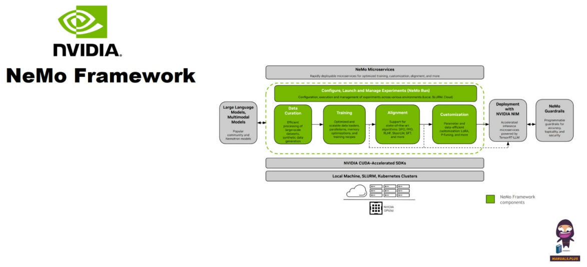
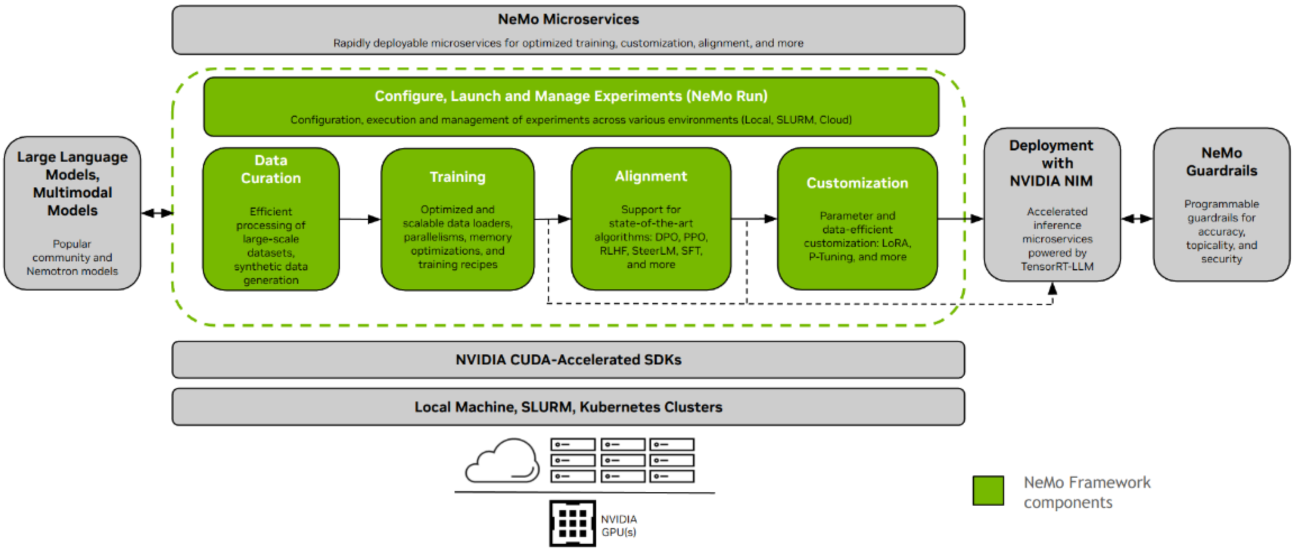

NVIDIA
NeMo Framework
NeMo Microservices
Rapidly deployable microservices for optimized training, customization, alignment, and more
Configure, Launch and Manage Experiments (NeMo Run)
Configuration, execution and management of experiments across various environments (Local, SLURM, Cloud)
Large Language Models, Multimodal Models
Popular community and Nemotron models
Data Curation
Efficient processing of large-scale datasets, synthetic data generation
Training
Optimized and scalable data loaders, parallelisms, memory optimizations, and training recipes
Alignment
Support for state-of-the-art algorithms: DPO, PPO, RLHF, SteerLM, SFT, and more
Customization
Parameter and data-efficient customization: LoRA, P-Tuning, and more
Deployment with NVIDIA NIM
Accelerated inference microservices powered by TensorRT-LLM
NeMo Guardrails
Programmable guardrails for accuracy, topicality, and security
NVIDIA CUDA-Accelerated SDKs
Local Machine, SLURM, Kubernetes Clusters
NVIDIA GPU(s)
NeMo Framework components
MANUALS.PLUS

NVIDIA NeMo Framework User Guide

Home » Nvidia » NVIDIA NeMo Framework User Guide

Contents

NVIDIA NeMo Framework


NeMo Microservices
Rapidly deployable microservices for optimized training, customization, alignment, and more
Configure, Launch and Manage Experiments (NeMo Run)
Configuration, execution and management of experiments across various environments (Local, SLURM, Cloud)
Large Language Models, Multimodal Models
Popular community and Nemotron models
Data Curation
Efficient processing of large-scale datasets, synthetic data generation
Training
Optimized and scalable data loaders, parallelisms, memory optimizations, and training recipes
Alignment
Support for state-of-the-art algorithms: DPO, PPO, RLHF, SteerLM, SFT, and more
Customization
Parameter and data-efficient customization: LoRA, P-Tuning, and more
Deployment with NVIDIA NIM
Accelerated inference microservices powered by TensorRT-LLM
NeMo Guardrails
Programmable guardrails for accuracy, topicality, and security
NVIDIA CUDA-Accelerated SDKs
Local Machine, SLURM, Kubernetes Clusters
NVIDIA GPU(s)
NeMo Framework components

Specifications

- **Product Name**: NVIDIA NeMo Framework
- **Affected Platforms:** Windows, Linux, macOS
- **Affected Versions:** All versions prior to 24
- **Security Vulnerability:** CVE-2025-23360
- **Risk Assessment Base Score:** 7.1 (CVSS v3.1)

Product Usage Instructions

Security Update Installation:

To protect your system, follow these steps:

1. Download the latest release from the NeMo-Framework-Launcher Releases page on GitHub.
2. Go to NVIDIA Product Security for further information.

Security Update Details:

The security update addresses a vulnerability in the NVIDIA NeMo Framework that could lead to code execution and data tampering.

Software Upgrade:

If you are using an earlier branch release, it is recommended to upgrade to the latest branch release to address the security issue.

Overview

NVIDIA NeMo Framework is a scalable and cloud-native generative AI framework built for researchers and developers working on Large Language Models, Multimodal, and Speech AI (e.g. Automatic Speech Recognition and Text-to-Speech). It enables users to efficiently create, customize, and deploy new generative AI models by leveraging existing code and pre-trained model checkpoints.

Setup Instructions: Install NeMo Framework

Large Language Models and Multimodal Models

NeMo Framework provides end-to-end support for developing Large Language Models (LLMs) and Multimodal

Models (MMs). It provides the flexibility to be used on-premises, in a data-center, or with your preferred cloud provider. It also supports execution on SLURM or Kubernetes enabled environments.

_images/nemo-llm-mm-stack.png

Data Curation
NeMo Curator [1] is a Python library that includes a suite of modules for data mining and synthetic data generation. They are scalable and optimized for GPUs, making them ideal for curating natural language data to train or fine-tune LLMs. With NeMo Curator, you can efficiently extract high-quality text from extensive raw web data sources.

Training and Customization

NeMo Framework provides tools for efficient training and customization of LLMs and Multimodal models. It includes default configurations for compute cluster setup, data downloading, and model hyperparameters, which can be adjusted to train on new datasets and models. In addition to pre-training, NeMo supports both Supervised Fine-Tuning (SFT) and Parameter Efficient Fine-Tuning (PEFT) techniques like LoRA, Ptuning, and more.

Two options are available to launch training in NeMo – using the NeMo 2.0 API interface or with NeMo Run.

- **With NeMo Run (Recommended):** NeMo Run provides an interface to streamline configuration, execution and management of experiments across various compute environments. This includes launching jobs on your workstation locally or on big clusters – both SLURM enabled or Kubernetes in a cloud environment.
  - Pre-training & PEFT Quickstart with NeMo Run
- **Using the NeMo 2.0 API:** This method works well with a simple setup involving small models, or if you are interested in writing your own custom dataloader, training loops, or change model layers. It gives you more flexibility and control over configurations, and makes it easy to extend and customize configurations programmatically.
  - Training Quickstart with NeMo 2.0 API
  - Migrating from NeMo 1.0 to NeMo 2.0 API

Alignment

- NeMo-Aligner [1] is a scalable toolkit for efficient model alignment. The toolkit has support for state-of-the-art model alignment algorithms such as SteerLM, DPO, Reinforcement Learning from Human Feedback (RLHF), and much more. These algorithms enable users to align language models to be more safe, harmless, and helpful.
- All the NeMo-Aligner checkpoints are cross-compatible with the NeMo ecosystem, allowing for further customization and inference deployment.

Step-by-step workflow of all three phases of RLHF on a small GPT-2B model:

- SFT training
- Reward model training
- PPO training

In addition, we demonstrate support for various other novel alignment methods:

- **DPO**: a lightweight alignment algorithm compared to RLHF with a simpler loss function.
- **Self-Play** Fine-Tuning (SPIN)
- **SteerLM:** a technique based on conditioned-SFT, with steerable output.

Check out the documentation for more information: Alignment Documentation

Multimodal Models

- NeMo Framework provides optimized software to train and deploy state-of-the-art multimodal models across several categories: Multimodal Language Models, Vision-Language Foundations, Text-to-Image models, and beyond 2D Generation using Neural Radiance Fields (NeRF).
- Each category is designed to cater to specific needs and advancements in the field, leveraging cutting-edge models to handle a wide range of data types, including text, images, and 3D models.

Note
We are migrating support for multimodal models from NeMo 1.0 to NeMo 2.0. If you want to explore this domain in the meantime, please refer to the documentation for the NeMo 24.07 (previous) release.

Deployment and Inference
NeMo Framework provides various paths for LLM inference, catering to different deployment scenarios and performance needs.

Deploy with NVIDIA NIM

- NeMo Framework seamlessly integrates with enterprise-level model deployment tools through NVIDIA NIM. This integration is powered by NVIDIA TensorRT-LLM, ensuring optimized and scalable inference.
- For more information on NIM, visit the NVIDIA website.

Deploy with TensorRT-LLM or vLLM

- NeMo Framework offers scripts and APIs to export models to two inference optimized libraries, TensorRT-LLM and vLLM, and to deploy the exported model with the NVIDIA Triton Inference Server.
- For scenarios requiring optimized performance, NeMo models can leverage TensorRT-LLM, a specialized library for accelerating and optimizing LLM inference on NVIDIA GPUs. This process involves converting NeMo models into a format compatible with TensorRT-LLM using the nemo.export module.
    - LLM Deployment Overview
    - Deploy NeMo Large Language Models with NIM
    - Deploy NeMo Large Language Models with TensorRT-LLM
    - Deploy NeMo Large Language Models with vLLM

Supported Models

Large Language Models

Large Language Models

Large Language Models	Pretraining & SFT	PEFT	Alignment	FP8 Training Convergence	TRT/TRTLLM	Convert To & From Hugging Face	Evaluation
Llama3 8B/70B, Llama3.1 405B	Yes	Yes	x	Yes (partially verified)	Yes	Both	Yes
Mixtral 8x7B/8x22B	Yes	Yes	x	Yes (unverified)	Yes	Both	Yes
Nemotron 3 8B	Yes	x	x	Yes (unverified)	x	Both	Yes
Nemotron 4 340B	Yes	x	x	Yes (unverified)	x	Both	Yes
Baichuan2 7B	Yes	Yes	x	Yes (unverified)	x	Both	Yes
ChatGLM3 6B	Yes	Yes	x	Yes (unverified)	x	Both	Yes
Gemma 2B/7B	Yes	Yes	x	Yes (unverified)	Yes	Both	Yes
Gemma2 2B/9B/27B	Yes	Yes	x	Yes (unverified)	x	Both	Yes
Mamba2 130M/370M/780M/1.3B/2.7B/8B/ Hybrid-8B	Yes	Yes	x	Yes (unverified)	x	x	Yes
Phi3 mini 4k	x	Yes	x	Yes (unverified)	x	x	x
Qwen2 0.5B/1.5B/7B/72B	Yes	Yes	x	Yes (unverified)	Yes	Both	Yes
StarCoder 15B	Yes	Yes	x	Yes (unverified)	Yes	Both	Yes
StarCoder2 3B/7B/15B	Yes	Yes	x	Yes (unverified)	Yes	Both	Yes
BERT 110M/340M	Yes	Yes	x	Yes (unverified)	x	Both	x
T5 220M/3B/11B	Yes	Yes	x	x	x	x	x

Vision Language Models

Vision Language Models

Vision Language Models	Pretraining & SFT	PEFT	Alignment	FP8 Training Convergence	TRT/TRTLLM	Convert To & From Hugging Face	Evaluation
NeVA (LLaVA 1.5)	Yes	Yes	x	Yes (unverified)	x	From	x
Llama 3.2 Vision 11B/90B	Yes	Yes	x	Yes (unverified)	x	From	x
LLaVA Next (LLaVA 1.6)	Yes	Yes	x	Yes (unverified)	x	From	x

Embedding Models

Embedding Models

Embedding Language Models	Pretraining & SFT	PEFT	Alignment	FP8 Training Convergence	TRT/TRTLLM	Convert To & From Hugging Face	Evaluation
SBERT 340M	Yes	x	x	Yes (unverified)	x	Both	x
Llama 3.2 Embedding 1B	Yes	x	x	Yes (unverified)	x	Both	x

World Foundation Models

World Foundation Models

World Foundation Models	Post-Training	Accelerated Inference
Cosmos-1.0-Diffusion-Text2World-7B	Yes	Yes
Cosmos-1.0-Diffusion-Text2World-14B	Yes	Yes
Cosmos-1.0-Diffusion-Video2World-7B	Coming Soon	Coming Soon
Cosmos-1.0-Diffusion-Video2World-14B	Coming Soon	Coming Soon
Cosmos-1.0-Autoregressive-4B	Yes	Yes
Cosmos-1.0-Autoregressive-Video2World-5B	Coming Soon	Coming Soon
Cosmos-1.0-Autoregressive-12B	Yes	Yes
Cosmos-1.0-Autoregressive-Video2World-13B	Coming Soon	Coming Soon

Note
NeMo also supports pretraining for both diffusion and autoregressive architectures `text2world` foundation models.

Speech AI

Developing conversational AI models is a complex process that involves defining, constructing, and training models within particular domains. This process typically requires several iterations to reach a high level of accuracy. It often involves multiple iterations to achieve high accuracy, fine-tuning on various tasks and domain-specific data, ensuring training performance, and preparing models for inference deployment.

_images/nemo-speech-ai.png

NeMo Framework provides support for the training and customization of Speech AI models. This includes tasks like Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) synthesis. It offers a smooth transition to enterprise-level production deployment with NVIDIA Riva. To assist developers and researchers, NeMo Framework includes state-of-the-art pre-trained checkpoints, tools for reproducible speech data processing, and features for interactive exploration and analysis of speech datasets. The components of the NeMo Framework for Speech AI are as follows:

Training and Customization
NeMo Framework contains everything needed to train and customize speech models (ASR, Speech Classification, Speaker Recognition, Speaker Diarization, and TTS) in a reproducible manner.

SOTA Pre-trained Models

- NeMo Framework provides state-of-the-art recipes and pre-trained checkpoints of several ASR and TTS models, as well as instructions on how to load them.
- Speech Tools
- NeMo Framework provides a set of tools useful for developing ASR and TTS models, including:
  - NeMo Forced Aligner (NFA) for generating token-, word- and segment-level timestamps of speech in audio using NeMo's CTC-based Automatic Speech Recognition models.
  - Speech Data Processor (SDP), a toolkit for simplifying speech data processing. It allows you to represent data processing operations in a config file, minimizing boilerplate code and allowing reproducibility and shareability.
  - Speech Data Explorer (SDE), a Dash-based web application for interactive exploration and analysis of speech datasets.
  - Dataset creation tool which provides functionality to align long audio files with the corresponding transcripts and split them into shorter fragments that are suitable for Automatic Speech Recognition (ASR) model training.
  - Comparison Tool for ASR Models to compare predictions of different ASR models at word accuracy and utterance level.
  - ASR Evaluator for evaluating the performance of ASR models and other features such as Voice Activity Detection.
  - Text Normalization Tool for converting text from the written form to the spoken form and vice versa (e.g. "31st" vs "thirty first").
- Path to Deployment
- NeMo models that have been trained or customized using the NeMo Framework can be optimized and deployed with NVIDIA Riva. Riva provides containers and Helm charts specifically designed to automate the steps for push-button deployment.

Other Resources

GitHub Repos

- **NeMo**: The main repository for the NeMo Framework
- **NeMo–Run**: A tool to configure, launch and manage your machine learning experiments.
- **NeMo-Aligner:** Scalable toolkit for efficient model alignment
- **NeMo-Curator:** Scalable data pre-processing and curation toolkit for LLMs

Getting Help
Engage with the NeMo community, ask questions, get support, or report bugs.

- NeMo Discussions
- NeMo Issues

Programming Languages and Frameworks

- **Python**: The main interface to use NeMo Framework
- **Pytorch**: NeMo Framework is built on top of PyTorch

Licenses

- NeMo Github repo is licensed under the Apache 2.0 license
- NeMo Framework is licensed under the NVIDIA AI PRODUCT AGREEMENT. By pulling and using the container, you accept the terms and conditions of this license.
- The NeMo Framework container contains Llama materials governed by the Meta Llama3 Community License Agreement.

Footnotes
Currently, NeMo Curator and NeMo Aligner support for Multimodal models is a work in progress and will be available very soon.

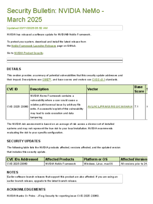
FAQ

Q: How can I check if my system is affected by the vulnerability?
A: You can check if your system is affected by verifying the version of the NVIDIA NeMo Framework installed. If it is below version 24, your system may be vulnerable.

Q: Who reported the security issue CVE-2025-23360?
A: The security issue was reported by Or Peles – JFrog Security. NVIDIA acknowledges their contribution.

Q: How can I receive future security bulletin notifications?
A: Visit the NVIDIA Product Security page to subscribe to security bulletin notifications and stay informed about product security updates.

Documents / Resources



Security Bulletin: NVIDIA NeMo - March 2025

Updated 03/11/2025 05:52 AM

NVIDIA has released a software update for NVIDIA® NeMo Framework.

To protect your system, download and install the latest release from the NeMo-Framework-Launcher Releases page on GitHub.

Go to NVIDIA Product Security.

DETAILS

This section provides a summary of potential vulnerabilities that this security update addresses and their impact. Descriptions use CWE™, and base scores and vectors use CVSS v3.1 standards.

CVE ID	Description	Vector	Base Score
CVE-2025-23360	NVIDIA Nemo Framework contains a vulnerability where a user could cause a relative path traversal issue by arbitrary file write. A successful exploit of this vulnerability may lead to code execution and data tampering.	AV:L/AC:L/PR:N/UI:R/S:U/C:N/I:H/A:H	7.1

The NVIDIA risk assessment is based on an average of risk across a diverse set of installed systems and may not represent the true risk to your local installation. NVIDIA recommends evaluating the risk to your specific configuration.

SECURITY UPDATES

The following table lists the NVIDIA products affected, versions affected, and the updated version that includes this security update.

CVE IDs Addressed	Affected Products	Platform or OS	Affected Version
CVE-2025-23360	NVIDIA NeMo Framework	Windows, Linux, macOS	All versions prior to 24.

NOTES

Earlier software branch releases that support this product are also affected. If you are using an earlier branch release, upgrade to the latest branch release.

ACKNOWLEDGEMENTS

NVIDIA thanks Or Peles - JFrog Security for reporting issue CVE-2025-23360.

NVIDIA NeMo Framework [pdf] User Guide
NeMo Framework, NeMo, Framework

References

- **User Manual**