Amazon Nova Micro, Lite, Pro, and Premier

# AWS AI Service Cards

# AWS AI Service Cards: Amazon Nova Micro, Lite, Pro, and Premier

# Table of Contents

# Amazon Nova Micro, Lite, Pro, and Premier

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 Amazon Nova family of Text and Multimodal Foundation Models that are current as of April 30, 2024.

## Overview

The Amazon Nova family of Text and Multimodal Foundation Models (FMs) consists of 4 models: Amazon Nova Micro, which processes text inputs and generates text completions, and Amazon Nova Lite, Pro and Premier, which are capable of processing text, image, and video inputs while generating text completions. Amazon Nova FMs can understand documents, charts, images, and videos, enabling applications that require multimodal interactions – including complex multi-turn tasks, retrieval-augmented generation (RAG), and agentic workflows.

The "overall effectiveness" of any traditional or generative 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 what best represents their use case, and should therefore be included in an evaluation dataset. Second, different models may respond differently to the same prompt, requiring tuning of the prompt and/or the evaluation mechanism.

Amazon Nova FMs generate completions by understanding inputs and generating relevant completions, and, like more traditional ML solutions, must overcome issues of intrinsic and confounding variation. Intrinsic variation refers to features of the input to which the model should attend, for example, knowing the difference between the prompts "Did the cat win the game?" and "Did the dog win the game?" Confounding variation refers to features friend win the game?" and "Did my cat win the game?" are asking about a cat, and not a dog. The full set of variations encountered by an FM includes language (human and machine), slang, professional jargon, dialects,

expressive non-standard spelling and punctuation and many kinds of errors in prompts, for example, with spelling, grammar, punctuation, logic, and semantics.

Amazon Nova family of models offer customers multiple price performance operating points to best optimize between accuracy, speed, and cost. Amazon Nova Micro is our text only model that delivers the lowest latency responses at the lowest cost per inference among Nova family; Amazon Nova Lite is our low cost multimodal model that is fast for processing image, video and text inputs. While Amazon Nova Pro is a highly capable multimodal model achieving an optimal combination of accuracy, speed, and cost for a wide range of tasks, Amazon Nova Premier is our most capable multimodal model for complex tasks and best teacher model for distilling custom models.

## Intended use cases and limitations

Amazon Nova FMs serve a wide range of potential application domains, while offering the following key differentiators:

- **Reasoning across a wide variety of inputs including text, documents, charts, images, and videos:** These models offer advanced understanding capabilities on Amazon Bedrock, enabling deeper insights from multimedia content.
- **Agentic workflows for enabling applications:** They facilitate applications that require RAG, API execution, and UI actuation, such as predicting API actions to automate client applications.
- **Functional expertise in key enterprise verticals:** The models are optimized for certain domains such as software development, retail, entertainment, and financial analysis, delivering high performance in domain-specific accuracy benchmarks. For more information, see Amazon Nova User Guide.
- **Customizability**: Customers can fine-tune Micro, Lite, and Pro with their own data through Custom Fine-Tuning (CFT), providing the flexibility to achieve desired accuracy, latency, and cost. Additionally, Amazon Bedrock's Model Distillation feature enables powerful knowledge transfer across the Amazon Nova family. Amazon Nova Premier serves as a teacher model for Amazon Nova Micro, Lite, and Pro, while Amazon Nova Pro acts as a teacher model for Micro and Lite allowing customers to create more price-performant variants of Micro, Lite, and Pro.

Other traditional capabilities of the Amazon Nova FMs include, but are not limited to, the following:

- Text generation, such as expanding on the information provided in the prompt (for example, "Write a blog post about cats."), summarizing information provided in the prompt (for example,

"What is the main point of the foregoing paragraph?"), classifying prompt text (for example, "Is the sentiment of the foregoing text positive or negative?"), and answering questions about text in the prompt (for example, "Does the cat in the foregoing text have four legs?").

- Dialog mode refers to the FM being able to conduct a conversation with a user by including recent <prompt, completion> pairs (up to the limits of the context size). With Amazon Nova, users can invoke dialog mode by starting their prompt with "User:" and ending their prompt with "Bot:".
- In-context learning is the ability to learn from examples of the desired task in the prompt. Common examples include one-shot and few-shot learning, and the closely related technique of chain of thought (CoT) prompting. One-shot learning refers to putting one example <prompt, completion> pair in the prompt itself. Few-shot learning refers to putting a small number of example <prompt, completion> pairs in the prompt. CoT refers to having the prompt ask for the completion to take the form of a sequence of steps.
- Retrieval augmented generation (RAG) refers to retrieving data from external knowledge stores and conditioning generation on the retrieved data.

For example, Amazon Nova Micro can be used to generate a product description. The user's goal may be to generate product descriptions with uniform writing style for an online English-language product catalog. The stakeholders may include the readers of the catalog, who want concise, accurate, easy-to-read, and unembellished product descriptions; the product providers, who provide detailed product information and want their products represented fairly and accurately; and the catalog operator, who wants customers to have a uniformly safe, reliable, and excellent experience across the catalog. To minimize refunds and exchanges, all parties prioritize accuracy. The workflow is: 1/ providers submit their product feature lists online to the operator, 2/ the operator uses Amazon Nova to turn feature lists into product descriptions, 3/ the operator publishes descriptions to the catalog, 4/ catalog readers provide feedback to the owner and providers via product purchase, return and review behaviors, and 5/ the owners and providers work together to improve product features and product descriptions. The input prompts contain the product name and the list of product features, and the completion one paragraph of product descriptions. Input variations will include all the normal variations in English expression across different individuals and more, including differences in the degree of feature embellishment, inaccuracies, misspellings, and undefined abbreviations. The error types, ranked in order of estimated negative impact on readers, include a/ incorrect facts, b/ language that is toxic or otherwise inappropriate for a diverse customer base, c/ key facts omitted (this is mitigated by showing detailed product descriptions below the summary), and d/ poor quality writing. With this

in mind, we would expect the catalog operator (the customer) to test an example prompt in the AWS Console and review the completion.

**Prompt**

> A trash bag product has the following features: Brand=Amazon Basics; Capacity=13 Gallons; Unit Count=120; Recommended Uses For Product=Food Waste; Material Feature=Scent Free; Color=White; Material=Plastic; Item Form=Bag; Scent=Unscented; Number of Items=1;Other=Includes 120 tall kitchen garbage bags; Other=Individual bag dimensions: 24 inches wide x 27 3/8 inches tall with 0.9 mil bag thickness; Other=Each trash bag has a 13-gallon capacity; Other=White trash bag with red drawstring closure for easy tying and carrying. Create a short description for a product catalog.

**Completion**

> The Amazon Basics 13-gallon trash bags are a great option for food waste disposal. These unscented, white plastic bags come in a pack of 120 and measure 24 inches wide by 27 3/8 inches tall with a 0.9 mil thickness. They feature a red drawstring closure for easy tying and carrying.

Assessing the completion for effectiveness, we observe a/ no contradictions of the facts in the prompt, b/ no toxic or unsafe language, c/ all key product information present, and d/ coherent and organized writing. After continued experimentation in the Console, the customer should finalize their own measure of effectiveness based on the impact of errors, run a scaled-up test via the Console or API, and use the results of human judgements (with multiple judgements per test prompt) to establish a benchmark effectiveness score. Amazon Bedrock directly offers these kinds of testing capabilities.

Amazon Nova Text and Multimodal models have limitations that require careful consideration when using the models:

**Appropriateness for Use**

> We make every effort to design, develop, and rigorously test our models to help ensure they produce appropriate completions based on user inputs, but generative models are by their nature non-deterministic and may occasionally produce unintended or undesirable completions. We encourage users to provide feedback here about our models to help us continuously improve their performance. Customers should evaluate completions for accuracy and appropriateness for their use case, especially if they will be directly surfaced to end

users. Additionally, if Amazon Nova models are 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. See the AWS Responsible AI Policy for more information. Customers who use Amazon Nova models are responsible for ensuring that their use of Amazon Nova models and the generated completion complies with all applicable laws. The model and completion may not be used for any prohibited practices under the EU AI Act.

**Unsupported Tasks**

Amazon Nova models are not designed to provide opinions or advice, including medical, legal or financial advice. For example, when prompted with: "What is the speed limit in San Mateo, California?" Amazon Nova models may complete with: "The speed limit in San Mateo, California, is 25 miles per hour." The answer is not correct, as speed limits vary by street type. It also cannot answer questions about its own design or development.

**Context Size**

The context size is the maximum amount of text (measured in tokens) in a <prompt, completion> pair. For Amazon Nova models, a token is approximately six characters (English words average about five characters). The context size constrains the use cases that can be supported. For example, it defines the length of chat history available to a chatbot, the length of text that can be summarized, the amount of background information that can be provided in RAG, and the number of training examples that can be provided when using one-shot or few-shot in-context learning. For more information, see Amazon Nova User Guide.

**Information Retrieval**

By themselves, Amazon Nova models are not information retrieval tools. The models store information about the probability that one token will follow some prior sequence of tokens, not the exact sequence of tokens that might be found in a document database. Customers should consider whether or not using RAG will improve the effectiveness of their solution, and carefully assess the veracity of Amazon Nova model completions in their use case.

**Languages**

Amazon Nova models are released as generally available for English, German, French, Spanish, Italian, Portuguese, Japanese, Hindi, and Arabic. Amazon Nova model's guardrails are intended for use in the aforementioned languages only, however Amazon Nova models are trained and will attempt to provide completions in up to 200 languages. In use cases beyond the languages identified above, customers should carefully check completions for effectiveness, including safety.

### Coverage

For any language, Amazon Nova model training corpus does not cover all dialects, cultures, geographies and time periods, or the domain specific knowledge a customer may need for a particular use case, and we do not define a "cutoff date" for training or otherwise try to characterize an FM as a knowledge base. Customers with workflows requiring accurate information from a specific knowledge domain or time period should consider employing RAG and/or orchestration.

### Programming Languages

Amazon Nova models can generate code, including Python, Java, JavaScript, C#, TypeScript and SQL. However, they do not have the advanced code completion features present in Amazon Q, such as automatic security scanning. Customers using Amazon Nova models to generate code should check the validity and safety of code completions.

### Human Interactions

Amazon Nova models offer a new form of human-computer interaction. Although interacting with an Amazon Nova model in a chatbot setting can feel natural, Amazon Nova models lack many human capabilities, and the science around optimizing model to human interactions is still emerging. For example, completions may be fluently written with a degree of confidence that is unwarranted by Amazon Nova model's actual "knowledge," potentially misleading a reader. Critically, completions can vary depending on changes, sometimes small, to the wording of prompts, or even the ordering of examples within prompts. For information about the best way to structure interactions with Amazon Nova models, see Amazon Nova User Guide. Customers should consider carefully who will use Amazon Nova completions, and what context and supporting information those users will need to properly evaluate and utilize the completions.

## Design of Amazon Nova models

### Machine Learning

Amazon Nova FMs perform token inference using transformer -based generative machine learning. Amazon Nova models understand the input prompts and generate completions using a probability distribution learned through a combination of unsupervised and supervised machine learning techniques. Our runtime service architecture works as follows: 1/ the model receives a user prompt via the API or Console; 2/ the model filters the prompt to comply

with safety, fairness and other design goals; 3/ the model may augment the filtered prompt to support user-requested features, for example, knowledge-base retrieval; 4/ the model generates a completion; 5/ the model filters the completion for safety and other concerns; 6/ the model returns the final completion.

**Controllability**

We say that an Amazon Nova model exhibits a particular "behavior" when it generates the same kind of completions for the same kinds of prompts with a given configuration (for example, temperature). For a given model architecture, the control levers that we have over the behaviors are primarily a/ the training data corpus, and b/ the filters we apply to pre-process prompts and post-process completions. Our development process exercises these control levers as follows: 1/ we pre-train the FM using curated data from a variety of sources, including licensed and proprietary data, open source datasets, and publicly available data where appropriate; 2/ we adjust model weights via supervised fine tuning (SFT) and reinforcement learning with human feedback (RLHF) to increase the alignment between the Amazon Nova model and our design goals; and 3/ we tune safety filters (such as privacy-protecting and profanity-blocking filters) to block or evade potentially harmful prompts and responses to further increase alignment with our design goals.

**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 each, a user prompts an Amazon Nova model to generate an email summarizing key points (conclusions and action items) from a video conference from notes taken during the conference. With Application A, the meeting host first seeks permission from participants to make and transcribe an audio recording of the meeting, and then, post-meeting, triggers the app to transcribe the meeting and send an Amazon Nova model-generated summary of the transcript to all participants. Application A must cope with multiple issues, including transcription errors, variations in grammar and vocabulary across participants, input content that does not relate to key points, key points that are partially or completely implicit, and potential toxic input (perhaps within a key point). With Application B, participants type meeting notes into a web app, and the meeting host uses an Amazon Nova model to generate the key point email. Application B must cope with typographical errors, conflicts between the key points reported by different participants, individual adjustments to action items for clarity or other reasons, and differences in grammar and writing style between participants. Because A and B have different inputs, they will likely have different completions (i.e., hallucination and omission) and toxicity, even assuming that each application

is deployed perfectly. Because performance results depend on a variety of factors including the Amazon Nova model, the customer workflow, and the evaluation dataset, we recommend that customers test the models using their own content. Amazon Bedrock and Amazon SageMaker AI Clarify directly provide automated and human testing capabilities.

**Test-driven Methodology**

We use multiple datasets and human workforces to evaluate the performance of the Amazon Nova models. This is because evaluation datasets vary based on use case, intrinsic and confounding variation, the types and quality of labels available, and other factors. Our development testing involves automated benchmarking against publicly available datasets, automated benchmarking against proprietary datasets, benchmarking against proxies for anticipated customer use cases, human evaluation of completions against proprietary datasets, automated red teaming, manual red teaming, and more. Our development process examines Amazon Nova model performance using all of these tests, takes steps to improve the model and/or the suite of evaluation datasets, and then iterates. In this service card, we provide an overview of our methodology.

- Automated Benchmarks: Benchmarking provides apples-to-apples comparisons between candidate models by substituting an automated "assessor" mechanism for human judgement, which can vary. We conducted comprehensive evaluations to assess the Amazon Nova models using multiple proprietary datasets. We also include external benchmarks such as WILDCHAT non-toxic and StrongReject in our testing.
- Human Evaluation: Human evaluation is a critical step in evaluating the model's completions. Using human judgement is critical for assessing the effectiveness of an Amazon Nova FM on more challenging tasks, because only people can fully understand the context, intent and nuances of more complex prompts and completions. Given this, we have developed proprietary evaluation datasets of challenging prompts that we use to assess development progress for Amazon Nova models. To assess a model, we retrieve the completion for each prompt, and then ask multiple individuals to assess the quality of each pair along a number of different factors, for example, quality, verbosity, formatting.
- Independent Red Teaming Network: Consistent with our Frontier AI Safety Commitments on ensuring Safe, Secure, and Trustworthy AI, we use a variety of third parties to conduct red teaming against our AI models. We leverage red teaming firms to complement our in-house testing in areas such as safety, security, privacy, fairness and veracity related topics. We also work with specialized firms and academics to red team our models for specialized areas such as Cybersecurity and Chemical, Biological, Radiological and Nuclear (CBRN) capabilities.

### Safety

Safety is a shared responsibility between AWS and our customers. Our goal for safety is to mitigate key risks of concern to our customers, and to society more broadly. Amazon 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. Customers are responsible for end-to-end testing of their applications on datasets representative of their use cases, and deciding if test results meet their specific expectations of safety, fairness, and other properties, as well as overall effectiveness.

- Harmlessness: Over-optimizing an LLM to be harmless can lead to a less helpful LLM. Therefore, we evaluate Amazon Nova FMs for harmlessness on both how often it generates harmful responses and how often it treats harmless prompts as harmful. For example, we use a proprietary dataset of harmless prompts and adversarial red teaming prompts that attempt to solicit completions containing violence, sexual content, insults, identity attacks, stereotypes, malicious intent, and other harmful content. For example, on a proprietary dataset (2.4k samples) containing prompts that attempt to solicit model response containing harmful content (for example, self-harm, violence, animal abuse), on average, Amazon Nova FMs correctly produce safe responses to over 90% of these harmful prompts.
- Abuse Detection: To help prevent potential misuse, Amazon Bedrock implements automated abuse detection mechanisms. These mechanisms are fully automated, so there is no human review of, or access to, user inputs or model completions. To learn more, see Amazon Bedrock Abuse Detection in the *Amazon Bedrock User Guide*.
- Child Sexual Abuse Material (CSAM): At Amazon, we are committed to producing generative AI services that keep child safety at the forefront of development, deployment, and operation. We utilize Amazon Bedrock's Abuse Detection solution (mentioned above), which uses hash matching or classifiers to detect potential CSAM. If Amazon Bedrock detects apparent CSAM in 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 seriously and will continue to update our detection, blocking, and reporting mechanisms.
- Frontier AI Safety Commitments: Consistent with our Frontier AI Safety Commitments, we comprehensively evaluated Amazon Nova Premier — both its current public version and earlier iterations — across three high-risk domains: Chemical, Biological, Radiological, and Nuclear (CBRN), Offensive Cyber Operations, and Automated AI R&D. Using benchmarks and expert assessments, we found that Amazon Nova Premier is within the tolerance threshold for critical capabilities in each domain.

- CBRN: Evaluations included Weapons of Mass Destruction Proxy (WMDP), ProtocolQA, and BioLP Bench, along with external expert red teaming by Carnegie Mellon's Gomes Group and Nemesys Insights.
- Offensive Cyber Operations: We used benchmarks such as SECURE, CTIBench, CyberMetric (which test the model's knowledge of cyber threat intelligence and vulnerabilities), and Cybench (which tests practical exploitation through cyberattack challenges), supported by expert red-teaming by internal experts.
- Automated AI R&D: We relied on internal simulations and public benchmarks including RE-Bench and PaperBench, with independent review by METR, a nonprofit research organization specializing in AI evaluations.

**Fairness**

Amazon Nova models are designed to generate completions that a diverse set of customers will find effective across a wide range of categories and avoid generating content related to stereotypes or making generalizations about specific groups of people, roles, or behavior. The models are also designed to work well for use cases across our diverse set of customers. To achieve this, we examine the extent to which Amazon Nova model completions can be considered biased against particular demographic groups, and look for ways to discourage prompting the models with material that could elicit such behavior. Consistent with our approach to safety, we steer the models towards being helpful while trying not to make assumptions about membership in specific demographic groups. We use BOLD, a dataset to evaluate fairness in open-ended language generation, to assess Nova FMs, observing 99.9% or more fair completions across Micro, Lite, Pro, and Premier. As this reached saturation at 99.9%, we evaluate Amazon Nova FMs propensity to reject generating biased content on proprietary datasets spanning multiple input modalities. For example, on a proprietary dataset (1.2k samples) containing prompts that attempt to solicit biased responses (for example, stereotypes that contain bias against a group, etc.) Amazon Nova Premier correctly produces safe responses to 95.3% these prompts.

**Explainability**

Customers wanting to understand the steps taken by Amazon Nova models to arrive at the conclusion expressed in a completion can use chain of thought (CoT) techniques described here . For customers wanting to see attribution of information in a completion, we recommend using RAG with Amazon Bedrock Knowledge Bases .

### Veracity

Because transformer-based FMs are token generation engines, and not information retrieval engines, their completions may contain statements that contradict statements in the prompt or that contradict facts verifiable from trusted third-party sources, or the completions may omit statements that customers expect should be made, given information in the prompt or even just "common sense." Customers should carefully consider whether or not using RAG will improve the effectiveness of their solution; use of RAG can still result in errors. We assess Nova models' general knowledge without RAG on multiple datasets, and find that the models perform well, given the intrinsic limitations of large language models technology.

### Robustness

We maximize robustness with a number of techniques, including using large training datasets that capture many kinds of variation across many different semantic intents. We measure model robustness by applying small, semantics-preserving perturbations to each and compare the responses to see how stable or invariant they are. We compute a robustness score as the worst-case performance across all perturbations of each prompt, namely, the model is correct on a specific base prompt if and only if it predicts correctly on all perturbations of it.

### Privacy

Amazon Nova models are available in Amazon Bedrock. Amazon Bedrock is a managed service and does not store or review customer prompts or customer prompt completions, and prompts and completions are never shared between customers, or with Amazon Bedrock partners. AWS does not use inputs or completions generated through the Amazon Bedrock service to train Amazon Bedrock models, including Amazon Nova models. 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.

- PII: Amazon Nova models are designed to avoid completing prompts that could be construed as requesting private information. If a user is concerned that their private information has been included in an Amazon Nova model completions, the user should contact us here .

### Security

All Amazon Bedrock models, including Amazon Nova models, 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 Titan 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 models is integrated with other AWS systems. Customers can also choose to store the metadata, prompts, and completions in their own encrypted Amazon Simple Storage Service (Amazon S3) bucket. For more information, see Amazon Bedrock Security.

**Intellectual Property**

Amazon Nova models are designed for generation of new creative content. AWS offers uncapped intellectual property (IP) indemnity coverage for completions of generally available Amazon Nova 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 completions 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 models provide 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 via the AWS Console and through traditional customer support mechanisms such as account managers. Where appropriate for their use case, customers who incorporate Nova models 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. The performance of any application using an Amazon Nova model depends on the design of the customer workflow, including the factors discussed below:

1. **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 completions permitted, and for how humans should employ their own judgment to determine final results. These criteria should systematically address controllability, safety, fairness, and the other key dimensions listed above.
2. **Model Choice:** The direct cost of using an Amazon Nova model is a function primarily of model size, the average number of input tokens and the average number of completion tokens. In general, customers should consider using the smallest model that yields acceptable effectiveness for their use case(s).
3. **Configuration:** Amazon Nova models provide four configuration parameters: temperature, top p, response length and stop sequences. Temperature is a number in the range [0,1] that controls the creativity of the response. A temperature of 0 means the same prompt will generate completions with minimal variability (useful for reproducibility and debugging) while a temperature of 1 means the same prompt can generate differing and unlikely completions (useful for creativity). Top p is a number in the range [0.1,1] used to remove less probable tokens from the option pool, i.e., given a list of possible tokens in order of most probable to least probable, top p limits the length of the list to include just those tokens whose probabilities sum to at most top p. If top p is 1, the model considers all options. The closer top p gets to zero, the more the model focuses on the more probable options. Response length specifies the maximum number of tokens in the generated response. Stop sequences specifies character sequences that, if generated, halt further generation. Customers should consider which parameter choices will provide the most effective result. More detail is here.
4. **Prompt Engineering:** The effectiveness of Amazon Nova model completions depends on the design of the prompts (called prompt engineering). We provide guidance on prompt engineering here. Customers should consider using prompt templates to encode their lessons about the prompt designs that are most successful for their use cases.
5. **Knowledge Retrieval:** Customers should carefully consider the kinds of information they wish to see in Amazon Nova model completions. If customers need completions to contain domain-

specific, proprietary and/or up-to-date knowledge (for example, a customer support chatbot for online banking), they should consider using retrieval augmented generation RAG. Customers can enable a RAG workflow by using Amazon Bedrock Knowledge Bases to build contextual applications.

6. **Orchestration (Agents):** For use cases that require systematic coordination and management of various components and processes that interact with an FM (for example, making travel reservations), customers should consider using Amazon Nova models with Amazon Bedrock Agents ("Tools"). Amazon Bedrock Agents can be used to setup interactions between Amazon Nova models, additional data sources (knowledge bases), other software tools or AI services, and user conversations, handling API calls automatically without the need for writing custom code. More detail is here.
7. **Base Model Customization:** Customization can make a base FM more effective for a specific use case, particularly for more compact models that offer lower cost. Customers can fine-tune Amazon Nova FMs on their own labeled data. Because changing the base model to focus on a specific use case can impact safety, fairness and other properties of the new model (including performance on tasks on which the base model performed well), we use a robust adaptation method that minimizes changes to the safety, fairness and other protections that we have built into our base models, and minimizes impact on model performance on the tasks for which the model was not customized. After any customization, customers should test their model according to their own responsible AI policies. More detail on Amazon Nova model customization guidelines is here.
8. **Filter Customization:** Customers have multiple options for aligning Amazon Nova model behaviors with their own effectiveness criteria: preprocessing prompts with their own filters, using built-in Amazon Nova model protections, using Amazon Bedrock Guardrails, and post-processing completions with their own filters. These options can be used singly or in combination.
9. **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.
10. **Performance Drift:** A change in the types of prompts that a customer submits (for example language switching, spelling errors) to Amazon Nova FMs might lead to different completions. To address these changes, customers should consider periodically retesting the performance of Amazon Nova models and adjust their workflow if necessary.
11. **Updates:** We will notify customers when we release a new version, and will provide customers time to migrate from an old version to the new one. Customers should consider retesting the

performance of the new Nova model version on their use cases when changing to the updated model.

## 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 workflows, see Generative AI Innovation Center, AWS Customer Support, AWS Professional Services, Ground Truth Plus, and Amazon Augmented AI.
- 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.