Nemotron-4-4B-Instruct

Nemotron-4 4B Instruct has been optimized for on-device inference. This model provides better roleplay, retrieval augmented generation (RAG), and function calling capabilities.

How to use Nemotron-4-4B-Instruct for NVIDA In-Game Inferencing (IGI) SDK

Please ensure that you have downloaded the NVIDIA IGI SDK and confirm that it runs on your environment.

Installing the Model into the IGI SDK

Double-click the `download.bat` file located under:

<IGI Root>/nvigi.models/nvigi.plugin.gpt.ggml/ {8E31808B-C182-4016-9ED8-64804FF5B40D}

Which uses `curl` to download and unpack the model in the appropriate place.

Alternatively, you can manually download the model and unzip the provided pack to a local drive; in it will be a directory that contains the model and model information needed by IGI. These model directories are named for the GUID that represents that model. In this case, {8E31808B-C182-4016-9ED8-64804FF5B40D}.

IGI stores its models in a hierarchical directory structure by plugin. In the IGI SDK pack, this directory is rooted at nvigi.models. Under this models directory will be the names of several plugins; the plugin that supports Nemotron-Mini-4B-Instruct is named nvigi.plugin.gpt.ggml, the GGML-based GPT plugin. Under that directory, place the {8E31808B-C182-4016-9ED8-64804FF5B40D} directory that was unzipped from the Nemotron-Mini-4B-Instruct pack. giving a structure of the form:

<IGI Root>/nvigi.models/nvigi.plugin.gpt.ggml/ {8E31808B-C182-4016-9ED8-64804FF5B40D}

of course, inside of {8E31808B-C182-4016-9ED8-64804FF5B40D} will be numerous files, including the all-important GGUF model file itself.

Using the Nemotron-Mini-4B-Instruct Model with EA2

With the model copied into the standard model tree in the EA2 pack, there are several methods of using the Nemotron-Mini-4B-Instruct model. These include the 3D sample and the command-line sample.

The 3D Sample

The easiest way is to use the 3D sample, which is found under /sample. This sample will automatically detect all available models for the supported plugins, so the Nemotron-Mini-4B-Instruct model should appear in the GPT plugin pulldown in the sample's UI as something like "ggml.cuda : nemotron-4-mini-4b-instruct_q4_0". Select this in the UI and you will be able to use the Nemotron-Mini-4B-Instruct model as you would any other GPT model as per the sample's README file.

The 3D sample also includes an updated prompt template in its code which matches the popular "Standard Chat Template" that works well with minitron and may be found in the sample's src/nvigi/NVIGIContext.cpp as fullPrompt.

The Command-line Basic Sample

The command-line sample is in the IGI SDK itself; it is built via the normal sample build methods as described in the SDK's README file, and is run per the instructions there. The source to the sample may be found in /sdk/source/samples/nvigi.basic.

Unlike the more complex 3D sample, the command-line sample sets the model to be loaded in code. So to try Nemotron-Mini-4B-Instruct in the command-line sample, a trivial code change is needed. In EA2 the file to be changed is /sdk/source/samples/nvigi.basic/basic.cpp. The line to be changed is where the GPT model GUID is set in instance creation. In EA2, this line looks like:

gptParams.common->modelGUID = "{D5E8DEB3-28C2-4B9E-9412-B9A012B23584}";

which loads a different model. To load Nemotron-Mini-4B-Instruct, simply change that GUID to the one for the Nemotron-Mini-4B-Instruct model we installed:

gptParams.common->modelGUID = "{8E31808B-C182-4016-9ED8-64804FF5B40D}";

With this change, rebuild the Basic sample as per the documentation, and the Nemotron-Mini-4B-Instruct model will run in the sample.

Note that the sample in EA2 also includes an updated prompt template that better matches Minitron, which may be found in basic.cpp as as fullPrompt.

Advanced Nemotron-Mini-4B-Instruct Topics

Context Size

The amount of context memory used when interacting with Nemotron-Mini-4B-Instruct can affect quality of the results, especially when using large input text for background information. This amount can be changed in both the 3D sample and the command-line sample by looking for instances of nvigi::GPTCreationParameters. These will tend to set the context size to:

```
params1->contextSize = 512;
```

Or similar. This can be increased to larger values, e.g.

```
params1->contextSize = 4096;
```

However, increasing this value can lead to the model requiring significantly more VRAM to run, possibly several GB more. So this change should be considered an advanced use-case, and the size-quality tradeoffs investigated with real experiments.

RAG Templates

The templating of prompts in the samples are basic "assistant-style" templates. If an application wishes to use RAG-style prompts, then a different template should be used. These can be edited in the two samples (NVIGIContext.cpp in the 3D sample, and basic.cpp in the basic sample) by searching the code for the nvigi::CpuData that is created to use as the nvigi::kGPTDataSlotUser slot for the GPT plugin. In the EA2 samples, this is of the form:

```
std::string fullPrompt =

"<extra_id_0>System\nAnswer this question in a helpful manner\n<extra_id_1>User\n
"

+ prompt

+ "\n<extra_id_1>Assistant\n";
```

Thus forming a Prefix, the user query and a Suffix. In the case of RAG or tool usage, the prompt template should be some form of:

```
<extra_id_0>System

{system prompt}

<tool> ... </tool>

<context> ... </context>

<extra_id_1>User

{prompt}

<extra_id_1>Assistant
```

```
<toolcall> ... </toolcall>

<extra_id_1>Tool

{tool response}

<extra_id_1>Assistant\n
```

Model License

NVIDIA Software and Model Evaluation License included. Model may be freely used for evaluation and pre-commercialization. Please contact your NVIDIA Developer Relations representative to obtain a commercial license to this AI Inference Manager plugin and model.

Model Card Overview

Description:

Nemotron-4 4B Instruct is a model for generating responses for roleplaying, retrieval augmented generation, and function calling. It is a small language model (SLM) optimized through distillation, pruning and quantization for speed and on-device deployment. VRAM usage has been minimized to approximately 2 GB, providing significantly faster time to first token compared to LLMs.

This model is ready for commercial use.

License/Terms of Use:

NVIDIA AI Foundation Models Community License Agreement

References

Please refer to the User Guide to use the model and use a suggested guideline for prompts.

Model Architecture:

Architecture Type: Transformer
Network Architecture: Decoder-only

Limitations

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. 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. This issue could be exacerbated without the use of the recommended prompt template. This issue could be exacerbated without the use of the recommended prompt template.

Input:

Input Type(s): Text (Prompt)
Input Format(s): String
Input Parameters: One Dimensional (1D)
Other Properties Related to Input: The model has a maximum of 4096 input tokens.

Output:

Output Type(s): Text (Response)
Output Format: String
Output Parameters: 1D
Other Properties Related to Output: The model has a maximum of 4096 input tokens. Maximum output for both versions can be set apart from input.

Prompt Format:

We recommend using the following prompt template, which was used to fine-tune the model. The model may not perform optimally without it.

Single Turn

```
<extra_id_0>System
{system prompt}

<extra_id_1>User
{prompt}
<extra_id_1>Assistant\n
```

Tool use

```
<extra_id_0>System
{system prompt}

<tool> ... </tool>
<context> ... </context>

<extra_id_1>User
{prompt}
<extra_id_1>Assistant
<toolcall> ... </toolcall>
<extra_id_1>Tool
{tool response}
<extra_id_1>Assistant\n
```

Software Integration: (On-Device)

Runtime(s): At least In-Game Inferencing (NVIGI) Version 1.0.0
Toolkit: NVIGI
See this document for details on how to integrate the model into NVIGI.
Supported Hardware Platform(s): GPU supporting DirectX 11/12 and Vulkan 1.2 or higher
[Preferred/Supported] Operating System(s):

- Windows

Software Integration: (Cloud)

Toolkit: NVIDIA NIM
See this document for details on how to integrate the model into NVIGI.
[Preferred/Supported] Operating System(s):

- Linux

Model Version(s)

Nemotron-4-4B-instruct

Training & Evaluation:

Training Dataset:

** Data Collection Method by dataset

- Hybrid: Automated, Human

** Labeling Method by dataset

- Hybrid: Automated, Human

Properties (Quantity, Dataset Descriptions, Sensor(s)):
Trained on approximately 10000 Game/Non-Playable Character (NPC) dialog turns from domain chat data.

Evaluation Dataset:

** Data Collection Method by dataset

- Hybrid: Automated, Human

** Labeling Method by dataset

- Human

Properties (Quantity, Dataset Descriptions, Sensor(s)):
Evaluated on approximately Game/NPC 1000 dialog turns from domain chat data.

Inference:

Engine: TRT-LLM
Test Hardware [Name the specific test hardware model]:

- A100
- A10g
- H100
- L40s

Supported Hardware Platform(s): L40s, A10g, A100, H100

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. When downloaded or used in accordance with our terms of service, 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. For more detailed information on ethical considerations for this model, please see the Model Card++ Explainability, Bias, Safety & Security, and Privacy Subcards. Please report security vulnerabilities or NVIDIA AI Concerns here.

Model Card ++ Safety

Field	Response
Model Application(s):	Non-Playable Character Conversation
Describe the life-critical impact (if present).	None Known
Use Case Restrictions:	Abide byNVIDIA AI Foundation Models Community License Agreement
Model and dataset restrictions:	The Principle of least privilege (PoLP) is applied limiting access for dataset generation and model

Field	Response
	development. Restrictions enforce dataset access during training, and dataset license constraints adhered to.

Model Card ++ Privacy

Field	Response
Generatable or reverse engineerable personally-identifiable information (PII)?	None
Was consent obtained for any personal data used?	Not Applicable
Personal data used to create this model?	Datasets used for fine-tuning did not introduce any personal data that did not exist in the base model.
How often is dataset reviewed?	Before Release
Is a mechanism in place to honor data subject right of access or deletion of personal data?	Not Applicable
If personal data is collected for the development of the model, was it collected directly by NVIDIA?	Not Applicable
If personal data is collected for the development of the model by NVIDIA, do	Not Applicable

Field	Response
you maintain or have access to disclosures made to data subjects?	
If personal data is collected for the development of this AI model, was it minimized to only what was required?	Not Applicable
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

Model Card ++ Explainability

Field	Response
Intended Application & Domain:	Game Non-Playable Character (NPC) Development
Model Type:	Generative Pre-Trained Transformer (GPT)
Intended User:	Enterprise developers building game NPCs.
Output:	Text String(s)
Describe how the model works:	Generates a response using the input text and context such as NPC background information.

Field	Response
Name the adversely impacted groups this has been tested to deliver comparable outcomes regardless of:	Not Applicable
Verified to have met prescribed NVIDIA quality standards:	Yes
Performance Metrics:	Accuracy, Latency, and Throughput
Potential Known Risks:	This model may produce output that is biased and toxic based on how it is prompted, producing socially unacceptable or undesirable text, even if the prompt itself does not include anything explicitly offensive. The model may also amplify biases and return toxic responses especially when prompted with toxic prompts.
Licensing:	NVIDIA AI Foundation Models Community License Agreement

Model Card ++ Bias

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