

Public Summary of Training Content: Amazon Nova 2 Lite

Version of the Summary: 1.0

Last update: July 27, 2026

1. General information

1.1. Provider identification

Provider name and contact details: Amazon Media EU S.à r.l.
38, avenue John F. Kennedy
L-1855 Luxembourg
Grand Duchy of Luxembourg

Authorised representative name and contact details: N/A (provider established in the Union)

1.2. Model identification

Versioned model name(s): Nova 2 Lite (amazon.nova-2-lite-v1:0). The Nova 2 Lite model card is available at <https://docs.aws.amazon.com/bedrock/latest/userguide/model-card-amazon-nova-2-lite.html>.

Model dependencies: N/A

Date of placement of the model on the Union market: December 2, 2025

1.3 Modalities, overall training data size and other characteristics

Modality <i>Select the modalities present in the training data, to the extent that they are identifiable</i>	Training data size <i>For each selected modality, select the range within which the estimated total training data size for that modality falls. Dynamic datasets may be excluded from the estimation.</i>	Types of content <i>For each selected modality, provide a general description of the type of content that has been included in the training data.</i>
Text	☐ Less than 1 billion tokens ☐ 1 billion to 10 trillion tokens ☒ More than 10 trillion tokens	Nova 2 Lite was trained on text content that included reference materials, technical documentation, source code, and general web content.
Image	☐ Less than 1 million images ☐ 1 Million to 1 billion images ☒ More than 1 billion images	Nova 2 Lite was trained on image content that included photographs, graphs, tables, and screenshots.
Audio	☐ Less than 10,000 hours ☐ 10,000 to 1 million hours ☒ More than 1 million hours	Nova 2 Lite was trained on audio content that included conversational speech and sound effects.

Video	☐ Less than 10,000 hours ☒ 10,000 to 1 million hours ☐ More than 1 million hours	Nova 2 Lite was trained on video content that included instructional and general-interest audiovisual material.
Other	N/A	

Latest date of data acquisition/collection for model training:

October 2025. Nova 2 Lite is not continuously trained.

Description of the linguistic characteristics of the overall training data:

The overall training data for Nova 2 Lite includes over 200 languages, covering a broad range of EU official languages, with particular emphasis among them on Dutch, French, German, Italian, Portuguese, and Spanish, as well as other major world languages including Arabic, English, Hebrew, Hindi, Japanese, Korean, Simplified Chinese, and Turkish.

Other relevant characteristics of the overall training data:

The overall training data for Nova 2 Lite includes datasets intended to enhance the model's capabilities across modalities in various categories, including math, coding, language, logical reasoning, and factual, scientific, and audio-visual understanding.

Additional comments (optional):

More information on Amazon's training approach can be found in Amazon's [Generative AI Development Disclosure](#).

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model?

Yes

If yes, specify the modality(ies) of the content covered by the datasets concerned:

Text, image, video, and audio

List of large publicly available datasets:

Large publicly available datasets include data curated from public web content as well as datasets covering science and research topics.

General description of other publicly available datasets not listed above:

Other publicly available datasets span text, image, video, and audio content covering domains such as math, science, general knowledge, and reasoning, and are largely focused on the same languages as the rest of the training corpus.

Additional comments (optional):

N/A

2.2 Private non-publicly available datasets obtained from third parties

2.2.1. Datasets commercially licensed by rightsholders or their representatives

Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?

Yes

If yes, specify the modality(ies) of the content covered by the datasets concerned:

Text, image, video, and audio

2.2.2. Private datasets obtained from other third parties

Have you obtained private datasets from third parties that are not licensed as described in Section 2.2.1, such as data obtained from providers of private databases, or data intermediaries?

Yes

If yes, specify the modality(ies) of the content covered by the datasets concerned:

Text, image, video, and audio

If publicly known, list private datasets obtained from other third parties:

Information regarding private datasets obtained from third parties is subject to confidentiality terms.

General description of non-publicly known private datasets obtained from third parties

Private datasets cover subject areas including math, science, coding, general knowledge, and visual understanding, and are focused on the same languages as the rest of the training corpus.

Additional comments (optional):

N/A

2.3 Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?

Yes

If yes, specify crawler name(s)/identifier(s):

Amazonbot

Purposes of the crawler(s):

Amazonbot is Amazon's web crawler used to improve its services, including this model.

General description of crawler behaviour:	Amazonbot respects the Robots Exclusion Protocol defined at https://www.rfc-editor.org/rfc/rfc9309.html (i.e., robots.txt protocol), honoring the user-agent and the allow/disallow directives. Amazonbot also respects page-level robots meta tags, including 'noarchive', 'noindex', and 'nofollow', as well as the link-level 'nofollow' directive. Amazonbot is not designed to circumvent captchas or paywalls or to access password-protected content. For more information see https://developer.amazon.com/amazonbot .
Period of data collection:	From November 2023 to December 2024
Comprehensive description of the type of content and online sources crawled:	Data was crawled from a wide range of publicly available sources including sites focused on science, math, coding, business, and economics. This includes a wide variety of languages, focused on the same languages as the rest of the training corpus.
Type of modality covered:	Text, image, video, and audio
Summary of the most relevant domain names crawled:	The most relevant domains crawled include reference and knowledge sites, educational sources, scientific and technical repositories, business and financial information sites, and general-interest and region-specific portals. These domains span a wide variety of languages, focused on the same languages as the rest of the training corpus.
Additional comments (optional):	N/A

2.4 User data

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	Yes
Was data collected from user interactions with the provider's other services or products used to train the model?	Yes
If yes, provide a general description of the provider's services or products that were used to collect the user data:	In accordance with relevant terms of service and privacy policies, including service-specific policies, Amazon may use data collected from users of Amazon services (for example, public Amazon.com (US) product reviews) to train AI models. For more information about how Amazon collects and uses personal information, please see the Amazon Privacy Notice for the applicable country.
Type of modality covered:	Text, image, video, and audio

Additional comments (optional):

We are committed to building AI responsibly, with appropriate safeguards for safety, accuracy, privacy, and security. For more information about our approach to responsible AI, see [Responsible AI at Amazon](#).

2.5 Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?

Yes

If yes, modality of the synthetic data:

Text, image, video, and audio

If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:

Amazon generated synthetic data using its own and third-party general-purpose AI models, including Amazon Nova Premier.

Information about other AI models, including provider's own AI model(s) not available on the market, used to generate synthetic data to train the model to which this Summary applies:

Amazon may use internal models or third-party models to generate or to augment training data.

Additional comments (optional):

N/A

2.6 Other sources of data

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?

Yes

If yes, provide a narrative description of these data sources and the data:

Proprietary datasets created internally, which include human preference data and single and multi-turn responsible AI demonstrations in multiple languages.

Additional comments (optional)

N/A

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

Are you a Signatory to the Code of Practice for general-purpose AI models that includes commitments to respect reservations of rights from the TDM exception or limitation?

Yes

Describe the measures implemented before model training to respect reservations of rights from the TDM exception or limitation before and during data collection, including the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained:

As described in Section 2.3, Amazonbot reads and follows reservations of rights expressed through the Robots Exclusion Protocol (robots.txt). Amazonbot is not designed to circumvent captchas or paywalls or to access password-protected content. For more information see <https://developer.amazon.com/amazonbot>.

Additional comments (optional):

N/A

3.2 Removal of illegal content

General description of measures taken:

Measures are taken to detect and remove Child Sexual Abuse Material (CSAM), non-consensual intimate imagery (NCII), and other forms of illegal content from training data. Datasets are scanned for known CSAM, and such content is reported to the National Center for Missing & Exploited Children (NCMEC).

3.3 Other information (optional)

Other relevant information about data processing (optional):

N/A
