Public Summary of Training Content for NVIDIA Nemotron Nano 12B v2

Version 1.0

Last update: 25/09/2026

1. General information

1.1. Provider identification

Provider name and contact details	NVIDIA Corporation 2788 San Tomas Expressway, Santa Clara, CA 95051, USA
Authorised representative name and contact details	NVIDIA GmbH Adenauerstraße 20 A4 52146 Würselen, Germany AI_Governance@nvidia.com

1.2. Model identification

Versioned model name(s)	**NVIDIA Nemotron Nano 12B v2 (v1.0)** • NVIDIA-Nemotron-Nano-12B-v2 — https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2 • NVIDIA-Nemotron-Nano-12B-v2-Base — https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2-Base
Model dependencies	These models are not a fine-tune or modification of a general-purpose AI model.
Date of placement of the model on the Union market	**18 August 2025:** NVIDIA-Nemotron-Nano-12B-v2-Base **29 August 2025:** NVIDIA-Nemotron-Nano-12B-v2

1.3. Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
[X] Text	[ ] Less than 1 billion tokens [ ] 1 billion to 10 trillion tokens **[X] More than 10 trillion tokens**	Web pages, dialogue, articles and other written materials; source code in 43 programming languages; mathematical and scientific text; legal and regulatory text; biomedical and academic papers; financial filings; patents; question answering, instruction tuning, function and tool calling, and alignment data; reasoning traces; and synthetically generated and synthetically rephrased text.
[ ] Image	[ ] Less than 1 million images [ ] 1 million to 1 billion images [ ] More than 1 billion images	Not Applicable
[ ] Audio	[ ] Less than 10,000 hours [ ] 10,000 to 1 million hours [ ] More than 1 million hours	Not Applicable

[ ] Video	[ ] Less than 10,000 hours [ ] 10,000 to 1 million hours [ ] More than 1 million hours	Not Applicable
[ ] Other		Not Applicable

Latest date of data acquisition/collection for model training	June 2025
Description of linguistic characteristics	The pre-training corpus includes English, Arabic, Chinese, Danish, Dutch, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish and Thai.
Other relevant characteristics	The corpus spans domains including legal and regulatory, mathematics, science, medicine, finance, software engineering and general knowledge. A substantial proportion is synthetically generated or synthetically rephrased.
Additional comments (optional)	

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model?	**[X] Yes** [ ] No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	**[X] Text** [ ] Image [ ] Video [ ] Audio [ ] Other
List of large publicly available datasets:	Common Crawl (https://commoncrawl.org/); CC-NEWS (https://commoncrawl.org/blog/news-dataset-available); Wikimedia (https://dumps.wikimedia.org/); FineWeb-2 (https://huggingface.co/datasets/HuggingFaceFW/fineweb-2); mC4 (https://huggingface.co/datasets/legacy-datasets/mc4); The Stack (https://huggingface.co/datasets/bigcode/the-stack); Public Software Heritage S3 (https://docs.softwareheritage.org/devel/swh-export/graph/dataset.html); The Common Pile v0.1 (https://huggingface.co/common-pile); MegaMath (https://huggingface.co/datasets/LLM360/MegaMath); FineMath (https://huggingface.co/datasets/HuggingFaceTB/finemath); OpenWebMath (https://huggingface.co/datasets/open-web-math/open-web-math); MathPile (https://github.com/GAIR-NLP/MathPile/); OpenWebText2 (https://openwebtext2.readthedocs.io/en/latest/); arXiv (https://info.arxiv.org/help/bulk_data/index.html); peS2o (https://huggingface.co/datasets/allenai/peS2o); PMC Open Access Subset (https://pmc.ncbi.nlm.nih.gov/tools/openftlist/); PubMed Abstracts (https://github.com/thoppe/The-Pile-PubMed); NIH ExPorter (https://exporter.nih.gov/ExPORTER_Catalog.aspx); BioRxiv (https://www.biorxiv.org/tdm); Stack Exchange Data Dump (https://archive.org/details/stackexchange); Reddit Dataset

	(https://files.pushshift.io/reddit/); SEC EDGAR (https://www.sec.gov/search-filings); Court Listener (https://www.courtlistener.com/help/api/bulk-data/); BigScience Workshop Datasets (https://github.com/bigscience-workshop/bigscience); LMSYS-Chat-1M (https://huggingface.co/datasets/lmsys/lmsys-chat-1m).
General description of other publicly available datasets not listed above:	Text datasets covering instruction tuning, reasoning and chain-of-thought, question answering, mathematics and competitive mathematics, source code and programming problems, function and tool calling, safety alignment and red-teaming, moral and social reasoning, multilingual corpora, open textbooks used in commercial-use-permitting subsets, and academic, biomedical, legal and financial document collections.
Additional comments (optional)	

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 **[X] No**
If yes, specify the modality(ies) of the content covered by the datasets concerned:	[ ] Text [ ] Image [ ] Video [ ] Audio [ ] Other

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?	**[X] Yes** [ ] No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	**[X] Text** [ ] Image [ ] Video [ ] Audio [ ] Other
If publicly known, list private datasets obtained from other third parties:	Global Regulation; Workbench
General description of non-publicly known private datasets obtained from third parties	Regulatory and legislative text corpora obtained from a private database provider, and a workplace-task corpus used for tool-use and agentic evaluation and training.
Additional comments (optional)	

2.3. Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	[ ] Yes **[X] No**
Crawler Names/Identifier:	
Purposes of the crawler(s):	
General description of crawler behaviour	
Period of data collection	
Comprehensive description of the type of content and online sources crawled:	
Type of modality covered:	[ ] Text [ ] Image [ ] Video [ ] Audio [ ] Other
Summary of the most relevant domain names crawled:	
Additional comments (optional)	

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 **[X] No**
Was data collected from user interactions with the provider's other services or products used to train the model?	[ ] Yes **[X] No**
If yes, provide a general description of the provider's services or products that were used to collect the user data:	N/A
Type of modality covered:	N/A
Additional comments (optional)	

2.5. Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	**[X] Yes** [ ] No
If yes, modality of the synthetic data:	**[X] Text** [ ] Image [ ] Video [ ] Audio [ ] Other
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	**DeepSeek:** DeepSeek-R1; DeepSeek-R1-0528; DeepSeek-V3; DeepSeek-V3-0324; DeepSeek-R1-Distill-Qwen-32B **Alibaba / Qwen:** Qwen2.5-0.5B-Instruct; Qwen2.5-14B-Instruct; Qwen2.5-32B-Instruct; Qwen2.5-32B-Instruct-AWQ; Qwen2.5-72B;

	Qwen2.5-72B-Instruct; Qwen2.5-Math-7B; Qwen2.5-Math-72B; Qwen3-30B-A3B; Qwen3-235B-A22B **Mistral:** Mixtral-8x7B-Instruct-v0.1; Mixtral-8x22B-v0.1; Mixtral-8x22B-Instruct-v0.1; Mistral-NeMo-12B-Instruct **Microsoft:** phi-4 **Google:** Gemma 3 PT 12B; Gemma 3 PT 27B **NVIDIA:** Nemotron-4-340B-Base; Nemotron-4-340B-Instruct; Nemotron-4-340B-Reward
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:	Nemotron-4-15B-Base and Nemotron-4-15B-Instruct, internal NVIDIA checkpoints not released to the market, were used to generate alignment data. A Gemma-3 4B-based filtering model fine-tuned on the Nemotron Content Safety Dataset v2 was used to quality-filter and safety-filter synthetically generated data.
Additional comments (optional)	

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?	**[X] Yes** [ ] No
If yes, provide a narrative description of these data sources and the data:	NVIDIA-created internal datasets used as seeds for alignment data generation, comprising a model-identity dataset and an internal multi-turn conversational dataset. Publicly available source code from GitHub using the GitHub REST API and publicly available source code from Amazon Web Services using the Amazon S3 API.
Additional comments (optional)	

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 **[X] No**
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:	NVIDIA implements measures to respect EU rights reservations relevant to text and data mining by: (1) respecting robots.txt directives at the domains accessed where those directives constituted a reservation of rights under Article 4(3) of Directive (EU) 2019/790; and (2) filtering datasets on any actionable metadata identifiers provided by rightsholders.

Additional comments (optional)	

3.2. Removal of illegal content

General description of measures taken	NVIDIA applies measures to avoid or remove illegal content under Union law from the training data, which may include human safeguards, use of guard models and model-based classifiers, content-safety checks and screening, automated filtering, keyword-based rules, and internal validation. Post-training measures may also include benchmark evaluations and ongoing monitoring.

3.3. Other information (optional)

Other relevant information about data processing (optional)	N/A

Public Summary of Training Content for NVIDIA Nemotron Nano v2 12B VL

1. General information

1.1. Provider identification

Provider name and contact details	NVIDIA Corporation 2788 San Tomas Expressway Santa Clara, CA 95051, USA
Authorised representative name and contact details	NVIDIA GmbH Adenauerstraße 20 A4 52146 Würselen, Germany AI_Governance@nvidia.com

1.2. Model identification

Versioned model name(s)	NVIDIA-Nemotron-Nano-12B-v2-VL-BF16 (https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-BF16) NVIDIA-Nemotron-Nano-12B-v2-VL-FP8 (https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-FP8) NVIDIA-Nemotron-Nano-12B-v2-VL-NVFP4-QAD (https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-12B-v2-VL-NVFP4-QAD)
Model dependencies	Developed from NVIDIA-Nemotron-Nano-12B-v2.
Date of placement of the model on the Union market	27 October 2025

1.3. Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
[X] Text	[ ] Less than 1 billion tokens [X] 1 billion to 10 trillion tokens [] More than 10 trillion tokens	Instruction tuning, reasoning and question answering, source code, Optical Character Recognition (OCR)-derived text, captions, document text, and synthetic text.
[X] Image	[ ] Less than 1 million images [X] 1 million to 1 billion images [ ] More than 1 billion images	Natural and digital images, business documents, diagrams, charts, tables, scientific documents, text-bearing images, and synthetically rendered OCR material.
[ ] Audio	☐ Less than 10 000 hours ☐ 10,000 to1 million hours ☐ More than 1 million hours	
[X] Video	[ ] Less than 10.000 hours [X] 10,000 to 1 billion hours [] More than 1 billion hours	Audiovisual content, including instructional and activity videos and general-interest audiovisual material, captions, question and answers.
[ ] Other		N/A

Latest date of data acquisition/collection for model training	2025

Description of linguistic characteristics	The training corpus includes a variety of languages, including English, Chinese, German, Spanish, French, Italian, Japanese, Korean, Dutch, and Portuguese.
Other relevant characteristics	
Additional comments (optional)	

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model?	[X] Yes [ ] No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	[X] Text [X] Image [X] Video [ ] Audio [ ] Other
List of large publicly available datasets:	See Appendix A.1

General description of other publicly available datasets not listed above:	Text datasets including instruction tuning, reasoning, question answering, safety alignment, and function calling. Image-and-text datasets containing OCR, document and image reasoning, visual question answering (VQA), captioning, visual instruction tuning, safety, grounding, and classification. Video-and-text datasets including video question answering, captioning, temporal and spatial reasoning, action recognition, and safety.
Additional comments (optional)	

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 [X] No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	[] Text [ ] Image [ ] Video [ ] Audio [ ] Other

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?	[X] Yes [ ] No

If yes, specify the modality(ies) of the content covered by the datasets concerned:	[X] Text [X] Image [ ] Video [ ] Audio [ ] Other
If publicly known, list private datasets obtained from other third parties:	N/A
General description of non-publicly known private datasets obtained from third parties	Invoice documents with question-answer annotations.
Additional comments (optional)	

2.3. Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	[X] Yes [ ] No
Crawler Names/Identifier:	NVIDIA Crawler Identifier: nvidiabot Additional third-party crawler identifiers unknown.
Purposes of the crawler(s):	Acquire or curate images for captioning and enterprise visual question answering (VQA); derive dense image-caption data from Common Crawl materials; and create video captioning, question-answering, and reasoning datasets.
General description of crawler behaviour	Automated crawlers were used to collect publicly available media from selected websites. The crawlers navigated publicly accessible pages, extracted relevant links and associated metadata, and collected eligible content for model training and evaluation.
Period of data collection	2023-2025
Comprehensive description of the type of content and online sources crawled:	Natural and digital images used for captioning and VQA, plus videos used for captioning, question answering, and reasoning.
Type of modality covered:	[X] Text [X] Image [X] Video [ ] Audio [ ] Other
Summary of the most relevant domain names crawled:	
Additional comments (optional)	

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 [X] No.

Was data collected from user interactions with the provider's other services or products used to train the model?	[] Yes [X] No.
If yes, provide a general description of the provider's services or products that were used to collect the user data:	
Type of modality covered:	
Additional comments (optional)	

2.5. Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	[X] Yes [ ] No
If yes, modality of the synthetic data:	[X] Text [X] Image [ ] Video [ ] Audio [ ] Other
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	DeepSeek-R1-0528; Qwen3-235B-A22B; Qwen3-30B; Qwen2.5-32B-Instruct-AWQ; Qwen2.5-14B-Instruct; Qwen2.5-VL-32B-Instruct
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:	
Additional comments (optional)	

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 [X] No
If yes, provide a narrative description of these data sources and the data:	
Additional comments (optional)	

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 [X] No
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:	NVIDIA implements measures to respect EU rights reservations relevant to text and data mining by: (1) respecting robots.txt directives at the domains accessed where those directives constituted a reservation of rights under Article 4(3) of Directive (EU) 2019/790; and (2) filtering datasets on any actionable metadata identifiers provided by rightsholders.
Additional comments (optional)	

3.2. Removal of illegal content

General description of measures taken	Image-based datasets were scanned against known child sexual abuse material (CSAM), with the inventory recording passed compliance scans or reuse of prior scans for identical source images where applicable. Safety-alignment datasets, rule-based filtering, and model-based filtering were also used. Datasets were subject to content-safety checks appropriate to their source and modality. Web-derived image-text data underwent automated NSFW screening designed to identify explicit imagery and profane or inappropriate accompanying text, with flagged samples removed before using the data for training.

3.3. Other information (optional)

Other relevant information about data processing (optional)	N/A

Appendix A. Datasets

A.1. Publicly available datasets

A.1.1. Text datasets

Qualifying source lineage	Public source
STEM MMLU-aux (subset)	Source
MetaMathQA (CoT)	Source

A.1.2. Image-and-text datasets

Qualifying source lineage	Public source
Open Images (captioning and bounding-box-derived subsets)	Source
Wikipedia-based Image Text (WIT)	Source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source

A.1.2. Image-and-text datasets

Qualifying source lineage	Public source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source
arXiv-derived OCR and equation data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source

A.1.2. Image-and-text datasets

Qualifying source lineage	Public source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source

A.1.2. Image-and-text datasets

Qualifying source lineage	Public source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source
Wikimedia enterprise HTML-derived OCR data	Source

A.1.2. Image-and-text datasets

Qualifying source lineage	Public source
Open Images (captioning and bounding-box-derived subsets)	Source
Open Images (captioning and bounding-box-derived subsets)	Source
Open Images (captioning and bounding-box-derived subsets)	Source
arXiv-derived OCR and equation data	Source

A.1.3. Video-and-text datasets

Qualifying source lineage	Public source
ActivityNet	Source
ActivityNet	Source
ActivityNet	Source
ActivityNet	Source
Kinetics	Source
EgoExoLearn	Source
Kinetics	Source
ActivityNet	Source

A.1.3. Video-and-text datasets

Qualifying source lineage	Public source
HACS	Source
ActivityNet	Source
EgoExoLearn	Source
EgoExoLearn	Source
HC-STVG	Source
HC-STVG	Source

Public Summary of Training Content for NVIDIA Nemotron 3 and 3.5 Family

Version 1.0
Last update: 10/08/2026

1. General information

1.1. Provider identification

Provider name and contact details	NVIDIA Corporation 2788 San Tomas Expressway, Santa Clara, CA 95051, USA
Authorised representative name and contact details	NVIDIA GmbH Adenauerstraße 20 A4 52146 Würselen, Germany AI_Governance@nvidia.com

1.2. Model identification

Versioned model name(s)	**Nemotron 3 Nano 30B A3B (v1.0)** • NVIDIA-Nemotron-3-Nano-30B-A3B-Base-BF16 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-Base-BF16 • NVIDIA-Nemotron-3-Nano-30B-A3B-BF16 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-BF16 • NVIDIA-Nemotron-3-Nano-30B-A3B-FP8 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-FP8

	• NVIDIA-Nemotron-3-Nano-30B-A3B-NVFP4 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-NVFP4 **Nemotron 3 Nano Omni 30B A3B Reasoning (v1.0)** • Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16 — https://huggingface.co/nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16 • Nemotron-3-Nano-Omni-30B-A3B-Reasoning-FP8 — https://huggingface.co/nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-FP8 • Nemotron-3-Nano-Omni-30B-A3B-Reasoning-NVFP4 — https://huggingface.co/nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-NVFP4 **Nemotron 3 Super 120B A12B (v1.0 GA / v1.0 Base)** • NVIDIA-Nemotron-3-Super-120B-A12B-Base-BF16 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-Base-BF16 • NVIDIA-Nemotron-3-Super-120B-A12B-BF16 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-BF16 • NVIDIA-Nemotron-3-Super-120B-A12B-FP8 —

	https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-FP8 • NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4 **Nemotron 3 Ultra 550B A55B (v1.0 GA)** • NVIDIA-Nemotron-3-Ultra-550B-A55B-Base-BF16 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-Base-BF16 • NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16 • NVIDIA-Nemotron-3-Ultra-550B-A55B-NVFP4 — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-NVFP4 • NVIDIA-Nemotron-3-Ultra-550B-A55B-GenRM — https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-GenRM **Nemotron 3.5** • https://huggingface.co/nvidia/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-NVFP4 • https://huggingface.co/nvidia/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-Base-BF16 •

	https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16 • https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-NVFP4-DSpark • https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-NVFP4-DFlash
Model dependencies	**Nemotron 3.5 Lightning 30B A3B** is developed from **NVIDIA-Nemotron-3-Nano-30B-A3B** (Mamba2-Transformer hybrid MoE backbone). **Nano 30B A3B, Super 120B A12B and Ultra 550B A55B** are not fine-tunes or modifications of a previously placed general-purpose AI model. FP8 and NVFP4 variants are post-training quantisations of the corresponding BF16 checkpoint and were not separately trained. **Nano Omni 30B A3B Reasoning** is developed from NVIDIA-Nemotron-3-Nano-30B-A3B (Mamba2-Transformer hybrid MoE backbone).
Date of placement of the model on the Union market	**15 December 2025:** Nemotron 3 Nano 30B A3B (all variants) **11 March 2026:** Nemotron 3 Super 120B A12B (aligned, FP8, NVFP4). **28 April 2026:** Nemotron 3 Nano Omni 30B A3B Reasoning (all variants)

	4 June 2026: Nemotron 3 Ultra 550B A55B (all variants) **11 August 2026:** Nemotron 3.5 Lightning 30B A3B (all variants)

1.3. Modalities, overall training data size and other characteristics

Modality	**Training data size**	**Types of content**
[X] Text	[ ] Less than 1 billion tokens [ ] 1 billion to 10 trillion tokens [X] More than 10 trillion tokens	Web pages, dialogue, articles and other written materials; source code in 43 programming languages; mathematical and scientific text; legal and regulatory text; biomedical text; competitive programming and software engineering; financial filings; academic papers and patents; question answering, instruction tuning, tool use, function and tool calling, structured-output and alignment data; reasoning traces; and synthetically generated and synthetically rephrased text.
[X] Image	[ ] Less than 1 million images [X] 1 million to 1 billion images [ ] More than 1 billion images	**Applies to Nemotron 3 Nano Omni Only:** Natural and digital images, business and scientific documents, PDFs, charts, plots, diagrams, tables, screenshots and graphical user interfaces, text-bearing images and

Modality	Training data size	Types of content
		OCR material, medical images, geometry and mathematics figures, visual question answering, captions, and synthetically rendered or re-captioned imagery.
[X] Audio	[ ] Less than 10,000 hours [X] 10,000 to 1 million hours [ ] More than 1 million hours	**Applies to Nemotron 3 Nano Omni Only:** Speech and general audio for automatic speech recognition, speech-to-text chat, audio question answering and audio captioning, including text-to-speech-synthesised material.
[X] Video	[ ] Less than 10,000 hours [X] 10,000 to 1 million hours [ ] More than 1 million hours	**Applies to Nemotron 3 Nano Omni Only:** Audiovisual content including instructional and activity video, general-interest video, egocentric and procedural footage, with captions, question-answer pairs and temporal/spatial reasoning annotations.
[ ] Other		**Not Applicable**

Latest date of data acquisition/collection for model training	**August 2026**
Description of linguistic characteristics	The pre-training corpus includes English, Arabic, Chinese, Czech, Danish, Dutch, Finnish, French, German, Hebrew, Hindi,

	Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish and Thai.
Other relevant characteristics	The corpus spans domains including legal and regulatory, mathematics, science, medicine, finance, software engineering and general knowledge. A substantial proportion is synthetically generated or synthetically rephrased.
Additional comments (optional)	

---

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model?	[X] Yes [ ] No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	[X] Text [X] Image [X] Video [X] Audio [ ] Other
List of large publicly available datasets:	Common Crawl (https://commoncrawl.org/); CC-NEWS (https://commoncrawl.org/blog/news-dataset-available); Wikimedia (https://dumps.wikimedia.org/); FineWeb-2 (https://huggingface.co/datasets/HuggingFaceFW/fineweb-2); finepdfs (https://huggingface.co/datasets/HuggingFaceFW/finepdfs); Essential-Web v1.0 (https://huggingface.co/datasets/EssentialA

	l/essential-web-v1.0); The Common Pile v0.1 (https://huggingface.co/common-pile); MegaMath (https://huggingface.co/datasets/LLM360/MegaMath); FineMath (https://huggingface.co/datasets/HuggingFaceTB/finemath); arXiv (https://info.arxiv.org/help/bulk_data/index.html); peS2o (https://huggingface.co/datasets/allenai/peS2o); Stack Exchange Data Dump (https://archive.org/details/stackexchange); Reddit Dataset (https://files.pushshift.io/reddit/); SEC EDGAR (https://www.sec.gov/search-filings); Court Listener (https://www.courtlistener.com/help/api/bulk-data/); LMSYS-Chat-1M (https://huggingface.co/datasets/lmsys/lmsys-chat-1m); WildChat-1M (https://huggingface.co/datasets/allenai/WildChat-1M); MiraData; laion-disco-12M; OpenImages (https://storage.googleapis.com/openimages/web/index.html); Conceptual Captions 3M; MINT-1T.

General description of other publicly available datasets not listed above:	**Text:** instruction tuning, reasoning, question answering, mathematics, competitive programming and software engineering, function and tool calling, safety alignment, translation and multilingual corpora, and academic/biomedical/legal/financial document collections. **Image-and-text:** OCR and document understanding, chart/table/diagram

	reasoning, visual question answering, captioning, content for visual grounding, computer-use and GUI interaction, visual instruction tuning, geometry and vision-critical STEM, and classification. **Video-and-text:** video question answering, captioning, temporal and spatial reasoning, procedural and egocentric activity understanding. **Audio-and-text:** automatic speech recognition, speech instruction following, audio captioning and audio question answering.
Additional comments (optional)	

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?	[X] Yes [ ] No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	[X] Text [X] Image [X] Video [X] Audio [ ] Other

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?	[X] Yes [ ] No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	[X] Text [X] Image [X] Video [X] Audio [ ] Other
If publicly known, list private datasets obtained from other third parties:	Global Regulation; TAUS Translation Memory; Scale HLE; HackerRank Coding; Granary; SIFT-50M; Mercor (agentic terminal-task data and agentic-search tool-use data); Turing Math Data Pack; Patronus (dialog memory data); "Droid Harness Pivot" data; "Chemistry Ether0" data.
General description of non-publicly known private datasets obtained from third parties:	Text; professional translation memories; expert-authored examination and reasoning items; competitive-programming problem sets and solutions; speech and speech-instruction; and vendor-produced agentic task, tool-use and dialogue datasets with human annotation.
Additional comments (optional)	

2.3. Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	[X] Yes [ ] No

Crawler Names/Identifier:	NVIDIA Crawler Identifier: nvidiabot Additional third-party crawler identifiers unknown.
Purposes of the crawler(s):	Acquire large-scale web text for pre-training; acquire permissively licensed source code for code pre-training; acquire images and image-text pairs for captioning, OCR, document understanding and visual question answering; and acquire video and audiovisual material for captioning, question answering and temporal reasoning.
General description of crawler behaviour	Automated crawlers were used to collect publicly available media from selected websites. The crawlers navigated publicly accessible pages, extracted relevant links and associated metadata, and collected eligible content for model training and evaluation. Each crawl was operated in accordance with the rate limits set by its respective source.
Period of data collection	2013–2026
Comprehensive description of the type of content and online sources crawled:	**Images and documents:** publicly accessible web images, documents and PDFs used for captioning, OCR and document/visual question answering. **Video and audio:** publicly accessible video and audiovisual material used for captioning, question answering and reasoning.
Type of modality covered:	[X] Text [X] Image [X] Video [X] Audio [ ] Other
Summary of the most relevant domain names crawled:	github.com; youtube.com
Additional comments (optional)	

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 [X] No
Was data collected from user interactions with the provider's other services or products used to train the model?	[ ] Yes [X] No
If yes, provide a general description of the provider's services or products that were used to collect the user data:	N/A
Type of modality covered:	N/A
Additional comments (optional)	

2.5. Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	[X] Yes [ ] No
If yes, modality of the synthetic data:	[X] Text [X] Image [X] Video [X] Audio [ ] Other
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	**Alibaba / Qwen:** Qwen2.5-0.5B-Instruct; Qwen2.5-14B-Instruct; Qwen2.5-32B-Instruct; Qwen2.5-72B; Qwen2.5-72B-Instruct; Qwen2.5-Math-7B; Qwen2.5-Math-72B; Qwen2.5-VL-32B-Instruct; Qwen2.5-VL-72B-Instruct; Qwen2-VL-72B; Qwen3-8B; Qwen3-30B-A3B (and Instruct-2507 / Thinking-2507); Qwen3-32B; Qwen3-235B-A22B (and Instruct-2507 / Thinking-2507);

	Qwen3-Next-80B-A3B (Instruct / Thinking); Qwen3-Coder-480B-A35B-Instruct; Qwen3-VL-30B-A3B (Instruct / Thinking-FP8); Qwen3-VL-32B-Instruct; Qwen3-VL-235B-A22B (Instruct / Thinking, incl. FP8); Qwen3-Omni-30B-A3B-Instruct; Qwen3-omni-captioner; Qwen3.5-122B-A10B; Qwen3.5-397B-A17B; QwQ-32B **DeepSeek:** DeepSeek-V3; V3-0324; V3.2; V3.2-Speciale; V4-Pro; R1; R1-0528; R1-Distill-Qwen-7B; R1-Distill-Qwen-32B; DeepSeek OCR **OpenAI:** gpt-oss-20b; gpt-oss-120b **Google:** Gemma-2-2b-it; Gemma-3-4b-it; Gemma-3-27b-it; Gemini 3 **Mistral:** Mixtral-8x7B-v0.1; Mixtral-8x22B-v0.1; Mixtral-8x22B-Instruct-v0.1; Mistral-7B-Instruct-v0.3; Mistral-Small-3.1-24B-Instruct-2503; Mistral-Small-3.2-24B-Instruct-2506; Mistral-NeMo-12B-Instruct; Ministral-3-14B-Reasoning-2512 **Microsoft:** phi-4; Phi-4-mini-instruct; Phi-3-small-8k-instruct; Phi-3-medium-4k-instruct; Florence-2-large **Zhipu / Z.ai:** GLM-4.1V-9B-Thinking; GLM-4.6; GLM-4.7 (incl. Flash and FP8); GLM-5 **Moonshot:** Kimi-K2-Instruct; Kimi-K2-Thinking; Kimi-K2.5 **MiniMax:** MiniMax-M1-40k; MiniMax-M1-80k; MiniMax-M2 **Other:** Goedel-Prover-V2-32B; Ring-1T; SwinDocSegmenter

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:	NVIDIA-Nemotron-3-Ultra-Teacher-STEM NVIDIA-Nemotron-3-Ultra-Teacher-General-Reasoning NVIDIA-Nemotron-3-Ultra-Teacher-Office-Work NVIDIA-Nemotron-3-Ultra-Teacher-Terminal-Use NVIDIA-Nemotron-3-Ultra-Teacher-Search NVIDIA-Nemotron-3-Ultra-Teacher-General-Agentic NVIDIA-Nemotron-3-Ultra-Teacher-Conversational-Tool-Use NVIDIA-Nemotron-3-Ultra-Teacher-Instruction-Following NVIDIA-Nemotron-3-Ultra-Teacher-Competitive-Coding NVIDIA-Nemotron-3-Ultra-Teacher-Chat NVIDIA-Nemotron-3-Ultra-Teacher-Agentic-Safety The teacher models were trained using the same broad, multidisciplinary data pool described above, with tailored data selection, mixture weighting, synthetic-data generation, and training recipes designed to emphasize the capabilities associated with each domain, including STEM, reasoning, office work, terminal and search use, agentic and conversational tool use, instruction following, competitive coding, chat, and agentic safety.
Additional comments (optional)	

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?	[X] Yes [ ] No
If yes, provide a narrative description of these data sources and the data:	NVIDIA-created and NVIDIA-generated reasoning and puzzle environments, and machine-translated corpora; internal competitive-coding, long-context and software-engineering reinforcement-learning environments; internal CUDA kernel task data; internal tool-use and agentic interaction data; hand-written prompt sets and model-identity data; and internal safety datasets.
Additional comments (optional)	

---

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 [X] No
Describe the measures implemented before model training to respect reservations of rights from the TDM	NVIDIA implements measures to respect EU rights reservations relevant to text and data mining by: (1) respecting robots.txt

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:	directives at the domains accessed where those directives constituted a reservation of rights under Article 4(3) of Directive (EU) 2019/790; and (2) filtering datasets on any actionable metadata identifiers provided by rightsholders.
Additional comments (optional)	

3.2. Removal of illegal content

General description of measures taken	Training datasets were passed through layered automated and human safeguards designed to reduce the presence of illegal, harmful or policy-violating content, including child sexual abuse material (CSAM), non-consensual intimate imagery (NCII), sexual content involving minors, weapons and criminal-planning content, harassment, hate speech, profanity, threats and incitement to violence, self-harm and suicide-related content, and graphic violence. A guard model fine-tuned on the Nemotron Content Safety Dataset v2 was used to exclude potentially illegal or harmful content from training and to quality-filter synthetically generated data. For multimodal data, two independent safety judge models were applied to flag and remove harmful samples. Image-based datasets were scanned against known CSAM, with the inventory recording passed compliance scans or reuse of prior scans for identical source images where applicable. Safety-alignment

	datasets, rule-based filtering, and model-based filtering were also used. Datasets were subject to content-safety checks appropriate to their source and modality. Web-derived image-text data underwent automated NSFW screening designed to identify explicit imagery and profane or inappropriate accompanying text, with flagged samples removed before using the data for training. Synthetic and simulation-generated data are evaluated through internal validation before inclusion. Benchmark evaluations and red-team testing are applied post-training to surface remaining safety gaps across world generation, reasoning, audio, and action tasks. Ongoing monitoring and dataset review continue post-release.

3.3. Other information (optional)

Other relevant information about data processing (optional)	