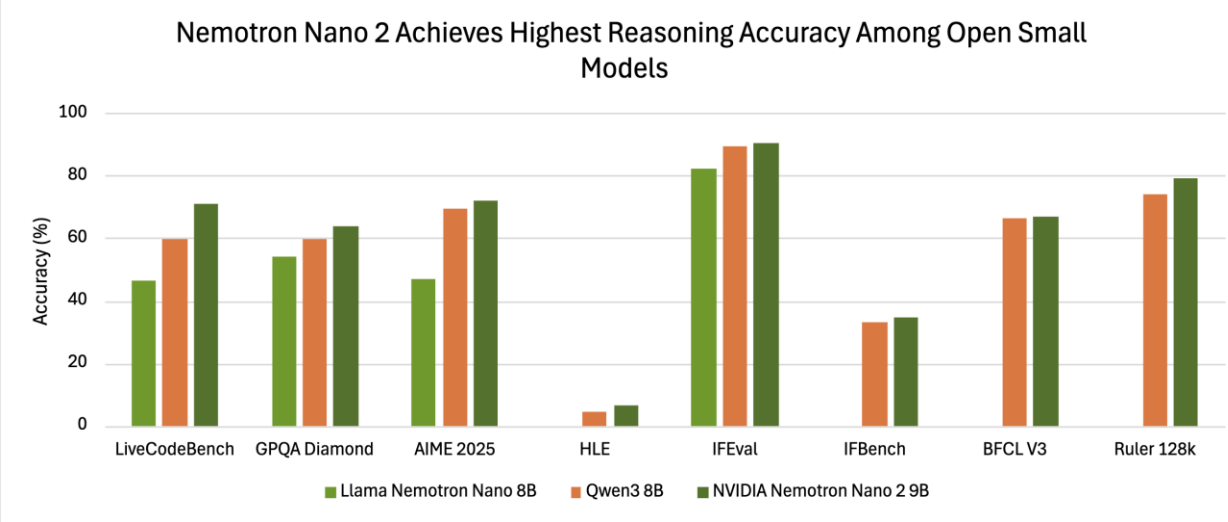
NVIDIA-Nemotron-Nano-9B-v2



Model Developer: NVIDIA Corporation

Model Dates:

June 2025 - August 2025

Data Freshness:

September 2024

The pretraining data has a cutoff date of September 2024.

Model Overview

NVIDIA-Nemotron-Nano-9B-v2 is a large language model (LLM) trained from scratch by NVIDIA, and designed as a unified model for both reasoning and non-reasoning tasks. It responds to user queries and tasks by first generating a reasoning trace and then concluding with a final response. The model's reasoning capabilities can be controlled via a system prompt. If the user prefers the model to provide its final answer without intermediate reasoning traces, it can be configured to do so, albeit with a slight decrease in accuracy for harder prompts that require reasoning. Conversely, allowing the model to

generate reasoning traces first generally results in higher-quality final solutions to queries and tasks.

The model uses a hybrid architecture consisting primarily of Mamba-2 and MLP layers combined with just four Attention layers. For the architecture, please refer to the Nemotron-H tech report. The model was trained using Megatron-LM and NeMo-RL.

The supported languages include: English, German, Spanish, French, Italian, and Japanese. Improved using Qwen.

This model is ready for commercial use.

Feature Voting

We want to hear from you! Share your ideas, vote on what matters, and help shape the future of Nemotron.

License/Terms of Use

Governing Terms: Use of this model is governed by the NVIDIA Open Model License Agreement.

Evaluation Results

Benchmark Results (Reasoning On)

We evaluated our model in **Reasoning-On** mode across all benchmarks, except RULER, which is evaluated in **Reasoning-Off** mode.

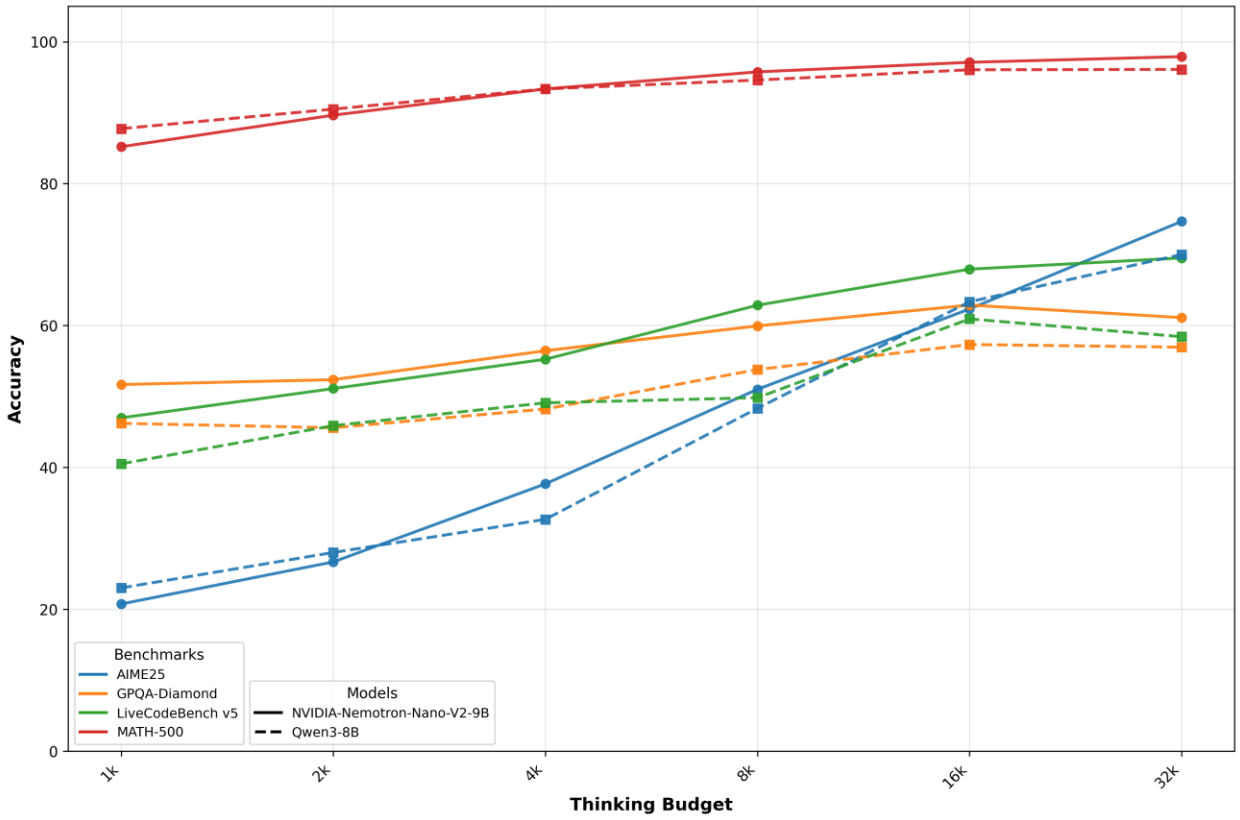
Benchmark	Qwen3-8B	NVIDIA-Nemotron-Nano-9B-v2
AIME25	69.3%	72.1%
MATH500	96.3%	97.8%

GPQA	59.6%	64.0%
LCB	59.5%	71.1%
BFCL v3	66.3%	66.9%
IFEval (Instruction Strict)	89.4%	90.3%
HLE	4.4%	6.5%
RULER (128K)	74.1%	78.9%

All evaluations were done using NeMo-Skills. We published a tutorial with all details necessary to reproduce our evaluation results.

Reasoning Budget Control

This model supports runtime “thinking” budget control. During inference, the user can specify how many tokens the model is allowed to "think".



Model Architecture

- Architecture Type: Mamba2-Transformer Hybrid
- Network Architecture: Nemotron-Hybrid

Deployment Geography: Global

Use Case

NVIDIA-Nemotron-Nano-9B-v2 is a general purpose reasoning and chat model intended to be used in English and coding languages. Other non-English languages (German, French, Italian, Spanish and Japanese) are also supported. Developers designing AI Agent systems, chatbots, RAG systems, and other AI-powered applications. Also suitable for typical instruction-following tasks.

Release Date: 08/18/2025

- Huggingface 08/18/2025 via https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-9B-v2
- API Catalog 08/18/2025 via https://build.nvidia.com/nvidia/nvidia-nemotron-nano-9b-v2

References

- NVIDIA Nemotron Nano 2: An Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model

Input

- Input Type(s): Text
- Input Format(s): String
- Input Parameters: One-Dimensional (1D): Sequences
- Other Properties Related to Input: Context length up to 128K. Supported languages include German, Spanish, French, Italian, Korean, Portuguese, Russian, Japanese, Chinese and English.

Output

- Output Type(s): Text
- Output Format: String
- Output Parameters: One-Dimensional (1D): Sequences up to 128K

Our models are designed and optimized to run on NVIDIA GPU-accelerated systems. By leveraging NVIDIA's hardware (e.g. GPU cores) and software frameworks (e.g., CUDA libraries), the model achieves faster training and inference times compared to CPU-only solutions.

Software Integration

- Runtime Engine(s): NeMo 25.07.nemotron-nano-v2
- Supported Hardware Microarchitecture Compatibility: NVIDIA A10G, NVIDIA H100-80GB, NVIDIA A100, Jetson AGX Thor
- Operating System(s): Linux