

**nvidia / GLM-5.2-NVFP4**

♡ like

283

Follow



NVIDIA

64k



Text Generation



Safetensors

Model Optimizer

glm_moe_dsa

nvidia

ModelOpt

GLM-5

quantized

4-bit precision

FP4

fp4

conversational



8-bit precision

modelopt



License: mit



→ Copy to bucket

NEW

**Model card**

Files



xet



Community

6

Downloads last month

1,625,761 **Safetensors** ⓘ

Model size

381B params

Tensor type

F32 · BF16 · F8_E4M3 · U8

{ Chat template

↗ Files info

Inference Providers NEW

Text Generation

This model isn't deployed by any Inference Provider.



Ask for provider support

Model tree for nvidia/GLM-5.2-NVFP4

Base model

zai-org/GLM-5.2

Quantized (119)

this model

Finetunes

2 models

Quantizations



4 models

Collection including nvidia/GLM-5.2-NVFP4

Inference Optimized Checkpoints (with Model Optimizer) Collection

A collection of generative models quantiz... • 84 items • Updated 4 days ago • △ 183

≡ [Model Overview](#)

[Description:](#)

The NVIDIA GLM-5.2 NVFP4 model is the quantized version of ZAI's GLM-5.2 model, which is an auto-regressive language model that uses an optimized transformer architecture. GLM-5.2 is a Mixture-of-Experts (MoE) model for reasoning and coding that uses sparse attention (with an IndexShare indexer) to support a long context. For more information, please check [here](#). The NVIDIA GLM-5.2 NVFP4 model is quantized with [Model Optimizer](#).

This model is ready for commercial or non-commercial use.

[License/Terms of Use:](#)

GOVERNING TERMS: Use of the model is governed by the [MIT License](#), same as the base model.

[Deployment Geography:](#)

Global

[Use Case:](#)

Developers looking to take off-the-shelf, pre-quantized models for deployment in AI Agent systems, chatbots, RAG systems, and other AI-powered applications.

[Release Date:](#)

Hugging Face 06/25/2026 via <https://huggingface.co/nvidia/GLM-5.2-NVFP4>

References

Nvidia Model Optimizer: <https://github.com/NVIDIA/Model-Optimizer>

Model Architecture:

Architecture Type: Transformers

Network Architecture: GLM-5.2 (GLMoeDsaForCausalLM)

Number of Model Parameters: 753B in total and 40B activated

Input:

Input Type(s): Text

Input Format(s): String

Input Parameters: One-Dimensional (1D)

Other Properties Related to Input: Context length up to 1M

Output:

Output Type(s): Text

Output Format: String

Output Parameters: 1D (One-Dimensional): Sequences

Other Properties Related to Output: None

Our AI models are designed and/or 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:

Supported Runtime Engine(s):

- SGLang

- vLLM

Supported Hardware Microarchitecture Compatibility:

- NVIDIA Blackwell

Preferred Operating System(s):

- Linux

The integration of foundation and fine-tuned models into AI systems requires additional testing using use-case-specific data to ensure safe and effective deployment. Following the V-model methodology, iterative testing and validation at both unit and system levels are essential to mitigate risks, meet technical and functional requirements, and ensure compliance with safety and ethical standards before deployment.

Model Version(s):

The model version is NVFP4 1.0 version and is quantized with nvidia-modelopt **v0.46.0**

Training, Testing, and Evaluation Datasets:

We calibrated the model using the dataset noted below, and performed evaluation using the benchmarks noted under Evaluation Datasets. We did not perform training or testing for this Model Optimizer release. The methods noted under Training and Testing Datasets below represent the data collection and labeling methods used by the third-party to train and test the underlying model.

Training Dataset:

Data Modality: Undisclosed

Data Collection Method by dataset: Undisclosed

Labeling Method by dataset: Undisclosed

Properties: Undisclosed

🔗 Testing Dataset:

Data Collection Method by dataset: Undisclosed

Labeling Method by dataset: Undisclosed

Properties: Undisclosed

🔗 Evaluation Dataset:

Datasets: GPQA Diamond, SciCode, IFBench, AA-LCR, τ^2 -Bench Telecom

Data Collection Method by dataset: Hybrid: Automated, Human

Labeling Method by dataset: Hybrid: Human, Automated

Properties: We evaluated the model on text-based reasoning, coding, long-context recall, and agentic tool-use benchmarks: GPQA Diamond contains 448 graduate-level multiple-choice questions written by domain experts in biology, physics, and chemistry; SciCode evaluates scientific coding capabilities; IFBench is a benchmark for evaluating instruction-following capabilities across diverse and structured task constraints; AA-LCR (Artificial Analysis Long Context Recall) evaluates a model's ability to accurately retrieve and recall information from long input contexts; τ^2 -Bench Telecom evaluates agentic tool-use and policy-adherence capabilities in dual-control telecom customer-service scenarios where the model interacts with a simulated user and external tools to resolve account issues.

🔗 Inference:

Acceleration Engine: SGLang, vLLM

Test Hardware: NVIDIA B200

NVIDIA B300

🔗 Post Training Quantization

This model was obtained by quantizing the weights and activations of GLM-5.2 to NVFP4 data type, ready for inference with SGLang and vLLM. Only the weights and activations of the linear operators within transformer blocks in MoE experts are quantized. The shared expert is not quantized.

[🔗](#) Usage

[🔗](#) SGLang

This checkpoint was served with the latest SGLang image (`lmsysorg/sglang:latest`). GLM-5.2's `glm_moe_dsa` architecture requires `transformers>=5.3.0`, which we installed in the container before launching the server. You can also use `lmsysorg/sglang:dev-glm52-nvfp4`

```
pip install -U "transformers>=5.3.0" && \
python3 -m sglang.launch_server \
  --model nvidia/GLM-5.2-NVFP4 \
  --tensor-parallel-size 8 \
  --quantization modelopt_fp4 \
  --tool-call-parser glm47 \
  --reasoning-parser glm45 \
  --trust-remote-code \
  --chunked-prefill-size 16384 \
  --mem-fraction-static 0.80
```

[🔗](#) vLLM

To serve this checkpoint with [vLLM](#), use the `vllm/vllm-openai:v0.23.0` image and run:

```
vllm serve nvidia/GLM-5.2-NVFP4 \
  --tensor-parallel-size 8 \
  --enable-expert-parallel \
  --trust-remote-code \
  --reasoning-parser glm45 \
  --tool-call-parser glm47 \
  --enable-auto-tool-choice \
  --kv-cache-dtype fp8_e4m3 \
  --host 0.0.0.0 --port 8000
```

Evaluation

The accuracy benchmark results are presented in the table below. AA-LCR was measured with SGLang; all other benchmarks were measured with vLLM.

Precision	GPQA Diamond	SciCode	IFBench	AA-LCR	τ^2 -Bench Telecom
baseline (FP8)	89.52	49.85	74.95	69.38	97.9
NVFP4	89.39	49.04	75.81	70.13	98.25

“Baseline: GLM-5.2-FP8. Benchmarked with temperature=1.0, top_p=0.95. GPQA Diamond used max_new_tokens=100000; all other benchmarks used max_new_tokens=64000.”

Model Limitations:

The base model was trained on data that contains toxic language and societal biases originally crawled from the internet. Therefore, the model may amplify those biases and return toxic responses especially when prompted with toxic prompts. The model may generate answers that may be inaccurate, omit key information, or include irrelevant or redundant text producing socially unacceptable or undesirable text, even if the prompt itself does not include anything explicitly offensive.

Ethical Considerations

NVIDIA believes Trustworthy AI is a shared responsibility and we have established policies and practices to enable development for a wide array of AI applications. Developers should work with their internal model team to ensure this model meets requirements for the relevant industry and use case and addresses unforeseen product misuse.

Please make sure you have proper rights and permissions for all input image and video content; if image or video includes people, personal health information, or intellectual

property, the image or video generated will not blur or maintain proportions of image subjects included.

For more detailed information on ethical considerations for this model, please see the [Model Card++ Bias, Explainability, Safety & Security, and Privacy Subcards](#).

Please report model quality, risk, security vulnerabilities or NVIDIA AI Concerns [here](#).

SUBCARDS:

Explainability

Field:	Response:
Intended Task/Domain:	Text generation, reasoning, summarization, and question answering.
Model Type:	Text and Image-to-text transformer
Intended Users:	This model is intended for developers, researchers, and customers building/utilizing LLMs, while balancing accuracy and efficiency.
Output:	Text String(s)
Describe how the model works:	Generates text by predicting the next word or token based on the context provided in the input sequence using multiple self-attention layers
Name the adversely impacted groups this has been tested to deliver comparable outcomes regardless of:	Not Applicable
Technical Limitations & Mitigation:	The model was trained on data that contains toxic language and societal biases originally crawled from the internet. Therefore, the model may amplify those biases and return toxic responses especially when prompted with toxic prompts. Therefore, before deploying any applications of this model, developers should

Field:	Response:
	perform safety testing and tuning tailored to their specific applications of the model.
Verified to have met prescribed quality standards?	Yes
Performance Metrics:	Accuracy, Throughput, and user-side throughput
Potential Known Risk	The model may generate answers that may be inaccurate, omit key information, or include irrelevant or redundant text producing socially unacceptable or undesirable text, even if the prompt itself does not include anything explicitly offensive.
Licensing:	Your usage is governed by the following GOVERNING TERMS: Use of the model is governed by the MIT License , same as the base model.

Bias

Field:	Response:
Participation considerations from adversely impacted groups <u>protected classes</u> in model design and testing:	None
Measures taken to mitigate against unwanted bias:	None

Safety & Security

Field:	Response:
Model Application(s):	Chat, Instruction Following, Chatbot Development, Code Generation, Reasoning
Describe life critical application (if	Not Applicable

Field:	Response:
present):	
Use Case Restrictions:	Abide by the GOVERNING TERMS: Use of the model is governed by the <u>MIT License</u> , same as the base model.
Model and Dataset Restrictions:	The Principle of least privilege (PoLP) is applied limiting access for dataset generation. Restrictions enforce dataset access during training, and dataset license constraints adhered to. Model checkpoints are made available on Hugging Face, and may become available on cloud providers' model catalog.

Privacy

Field:	Response:
Generatable or Reverse engineerable personal data?	No
Personal data used to create this model?	No
Was consent obtained for any personal data used?	Not Applicable
How often is dataset reviewed?	Before Release
Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	No
Is there provenance for all datasets used in training?	Yes
Does data labeling (annotation, metadata) comply with privacy laws?	Yes
Is data compliant with data subject requests for data correction or removal, if such a request was made?	Not Applicable

Field:	Response:
Applicable NVIDIA Privacy Policy	https://www.nvidia.com/en-us/about-nvidia/privacy-policy/

 System theme

Company

TOS

Privacy

About

Careers

Website

Models

Datasets

Spaces

Pricing

Docs

