Amazon Nova 2 Lite

AWS AI Service Cards

AWS AI Service Cards: Amazon Nova 2 Lite

Table of Contents

Amazon Nova 2 Lite

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 release of Amazon Nova 2 Lite that is current as of December 2, 2025.

Overview of Amazon Nova 2 Lite

Amazon Nova 2 Lite is a proprietary Foundation Model (FM) designed for enterprise use cases. It processes text, image, and video inputs (together, a "prompt") while generating text output (a "completion"). Amazon Nova 2 Lite 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. This AI Service Card applies to the use of Amazon Nova 2 Lite via Amazon Bedrock. Amazon Nova 2 Lite 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 the content in the service card is applicable to the use of customized Nova 2 Lite models via SageMaker AI, which allows you to adapt Nova models to your specific business needs, or models hosted on https://nova.amazon.com, a publicly available website where individuals may try certain Nova models.

Amazon Nova 2 Lite is said to be "effective" if a skilled human evaluator decides that the generated text has: 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. Otherwise, the model is considered to be "ineffective". After continued experimentation in Amazon Bedrock 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.

The "overall effectiveness" of any foundation 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 prompts and responses will best represent their use case and should therefore be included in an evaluation dataset. Second, different large language models may respond differently to the same prompt, requiring tuning of the prompt and/or the evaluation mechanism.

As with all ML solutions, Amazon Nova 2 Lite is designed to understand relevant differences in input (such as distinguishing between similar queries) while ignoring irrelevant variations (such as minor differences in phrasing). Relevant differences refer 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?' Irrelevant variations refer to features of the input that the model should ignore, for example, understanding that the text prompts 'a jumping cat' and 'the jumping cat' should return the same result, since there should be no semantic difference between 'a' and 'the'. 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. Since different Amazon Nova 2 Lite prompts will output different results, customers should experiment as necessary to understand how best to adjust prompts to achieve an effective result.

Intended use cases and limitations

Amazon Nova 2 Lite serves a wide range of potential application domains and offers the following core capabilities:

- **Long Context:** Amazon Nova 2 Lite supports a 1M token context window
- **Reasoning across a wide variety of inputs including text, documents, charts, images, and videos:** Amazon Nova 2 Lite offers advanced understanding capabilities, enabling deeper insights from multimedia content
- **Reasoning Controls:** Amazon Nova 2 Lite is a hybrid reasoning model, so you can turn reasoning on/off depending on the task
- **Reasoning Budget:** Amazon Nova 2 Lite supports a reasoning budget which allows you to configure the amount of reasoning needed for tasks by changing the setting to "low", "medium", or "high"

- **Multimodal Reasoning:** Amazon Nova 2 Lite will integrate multimodal inputs into its reasoning process, improving performance on tasks requiring understanding of charts, diagrams, or code screenshots
- **Improved Instruction Following:** Amazon Nova 2 Lite improves in its inherent instruction following capabilities in comparison to previous generations
- **Sequential Task Decomposition and Function Calling:** Amazon Nova 2 Lite will automatically decompose complex problems into manageable sub-tasks
- **System Tools:** Amazon Nova 2 Lite provides two different system tools that the model can use when responding to a user's query
  - Amazon Nova Grounding allows the model to access the web and search for live information
  - The Code Interpreter allows the model to perform calculations using a python interpreter
- **Agentic workflows for enabling applications:** Amazon Nova 2 Lite facilitates applications that require RAG, API execution, and User Interface (UI) actuation, such as predicting API actions to automate client applications
- **Customizability:** Amazon Nova 2 Lite supports Supervised Fine-Tuning (SFT) and Reinforcement Fine-Tuning (RFT) which allows you to adapt the model to your business-specific needs, see the Amazon Nova User Guide.

The components differ in the parameters (for example, the reasoning budget) required to invoke them. For more information about these specifications, see the Amazon Nova User Guide.

When assessing a FM 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 2 Lite as a creative tool to help a category manager generate product descriptions. The **business goal** is to generate product descriptions with uniform writing style for an online English-language product catalog. The **stakeholders** include the readers of the catalog who wants to access concise, accurate, easy-to-read, and unembellished product; the product providers, who want to 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. The **workflow** is: 1/ product providers submit their product feature lists online to the operator, 2/ the operator uses Amazon Nova 2 Lite 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. **Input prompts** contain information regarding the product name and the list of product features, and completion is a one paragraph product description. Input **variations** 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. Key potential **error types** for consideration include: a/ factual inaccuracies, b/ inappropriate or harmful language, c/ omitted information, and d/ suboptimal writing quality. The relative importance of each error type will vary by use case and should be evaluated by customers based on their specific application. With this in mind, we would expect the catalog operator (the customer) to test an example prompt in the Amazon Bedrock 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 Amazon Bedrock Console, the customer should finalize their own measure of effectiveness based on the impact of errors, run a scaled-up test via the Amazon Bedrock Console or API and use the results of human judgements (with multiple judgements per test prompt) to establish a benchmark effectiveness score.

Appropriateness for Use

Because its output is probabilistic, Amazon Nova 2 Lite may produce 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 2 Lite 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 Responsible AI at AWS Policy.

Unsupported Tasks

Amazon Nova 2 Lite is not designed to provide opinions or specialized advice in domains such as medicine, law, or finance. Customers should not rely on the model for professional guidance in these areas.

Amazon Nova 2 Lite is not intended to support any prohibited practices under the EU AI Act or any other relevant law. Amazon Nova 2 Lite can be integrated into an array of systems such as conversational assistants, content moderation systems, agentic AI systems, predictive systems, cybersecurity, and human-AI collaboration systems. For more technical information about how Amazon Nova 2 Lite may be integrated into AI systems, see the Amazon Nova User Guide. All Amazon Nova 2 Lite use cases must comply with the AWS Acceptable Use Policy.

Amazon Nova 2 Lite has limitations requiring careful consideration.

Inputs and Outputs

Amazon Nova 2 Lite supports a model input size of 1 million tokens which enables customers to analyze large codebases, documents up to 400 pages, or 90-minute-long videos in a single prompt. For more information about the maximum size of the input and output modalities, see the Amazon Nova User Guide.

Extended Thinking

Amazon Nova 2 Lite introduces extended thinking capabilities that enable the model to engage in deeper reasoning for complex problems with three effort levels: low, medium, and high. Extended thinking is designed for tasks requiring deep, systematic analysis including multi-step mathematical proofs, complex debugging, system architecture design, document synthesis across sources, and agentic workflows requiring orchestration of multiple tool calls. Extended thinking should remain disabled for straightforward tasks including simple factual questions, basic calculations, and speed-critical applications where latency matters. When using extended thinking

for agentic workflows consider using a planning framework SDK such as Strands Agents to make the planning and execution process of your agent systems more robust.

Reasoning Visibility

When extended thinking is enabled in Amazon Nova 2 Lite, the reasoning content displays as [REDACTED] rather than showing the model's step-by-step thinking process, though reasoning capabilities are actively contributing to improved output quality. Customers are billed for reasoning tokens as output tokens because this computational work directly improves the quality and accuracy of responses. While the step-by-step reasoning process is not displayed, customers benefit from enhanced outputs generated through this deeper analysis. Customers can confirm extended thinking is working by checking for reasoningContent blocks in the response and observing improved quality on complex tasks. For customers requiring transparency into model reasoning for debugging or auditing purposes, this limitation should be carefully considered when designing workflows.

Web Grounding

Amazon Nova 2 Lite provides access to Web Grounding to retrieve and incorporate publicly available information with citations as context for responses. Customers are responsible for retaining and displaying citations and links to source material in output provided to end users. Model-generated queries stay within AWS infrastructure and are never sent to the broader internet, and runtime filtering protects against indirect prompt injection and misinformation. The expansive internal web search index and knowledge graphs prioritize trustworthy and high-quality sources and filter malicious content on ingress.

Supported Natural Languages

Amazon Nova 2 Lite supports over 200 languages, with optimization for 15 languages: English, German, Spanish, French, Italian, Japanese, Korean, Arabic, Simplified Chinese, Russian, Hindi, Portuguese, Dutch, Turkish, and Hebrew. Amazon Nova model's guardrails are intended for use in optimized languages only. In use cases beyond the languages identified above, customers should carefully check completions for performance and reliability.

Coverage and Knowledge Cutoff Date

For any language, the 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 try to characterize a FM as a knowledge base. Customers with

workflows requiring accurate information from a specific knowledge domain or time period should consider employing web grounding, RAG, or tool use orchestration.

Human Interactions

Amazon Nova 2 Lite offers a new form of human-computer interaction. Although interacting with Amazon Nova 2 Lite in a chatbot setting can feel natural, Amazon Nova 2 Lite 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 2 Lite 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 2 Lite completions, and what context and supporting information those users will need to properly evaluate and utilize the completions.

Design of Amazon Nova 2 Lite

Machine Learning

Amazon Nova 2 Lite performs token inference using transformer-based generative machine learning. The model understands the input prompts and generates 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, security, and other design goals; 3/ the model may augment the filtered prompt to support user-requested features, for example, grounding to access the web and search for live information; 4/ the model generates a completion; 5/ the model filters the completion for safety and other concerns; 6/ the model returns the final completion. Amazon Nova 2 Lite is available pursuant to the AWS Customer Agreement or other relevant agreements with AWS.

Controllability

We say that Amazon Nova 2 Lite 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 2 Lite 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 Amazon Nova 2 Lite 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 2 Lite 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 Amazon Nova 2 Lite 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 – including potential inaccuracies, omissions and toxicity – 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 Amazon Nova 2 Lite, the customer workflow, and the evaluation dataset, we recommend that customers test the model using their own content. Amazon Bedrock and SageMaker AI Clarify directly provide automated and human testing capabilities.

Test Driven Methodology

We use multiple datasets and human teams to evaluate the performance of Amazon Nova 2 Lite. 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 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 2 Lite 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 Amazon Nova 2 Lite using multiple proprietary datasets. We also include external benchmarks such as WILDCHAT non-toxic and WMDT 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 Amazon Nova 2 Lite 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 2 Lite. 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 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 Nova 2 Lite is designed to prevent the generation of harmful content, including content that may cause physical or emotional harm, and content that may harass, harm, or encourage harm to individuals or specific groups, especially children. Amazon is committed to producing generative AI services that keep child safety at the forefront of development, deployment, and operation. We conduct testing and implement mitigations to prevent Amazon Nova 2 Lite from generating inappropriate content related to children. Amazon

Nova 2 Lite is designed to block problematic inputs and completions. In a case where a customer asks Amazon Nova 2 Lite to generate a completion and no safety filters are triggered, it returns the completion. In the case where the model cannot complete a prompt due to the possibility of generating a harmful output, it will display a message stating it cannot generate a response to that prompt. For example, the system will not provide responses to prompts such as 'tell me how to make a bomb', 'how do I produce meth', or 'how do I harm animals'.

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. Customers who customize Amazon Nova 2 Lite are responsible for testing their customized models and implementing appropriate guardrails to ensure the models continue to meet their safety and responsible AI requirements.

Harmlessness

We evaluate the capability of Amazon Nova 2 Lite to not respond with potentially harmful content. We test Nova 2 Lite on several proprietary datasets, covering different modalities and languages, of harmless prompts and adversarial red teaming prompts that attempt to elicit responses that contain violence, sexual content, insults, identity attacks, stereotypes, malicious intent, and other harmful content. In one of the proprietary datasets containing 6.4k prompts designed to elicit harmful responses, such as those related to self-harm and animal abuse, Amazon Nova 2 Lite correctly provides safe responses to over 98% of these prompts.

Toxicity is a common, but narrow form of harmfulness, on which individual opinion varies widely. We assess Amazon Nova 2 Lite's ability to avoid responding with content that contains potentially toxic content through automated testing on multiple datasets, and find that it performs well on common toxicity types. For example, on a proprietary toxic prompts dataset containing 8.5K prompts which we classified into sub-categories (for example, violence and gore, insults and stereotype, hate symbols, sexual content), Amazon Nova 2 Lite's end-to-end guardrails provide safe responses to over 95% of the 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 outputs. To learn more, see Amazon Bedrock Abuse Detection in the Amazon Bedrock User Guide.

Child Sexual Abuse Material (CSAM)

Amazon Nova 2 Lite utilizes 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 Amazon Nova 2 Lite 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)

Compared to information available via internet searches, science articles, and paid experts, we see no indications that Amazon Nova 2 Lite 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

Amazon Nova 2 Lite is 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 behaviors. The model is 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 2 Lite 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 evaluate Amazon Nova 2 Lite's propensity to reject generating biased content on proprietary datasets spanning multiple input modalities. For example, on a proprietary dataset containing 1.4k prompts that attempt to solicit biased responses (for example, stereotypes that contain bias against a group, etc.) Amazon Nova 2 Lite produces safe responses to 95.8% these prompts.

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 prompt 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.

Explainability

Customers wanting to understand the steps taken by Amazon Nova 2 Lite to arrive at the conclusion expressed in a completion can use Chain of Thought (CoT) techniques. For customers wanting to see attribution of information in a completion, we recommend using RAG with Amazon Bedrock Knowledge Bases.

Privacy

Amazon Nova 2 Lite is available in Amazon Bedrock. Amazon Bedrock is a managed service and does not store or review customer prompts or completions, and prompts and completions 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 2 Lite. 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 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 completion, the user should contact us here.

Security

All Amazon Bedrock models, including Amazon Nova 2 Lite, 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 2 Lite 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 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, CloudWatch can help track usage metrics that are required for audit purposes, and CloudTrail can help monitor API activity and troubleshoot issues as Amazon Nova 2 Lite is integrated with other AWS systems. Customers can also choose to store the metadata, prompts, and completions in their own encrypted Amazon S3 bucket. For more information, see Amazon Bedrock Security.

Intellectual Property

Amazon Nova 2 Lite is designed for generation of new creative content. We use guardrails to prevent customers from using our services to violate the rights of others. AWS offers uncapped intellectual property (IP) indemnity coverage for outputs 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 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. If you are a rightsholder/authorized representative and have a complaint regarding our commitments under the Copyright Chapter of the Code of Practice for General-Purpose AI Models under the EU AI Act, you may contact us at gpai-models@amazon.com. Please be sure to include enough detail for us to investigate your complaint.

Transparency

Amazon Nova 2 Lite provides information to customers in the following locations: this Service Card, the Amazon Nova Technical Report, 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 2 Lite 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 2 Lite 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 2 Lite provides multiple configuration parameters including temperature, top p, response length, stop sequences, the reasoning budget, web grounding, and code interpreter. 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. Additionally, customers can configure reasoning with the budget set to "low", "medium", or "high" to control reasoning depth. Customers can also enable system tools including Web Grounding for web search and Code Interpreter for calculations. Customers should consider which parameter choices will provide the most effective result for their use case. See the Amazon Nova User Guide for additional details.

- **Prompt Engineering:** The effectiveness of Amazon Nova 2 Lite completions depends on the design of the prompts (called prompt engineering). We provide guidance on prompt engineering in the Amazon Nova User Guide. Customers should consider using prompt templates to encode their lessons about the prompt designs that are most successful for their use cases.
- **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 details on Amazon Nova model customization guidelines are available in the Amazon Nova 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 prompts that a customer submits (for example language switching, spelling errors) to Amazon Nova 2 Lite might lead to different outputs. To address these changes, customers should consider periodically retesting the performance of Amazon Nova 2 Lite and adjust their workflow if necessary.
- **AI Agents (Orchestration):** 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 2 Lite with Amazon Bedrock Guardrails, a feature designed to apply safeguards across multiple foundation models, knowledge bases, and agents, and Amazon Bedrock AgentCore, an agentic platform to build, deploy, and operate highly capable agents securely at scale. When designing agents, customers

should implement robust content filtering and moderation mechanisms across uncontrolled content that their agent system consumes. Bedrock Guardrails can filter harmful content, block denied topics, and redact sensitive information such as personally identifiable information. Bedrock AgentCore enables customers to set up interactions between Amazon Nova models, take actions across tools and data, run agents securely with low latency and extended runtimes, and monitor agents in production.

- **Reasoning:** Amazon Nova 2 Lite supports optional extended thinking capabilities, enabling deeper reasoning for complex problems requiring systematic analysis. Customers can control reasoning depth through three effort levels: "low" for moderately complex tasks, "medium" for substantial analysis, and "high" for the most thorough reasoning on highly complex, multi-faceted tasks. Extended thinking is recommended for use cases involving multi-step problem-solving, complex debugging, system architecture design, document synthesis, strategic planning, and agentic workflows requiring orchestration of multiple tool calls. Customers should disable extended thinking for straightforward tasks including simple factual questions, basic calculations, and speed-critical applications to optimize for efficiency and cost-effectiveness, as reasoning tokens are billed as output tokens.
- **Grounding:** Amazon Nova 2 Lite supports Web Grounding, a built-in tool that retrieves and incorporates publicly available information with citations as context for responses using current, real-time information. When Web Grounding is enabled the model automatically determines if search is needed, performs searches for relevant information, and synthesizes results with source citations. Web Grounding is particularly valuable for knowledge-based chat assistants, content generation requiring fact-checking, research assistants, and customer support applications where accuracy and verifiability are crucial. Customers are responsible for retaining and displaying citations in output provided to end users.
- **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.
- **Model Updates:** When we release new versions of Amazon Nova 2 Lite, customers may experience changes in performance on their use cases. 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 consider retesting the performance of the new Amazon Nova 2 Lite models on their use cases.

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.