Amazon Nova Multimodal Embeddings

AWS AI Service Cards

AWS AI Service Cards: Amazon Nova Multimodal Embeddings

Table of Contents

Amazon Nova Multimodal Embeddings

An AWS AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the responsible design and use of the service. A Service Card will evolve as AWS receives customer feedback, and as the service progresses through its lifecycle. AWS recommends that customers assess the performance of any AI service on their own content for each use case they need to solve. For more information, please see AWS Responsible Use of AI Guide and the references at the end. Please also be sure to review the AWS Responsible AI Policy, AWS Acceptable Use Policy, and AWS Service Terms for the services you plan to use.

This Service Card applies to the releases of Amazon Nova Multimodal Embeddings that are current as of 10/28/2025.

Overview of Amazon Nova Multimodal Embeddings

Amazon Nova Multimodal Embeddings is a proprietary multimodal foundation model (FM) designed for enterprise use cases. Amazon Nova Multimodal Embeddings enables developers to create intuitive applications by converting different content types—such as text, documents, images, video, and audio - into a universal numerical format known as an embedding. This format is understood by artificial intelligence systems allowing them to compare information across different datatypes. For example, an embedding model can understand that the text "dog" and an image of a dog are similar concepts. This ability to translate information across modalities enables embedding models to create powerful search applications by allowing users to search based off intention and not by key-word matching. Customers can use Amazon Nova Multimodal Embeddings for tasks such as multimodal semantic search, agentic retrieval-augmented generation (RAG), and classification. For example, an e-commerce product search engine can create embeddings from images that can power intuitive search experiences where a user can describe what they are looking for in natural language as opposed to searching for key words (for example, "what are great outfit ideas for an occasion where someone is going to an outdoor wedding in <insert location>").

This AI Service Card applies to the use of Amazon Nova Multimodal Embeddings via the Amazon Bedrock API. Each Amazon Nova Multimodal Embeddings is a managed sub-service of Amazon Bedrock customers can focus on executing prompts without having to provision or manage any infrastructure such as instance types, network topology, and endpoint. Not all of the content in

the service card is applicable to models hosted on https://nova.amazon.com, a publicly available website where individuals may try certain Amazon Nova models.

Amazon Nova Multimodal Embeddings <model input, generated embeddings>, {pair} is said to be "effective" if an evaluator decides that the embedding provides a strong representation of the input content. For example, the evaluator can determine if an image of a "sunny day on the beach" is correctly represented by comparing it to the corresponding text embedding. The evaluator will assess quality by determining if it 1/ exceeds a predefined quality threshold, such as a common retrieval metrics, 2/ assessing similarity scores between like concepts with similar and varied descriptions3/ confirming that the embeddings close in meaning occupy are close together in an embedding space. Otherwise, the embedding pair is considered to be "ineffective". Amazon Nova Multimodal Embeddings do not provide a confidence score for the embeddings they generate; a customer's workflow must decide if the model is effective using human judgment, whether human judgement is applied on a case-by-case basis or is applied via the customer's choice of an acceptable score on an automated test. The "overall effectiveness" of any embedding model for a specific use case is based on the percentage of use-case specific inputs for which the model returns an effective result.

Customers should define and measure effectiveness for themselves for the following reasons. First, the customer is best positioned to know which modalities and domains will best represent their use case, and should therefore be included in an evaluation dataset. Second, different embedding models may respond differently to the same input and use-case and hence specific evaluations may need to be conducted to take it into account.

As with all ML solutions, Amazon Nova Multimodal Embeddings must overcome issues of intrinsic and confounding variation. Intrinsic variation refers to features of the input that the model should be able to discern between, for example, knowing the difference between an image of a house cat and that of a lion. Confounding variation refers to input features that unintentionally influence the embedding, even though they are irrelevant to the true meaning of the content. For example, in image embeddings, background and lighting conditions can cause two different images to appear more similar than they should. An image of a lion in a grassy savanna may be embedded close to an image of an elephant in a similar savanna setting—not because they depict the same animal, but because it also considers the shared background in as part of the information to include in the embedding. The full set of variations encountered in the set of model inputs include professional jargon, lighting, camera angle, occlusions (partial visibility of main object), color vs. grayscale, writing style, spelling errors, and speaker accent.

Intended use cases and limitations

Amazon Nova Multimodal Embedding serves a wide range of potential application domains but are optimized for use-cases such as multimodal RAG, semantic search, and clustering:

- **Agentic Retrieval-Augmented Generation (RAG):** customers can use Amazon Nova Multimodal Embeddings for RAG-based applications where the model serves as the embedding for the retrieval task. The input can be text from documents, images, or document images that interleave text with infographics, video and audio. The embedding enables the application developer to retrieve the most relevant information from a knowledge base that can be provided to an LLM system for improved responses.
- **Semantic Search:** customers can generate embeddings from text, images, document images, video and audio to power search applications that are stored in a vector index, a specialized embedding space that reduces the number of comparisons needed to return an effective result. Because the model is able to capture the nuance of a user's query within the embedding, it enables advanced search queries that do not have to rely on key-word matching.
- **Clustering:** customers can use Amazon Nova Multimodal Embeddings to generate embeddings from text, images, document images, video and audio. Clustering algorithms can group together items that are close to each other based on distance or similarity. For example, a customer in media management may want to categorize their media assets across similar themes. In this case, the embeddings can be used to cluster together similar assets without the need for meta data associated with each asset.

The model provides optionality of different parameters (for example, embedding lengths and segment sizes) that allow customers to tailor the embeddings to a specific use-case. For more information about these specifications, see the Amazon Nova User Guide.

When assessing an embedding model for a particular use case, we encourage customers to specifically define the use case, i.e., by considering at least the following factors: the **business problem** being solved; the **stakeholders** in the business problem and deployment process; the **workflow** that solves the business problem, with the model and oversight as components; key system **inputs and outputs**; the expected intrinsic and confounding **variation**; and the types of **errors** possible and the relative impact of each.

Consider the following use case of utilizing Amazon Nova Multimodal Embeddings as a tool to help a multimodal semantic search application for a digital asset management company. The **business goal** is to find and retrieve assets (images, videos, audio, documents, etc.) based on

their meaning or content, rather than just keywords or metadata, across multiple types of media. The **stakeholders** include the application end-user, who wants to find and retrieve the most relevant assets, and the search application developer who wants to use embeddings of the assets to enable the search and retrieval experience. We will use an example of a video clip as the asset. The **workflow** is: 1/ the application developer uses Amazon Nova Multimodal Embeddings to generate embeddings for the video clips; 2/ The embeddings are stored in a vector database; 3/ when the end-user searches for a video, the application developer uses Amazon Nova Multimodal Embeddings to generate a corresponding embedding for the search query; 4/ the application then compares how close in similarity the embedding of the search query is compared to the video clips.. It retrieves the video clip that best represents the search query based on a similarity metric (such as cosine similarity) 5/ the shows the corresponding video clip to the end-user. **Inputs** to generate embeddings can be in text, image, document image, video, or audio form. The inputs refer to both the items that are used to create the index and the end-user search query. **Outputs** from the model refer to the embedding which is leveraged by the application developer to retrieve the asset that best matches the query which is displayed to the end-user. Input **variations** include input video camera angle, input image color gradient, input text length, and input document image quality. The **error types**, ranked in order of estimated negative impact on stakeholders, include: 1/ retrieved items are not relevant to the query; 2/ the most relevant item is not retrieved; 3/ retrieved document contains the opposite or contradictory information of the query; 4/ inconsistent retrieval results on similar queries. With this in mind, we would expect the Application developer to test a set of example asset and query pairs by generating embeddings in the Bedrock API.

An example input could be a/ a text string such as *"The quick brown fox jumps over the lazy dog"*, b/ an image of a bowl of fruit, c/ a video scene of a car race, d/ stock audio footage. The model output is a vector in one of four lengths: 3072, 1024, 384, 256. To evaluate the quality of the model's output, the application developer would compare the similarity score of different inputs items. Similar items should result in a higher score whilst dissimilar items should result in a lower score. After continued experimentation in the Bedrock API, the application developer should finalize their own measure of effectiveness based on the impact of errors, run a scaled-up test via the Bedrock API and use the results of human judgements (with multiple judgements per test prompt) to establish a benchmark effectiveness score.

Amazon Nova Multimodal Embeddings is not intended to support any prohibited practices under the EU AI Act or any other relevant law. Nova Multimodal Embeddings can be integrated into an array of systems such as semantic search, Digital Asset Management (DAM) platforms, RAG applications, e-commerce recommendations, document clustering, and text classification. For more

technical information about how Amazon Nova Multimodal Embeddings may be integrated into AI systems, see the Amazon Nova User Guide. All Amazon Nova Multimodal Embeddings use cases must comply with the AWS Acceptable Use Policy.

Amazon Nova Multimodal Embeddings has a number of limitations requiring careful consideration.

Appropriateness for Use

Because its output is probabilistic Amazon Nova Multimodal Embeddings may result in inaccurate or inappropriate content. Customers should evaluate outputs for accuracy and appropriateness for their use case, especially if they will be directly surfaced to end users. Additionally, if Amazon Nova Multimodal Embeddings is used in customer workflows that produce consequential decisions, customers must evaluate the potential risks of their use case and implement appropriate human oversight, testing, and other use case-specific safeguards to mitigate such risks. For more information, see the AWS Responsible AI Policy.

Inputs and Outputs

The model supports input texts of up to 8,172 tokens, images of up to 1 image per pair, video of up to 30s, and audio of up to 30s. It accepts inputs via both sync and async APIs (with optionality for batch inference) and supports four output embedding lengths (with 3072-dimension as default), 1024, 256, 384), see Amazon Nova User Guide.

Supported Natural Languages

We curated data from over 200 languages for Amazon Nova Multimodal Embeddings, It accepts inputs via both sync and async APIs (with optionality for batch inference) and supports four output embedding lengths (with 3072-dimension as default, 1024, 256, 384), see Amazon Nova User Guide.

Programming Languages

Amazon Nova Multimodal Embeddings is not trained on programming languages.

Knowledge Cutoff Date

Amazon Nova Multimodal Embeddings is trained on data up until the model release date of October 28, 2025. The model will not be aware of new information beyond this date.

Design of Amazon Nova Multimodal Embeddings

Machine Learning

Amazon Nova Multimodal Embeddings converts model inputs into embedding using machine learning, specifically, using neural networks built on a transformer architecture. At a high level, the core model works by taking in the inputs and encoding them into numerical vectors. The vectors (optionally) pass through a dimension reduction layer to output the embedding depending on the users selected dimension length. Amazon Nova Multimodal Embeddings is available pursuant to the AWS Customer Agreement or other relevant agreements with AWS.

Controllability

We say that an Amazon Nova Multimodal Embeddings model exhibits a particular "behavior" when it generates the same kind of output for the same kinds of prompts and configuration (for example, segmentation size and embedding length). For a given model architecture, the control levers that we have over the behaviors are primarily the unlabeled pre-training data corpus and the contrastive learning, meaning that the model can understand when information is either similar or exactly the same. Our development process exercises these control levers as follows: 1/ We pre-train Amazon Nova Multimodal Embeddings using curated data from a variety of sources, which may include licensed and proprietary data, open-source datasets, and publicly available data where appropriate. By constructing a well-balanced data mixture, the model can learn generalizable patterns; 2/ During the contrastive learning stage, we shape the embedding space so that the embeddings of semantically similar or dissimilar embeddings are correspondingly nearer or further apart.

Performance Expectations

Intrinsic and confounding variation differ between customer applications. This means that performance will also differ between applications, even if they support the same use case. Consider two applications A and B. With Application A, a media application first uses an LLM to generate detailed text captions for a video, then converts those captions into a text embedding. Application A must address multiple issues such as captioning errors, variations in text quality and level of detail and text embedding accuracy. With Application B, a media application generates native video embeddings. Application B must contend with shorter context lengths and missed events. Because A and B have differing kinds of inputs, they will likely have different embedding qualities even if each application is deployed perfectly. As a result, the overall utility of Amazon Nova Multimodal Embeddings will depend both on the model and on workflows it enables. Performance results depend on a variety of factors including Amazon Nova Multimodal

Embeddings itself, the customer workflow, and the evaluation dataset, we recommend that customers test Amazon Nova Multimodal Embeddings using their own content.

Test Driven Methodology

We use multiple datasets and human teams to evaluate the performance of Amazon Nova Multimodal Embeddings models. No single evaluation dataset suffices to completely capture performance. This is because evaluation datasets vary based on use case, intrinsic and confounding variation, the quality of ground truth available, and other factors. Our development testing involves automated testing against publicly available and proprietary datasets, benchmarking against proxies for anticipated customer use cases, human evaluation of generations against proprietary datasets, manual red-teaming, and more. Our development process examines Amazon Nova Multimodal Embeddings's performance using all of these tests, takes steps to improve the model and the suite of evaluation datasets, and then iterates. In this Service Card, we provide examples of test results to illustrate our methodology.

Automated Evaluations provide a quantitative mechanism for evaluating the model's performance on tasks such as retrieval against candidate models in a like-for-like manner. Automated assessments can take several forms. One form is to see if the numerical representation of a text caption is sufficiently close to the numerical representation of the input item to be retrieved. One industry standard evaluation metric is net discounted cumulative gain (NDGC@k).

Safety

Safety is a shared responsibility between AWS and our customers. Our goal for safety is to mitigate key risks of concern to our enterprise customers, and to society more broadly. We align the behaviors of our foundation models with internal design policies and our commitment to responsible AI development practices. Amazon is committed to producing generative AI services that keep child safety at the forefront of development, deployment and operation, and conduct testing and implement mitigations to prevent Amazon Nova Multimodal Embeddings from accepting image inputs of inappropriate content related to children. In the case where a customer tries to retrieve an image using an embedding generated with Amazon Nova Multimodal Embeddings if no safety filters are triggered, it returns an image. In the case where the model cannot complete a prompt due to the possibility of generating an undesired output, it will not generate an embedding and return an error message.

Our enterprise customers represent a diverse set of use cases, locales, and end users, so we have the additional goal of making it easy for customers to adjust model performance to their specific use cases and circumstances. AWS offers services and tools to help customers identify

and mitigate safety risks, such as Amazon Bedrock Guardrails and Amazon Bedrock Model Evaluations. Customers are responsible for end-to-end testing of their applications on datasets representative of their use cases and any additional safety mitigations, and deciding if test results meet their specific expectations of safety, fairness, and other properties, as well as overall effectiveness.

Child Sexual Abuse Material (CSAM)

Amazon Nova Multimodal Embeddings utilizes Amazon Bedrock Abuse Detection solution, which uses hash matching or classifiers to detect potential CSAM. If Amazon Bedrock detects apparent CSAM in Amazon Nova Multimodal Embeddings user image inputs it will block the request, display an automated error message and may also file a report with the National Center for Missing and Exploited Children (NCMEC) or a relevant authority. We take CSAM commitments seriously and will continue to update our detection, blocking, and reporting mechanisms.

Chemical, Biological, Radiological, and Nuclear (CBRN)

Chemical, Biological, Radiological, and Nuclear (CBRN): Compared to information available via internet searches, science articles, and paid experts, we see no indications that Amazon Nova Multimodal Embeddings increases access to information about chemical, biological, radiological or nuclear (CBRN) threats. We continue to assess for CBRN risk, and engage with other third-party researchers or vendors to share, learn about, and mitigate possible CBRN threats and vulnerabilities.

Fairness

Despite the challenges in building training datasets that capture every real-world scenario, we aimed to combat societal bias and cultural appropriation. We tested Amazon Nova Multimodal Embeddings' ability to moderate these outcomes using a proprietary dataset of aggregated red teaming iterations that depict bias, stereotyping, and hate against individuals and groups. When users provide no guidance about the desired attributes of an object or person, it is unclear how to judge search query embeddings. For example, for the prompt "basketball players", some users might prefer a team with similar demographic attributes and other might want a distribution of attributes (for example, gender) matching some distribution they have in mind. Given this ambiguity, when there is no information included in the prompt, Amazon Nova Multimodal Embeddings is designed to return diverse results, but without specifying a desired distribution. Amazon Nova Multimodal Embeddings is designed with controls to help address this. For example, when retrieving images with the phrase “engineer” or “nurse”, the controls will return a more balance gender distribution than before. Given the intrinsic ambiguity of

generating embeddings with associations to people, groups of people, or occupations without guidance, we recommend that customers consider specifying attributes in the search query. We acknowledge that while the system is designed for fair results, responsible development is a constant evolution and we welcome feedback on fairness findings that can be used to improve the overall system.

Robustness

Amazon Nova Multimodal Embeddings is designed to accept text, images, document images, video, and audio as inputs. The output is limited to embeddings. Amazon Nova Multimodal Embeddings is designed to perform consistently well across a wide range of use cases. We evaluated the embedding quality on each modality and also across modalities. We measured model robustness using the following benchmarks: 1/ Text: Massive Multilingual Text Embedding Benchmark (MMTEB) (Multilingual, v1); 2/ Image: M-BEIR (MSCOCO Flikr30k); 3/ Document image: ViDoRe v2 benchmark; 4/ Video: ActivityNet, DiDeMo, MSRVTT; 5/ Audio: AudioCaps.

Explainability

Customers wanting to check whether similar inputs are positioned closely in the embedding space can measure cosine similarities between the output vectors.

Privacy

Amazon Nova Multimodal Embeddings is available in Amazon Bedrock. Amazon Bedrock is a managed service and does not store or review customer prompts or customer image generations, and prompts and generations are never shared between customers, or with Amazon Bedrock third party model providers. AWS does not use inputs or outputs generated through the Amazon Bedrock service to train Amazon Bedrock models, including Amazon Nova Multimodal Embeddings. For more information, see Section 50.3 of the AWS Service Terms and the AWS Data Privacy FAQs. For service-specific privacy information, see Security in the Amazon Bedrock FAQs. Amazon Nova models are designed to avoid completing prompts or generating outputs that would violate our content policies (for example, prompts seeking private information). If a user is concerned that their personal information has been included in an Amazon Nova model output, the user should contact us here.

Security

All Amazon Bedrock models, including Amazon Nova Multimodal Embeddings, come with enterprise security that enables customers to build generative AI applications that support common data security and compliance standards, including GDPR and HIPAA. Customers can use AWS PrivateLink to establish private connectivity between customized Amazon Nova

models and on-premises networks without exposing customer traffic to the internet. Customer data is always encrypted in transit and at rest, and customers can use their own keys to encrypt the data, for example, using AWS Key Management Service (AWS KMS). Customers can use AWS Identity and Access Management (IAM) to securely control access to Amazon Bedrock resources. Also, Amazon Bedrock offers comprehensive monitoring and logging capabilities that can support customer governance and audit requirements. For example, Amazon CloudWatch; can help track usage metrics that are required for audit purposes, and AWS CloudTrail can help monitor API activity and troubleshoot issues as Amazon Nova Multimodal Embeddings is integrated with other AWS systems. Customers can also choose to store the metadata, and embedding output in their vector store of choice and in their own encrypted Amazon Simple Storage Service (Amazon S3) bucket. For more information, see Amazon Bedrock Security.

Intellectual Property

AWS offers uncapped intellectual property (IP) indemnity coverage for outputs of generally available Amazon Nova Multimodal Embeddings models (see Section 50.10 of the AWS Service Terms). This means that customers are protected from third-party claims alleging IP infringement or misappropriation (including copyright claims) by the outputs generated by these Amazon Nova models. In addition, our standard IP indemnity for use of the Services protects customers from third-party claims alleging IP infringement (including copyright claims) by the Services (including Amazon Nova models) and the data used to train them.

Transparency

Amazon Nova Multimodal Embeddings provides information to customers in the following locations: this Service Card, AWS documentation, AWS educational channels (for example, blogs, developer classes), and the AWS Console. We accept feedback through customer support mechanisms such as account managers. Where appropriate for their use case, customers who incorporate Amazon Nova Multimodal Embeddings in their workflow should consider disclosing their use of ML to end users and other individuals impacted by the application, and customers should give their end users the ability to provide feedback to improve workflows. In their documentation, customers can also reference this Service Card.

Governance

We have rigorous methodologies to build our AWS AI services responsibly, including a working backwards product development process that incorporates Responsible AI at the design phase, design consultations, and implementation assessments by dedicated Responsible AI science and data experts, routine testing, reviews with customers, best practice development, dissemination, and training.

Deployment and performance optimization best practices

We encourage customers to build and operate their applications responsibly, as described in AWS Responsible Use of AI Guide. This includes implementing Responsible AI practices to address key dimensions including controllability, safety, fairness, veracity, robustness, explainability, privacy, security, transparency, and governance.

Workflow Design

The performance of any application using Amazon Nova Multimodal Embeddings depends on the design of the customer workflow, including the factors discussed below:

- **Effectiveness Criteria:** Customers should define and enforce criteria for the kinds of use cases they will implement, and, for each use case, further define criteria for the inputs and outputs permitted, and for how humans should employ their own judgment to determine final results. These criteria should systematically address controllability, safety, fairness, and the key dimensions listed above.
- **Configuration:** Amazon Nova Multimodal Embeddings provides configuration options on embedding dimensions, embedding segment size, and the embedding purpose that provide granularity to customize the embedding quality for a specific use-case. For more information, see model parameters in the Amazon Bedrock User Guide.
- **Human Oversight:** If a customer's application workflow involves a high risk or sensitive use case, such as a decision that impacts an individual's rights or access to essential services, human review should be incorporated into the application workflow where appropriate.
- **Performance Drift:** A change in the types of inputs that a customer submits (for example, change in image quality) to Amazon Nova Multimodal Embeddings might lead to different outputs. To address these changes, customers should consider periodically retesting the performance of Amazon Nova Multimodal Embeddings and adjust their workflow if necessary.
- **Model Updates:** When we release new versions of Amazon Nova Multimodal Embeddings, we will notify customers when we release a new version, and will provide customers with time to migrate from an old version to the new one. Customers should retesting the performance of updated version.

Further information

- For service documentation, see Amazon Nova, Amazon Bedrock Documentation, Amazon Bedrock Security and Privacy, Amazon Bedrock Agents, and Amazon Nova User Guide.
- For details on privacy and other legal considerations, see the following AWS policies: Acceptable Use, Responsible AI, Legal, Compliance, and Privacy.
- For help optimizing a workflow, see Generative AI Innovation Center, AWS Customer Support, AWS Professional Services, Ground Truth Plus, and AWS Well-Architected.
- For other tools to help customers work with foundation models, see Amazon Bedrock, Amazon Bedrock Guardrails, Amazon Bedrock Guardrails automated reasoning checks, Amazon Q developer, and Nova Understanding Models.
- If you have any questions or feedback about AWS AI service cards, please complete this form.

Glossary

Controllability: Steering and monitoring AI system behavior.

Privacy & Security: Appropriately obtaining, using and protecting data and models.

Safety: Preventing harmful system output and misuse.

Fairness: Considering impacts on different groups of stakeholders.

Explainability: Understanding and evaluating system outputs.

Veracity & Robustness: Achieving correct system outputs, even with unexpected or adversarial inputs.

Transparency: Enabling stakeholders to make informed choices about their engagement with an AI system.

Governance: Incorporating best practices into the AI supply chain, including providers and deployers.