to block harmful content and denied topics to filter undesirable queries and responses.

The chatbot serves 1000 user queries per hour. Each user query has an average input length of 200 characters and receives a FM response of 1500 characters.

Each user query of 200 characters correspond to 1 text unit.

Each FM response of 1,500 characters correspond to 2 text units.

Text units processed each hour = (1 + 2) * 1000 queries = 3000 text units

Total cost incurred per hour for content filters and denied topic = 3000 * ($0.75 + $1.00) / 1000 = $5.25

Example 2: Call center transcript summarization

An application developer creates an application to summarize chat transcripts between users and support agents. It uses sensitive information filter to redact personally identifiable information (PII) in the generated summaries for 10,000 conversations.

Each generated summary has an average of 3,500 characters that corresponds to 4 text units.

Total cost incurred to summarize 10,000 conversations = 10000 * 4 * ($0.1/1000) = $4

Amazon Bedrock Knowledge Bases

Pricing Example 1 (Reranking using Amazon Rerank 1.0 model)

In a given month, you make 2 million requests to Rerank API using Amazon Rerank 1.0 model – 1 million requests contain fewer than 100 documents each and hence will be charged for one request each. The remaining 1 million requests contain 120-150 documents, and hence each request will be charged for 2 requests.

Price for one request = $0.001

Total charge = 1,000,000 * $0.001 + 1,000,000*2*$0.001= $3000

Pricing Example 2: (Structured data retrieval)

An application developer creates a support chatbot that queries structured data stored in Amazon Redshift. The developer creates a Bedrock Knowledge Base and connects to Amazon Redshift. The chatbot serves 10000 user queries per hour. Each user query will cost $0.002 per GenerateQuery API to generate SQL from user query.

Total cost incurred for generating SQL per hour = $0.002*10000 = $20.

Total cost incurred in month = $20*24*30 = $1440

Custom Model Import

Pricing Example: An application developer imports a customized Llama 3.1 type model that is 8B parameter in size with a 128K sequence length in us-east-1 region and deletes the model after 1 month. This requires 2 Custom Model Units. So, the price per minute will be $0.1570 because 2 Custom Model Units are required. The model storage costs for 2 Custom Model Units would be $3.90 for the month.

There is no charge to import the model. The first successful invocation is at 8:03 AM, at which time the metering starts. The 5-minute metering windows are from 8:03 AM - 8:07 AM; 8:07 AM - 8:11 AM, and so on. If there is at least one invocation during any 5-minute period, the window will be considered active for billing. If there is no invocation from 8:07 AM - 8:11 AM, the metering will stop at 8:11 AM. In this case, the bill would be calculated as follows: $0.1570 * 5 minutes * 3 five minute windows = $2.355.

Flows

Example: News summarization

An application developer creates a flow to automate news summarization for traders. The flow includes an Input node that takes in an array of 10 S3 locations for articles from 10 major news agency (1 node transition). It then uses an iterator node to iterate through the 10 locations, retrieve the file from each S3 location using the S3 retrieval node, and invoke a model with a prompt node to summarize each file (+ 10 files x 3 node transitions). It then collects all the results using a collector node, write the results to S3 using S3 storage node, and complete in an Output node (+ 3 node transitions). They run this flow every half hour of every week day.

The number of node transition per flow execution is: 1 + 10*3 + 3 = 34 node transitions/flow execution

The number of flow execution per month is: 24 hours *2* 5 days * 4 weeks = 960 flow executions/month.

Total monthly bill is: 34 * 960 * $0.035/1000 = $1.14

Additional charges

The bill will also include additional charges for AWS services used in the workflow execution, including Amazon S3 usages in the retrieval and storage nodes, and Amazon Bedrock foundation model usage in the prompt node.

Data Automation

Pricing example 1:

Let's say you process a 1,000 page document using BDA Custom Output. All 1,000 pages are processed using blueprint 1 which has 15 fields. The per page price for any blueprint with 30 fields or less is $0.040. The total cost would be $40.

Total pages processed = 1,000

Price per page for blueprints with less than 30 fields = $0.040

Total charge = 1,000 * $0.040 = $40

Pricing example 2:

Let's say you process 2 documents using BDA Custom Output. Document 1 has 40 pages and is processed using blueprint 1 which has 20 fields. Document 2 has 10 pages and is processed using blueprint 2, which has 40 fields. The per page price of blueprint 1 is $0.040 since it

contains 30 fields or less. The per page price of blueprint 2 is $0.045. The processing cost for Document 1 using blueprint 1 is $1.60. The processing cost for Document 2 using blueprint 2 is $0.45. The total cost of processing both documents would be $2.05.

Total pages processed = 50

Price per page for Blueprint 1 with less than 30 fields = $0.040

Price per page for Blueprint 2 with 40 fields = $0.040 + (# of additional fields above 30 *$0.0005 per field)

Number of additional fields above 30 = 40 - 30 = 10

Price per page for Blueprint 2 with 40 fields = $0.040 + (10 *$0.0005 per field) = $0.045

Charge for Document 1 using Blueprint 1 = 40 pages x $0.040 per page = $1.6

Charge for Document 2 using Blueprint 2 = 10 pages x $0.045 per page = $0.45

Total charge = Charge for Document 1 + Charge for Document 2 = $1.6 + $0.45 = $2.05

Pricing example 3:

Let's say you process a 60 minute video using BDA Standard Output. The per minute price for video standard output is $0.050. The total cost would be $3.00.

Total minutes processed = 60

Price per minute for video standard output = $0.050

Total charge = 60 * $0.050 = $3.00

Pricing example 4:

Let's say you process 2,000 images using BDA Custom Output. The first 1,000 images are processed using blueprint 1, which has 10 fields. The last 1,000 pages are processed using blueprint 2, which has 40 fields. The per image price for blueprint 1 is $0.005, since it contains 30 fields or less. The per image price of blueprint 2 is $0.01. The processing cost for the first 1,000 images using blueprint 1 is $5.00. The processing cost for the second 1,000 images using blueprint 2 is $10.00. The total cost of processing all 2,000 images would be $15.00

Cost for first 1000 images = 1,000 images * $0.005 per image = $5.00

Cost for second 1,000 images = 1,000 images * ($0.005 + (# of additional fields above 30 *$0.0005 per field))*

*= 1,000 * ($0.005 + ((40-30)*$0.0005))*

= 1,000 * ($0.005 + (10*$0.0005)) = $10.00

Total cost = $5.00 + $10.00 = $15.00

Pricing example 5:

Let's assume that you want to use Bedrock Data Automation Standard Output to process 15,000 minutes of meeting audio recordings in your organization. The total cost of processing all 15,000 audio minutes would be $90.

Total minutes processed = 15,000 minutes

Total charge = 15,000 min × $0.006 = $90

Pricing Example 6:

Let's say you setup Bedrock Knowledge Bases to use Bedrock Data Automation as a parser and then ingest a 1000 page document. Note, that the Bedrock Knowledge Bases and Bedrock Data Automation integration uses standard output. The per page price for standard output is $0.010. The total cost would be $10.

Total pages processed = 1,000

Price per page for standard output = $0.010

Total charge = 1,000 * $0.010 = $10

OpenSearch Vector Store Pricing Calculator

https://calculator.aws/#/createCalculator/OpenSearchService

OCUs (OpenSearch Compute Units)

- **OpenSearch Compute Units (OCUs)** are a measure of the compute capacity available to an OpenSearch domain. They are used for allocating a balanced amount of CPU, memory, and storage resources to meet your workload requirements.
- OCUs help standardize and scale OpenSearch cluster capacity to ensure consistent performance as workloads increase.
- Each OCU typically consists of a certain amount of **vCPUs, RAM**, and **storage**. This enables predictable scaling by adding or removing OCUs based on your performance and scalability requirements.

Search and Query OCUs

- **Search and Query OCUs** are a type of OCU specifically optimized for handling search and query requests in an OpenSearch domain.

- These OCUs are configured to ensure that the domain has sufficient resources for **processing searches, handling complex queries**, and managing data indexing operations efficiently.
- By increasing **Search and Query OCUs**, you effectively increase the capacity and speed at which your OpenSearch domain can respond to search requests and process data.

EDT &
PARTNERS

www.edtpartners.com
+1 4156887260 (U.S)
+44 1274089806 (Europe)
info@edtpartners.com