Model Overview

Description:

The Meta Llama 3.2 collection of multilingual large language models (LLMs) is a collection of pre-trained and instruction-tuned generative models in 1B and 3B sizes (text in/text out). The Llama 3.2 instruction-tuned text only models are optimized for multilingual dialogue use cases, including agentic retrieval and summarization tasks. They outperform many of the available open source and closed chat models on common industry benchmarks.

Llama 3.2 models are ready for commercial use.

Model Developer: Meta

Third-Party Community Consideration

This model is not owned or developed by NVIDIA. This model has been developed and built to a third-party's requirements for this application and use case; see Llama 3.2 Model Card.

License/Terms of Use:

Use of Llama 3.2 is governed by the Llama 3.2 Community License (a custom, commercial license agreement).

Reference(s):

Llama 3.2 Github

Model Architecture:

- **Architecture Type:** Transformer

- ** Network Architecture:** Llama 3.2

Input:

- **Input Type(s):** Text
- **Input Format(s):** String
- ** Input Parameters:** 1D
- *_Other Properties Related to Input:_* Official Supported languages are English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. Llama 3.2 has been trained on a broader collection of languages than these 8 supported languages.

Output:

- **Output Type(s):** Text
- **Output Format:** String
- **Output Parameters:**1D
- **Other Properties Related to Output:** Official Supported languages are English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.

Supported Hardware Microarchitecture Compatibility:

- NVIDIA Ada

Supported Operating System(s):

- Windows

Model Version(s):

Llama 3.2 3B 1.0

Training Datasets:

Training Dataset:

- **Data Collection Method by dataset:** Unknown

- **Labeling Method by dataset** : Unknown
- **Properties:** Llama 3.2 was pre-trained on up to 9 trillion tokens of data from publicly available sources. For the 1B and 3B Llama 3.2 models, we incorporated logits from the Llama 3.1 8B and 70B models into the pre-training stage of the model development, where outputs (logits) from these larger models were used as token-level targets. Knowledge distillation was used after pruning to recover performance. In post-training we used a similar recipe as Llama 3.1 and produced final chat models by doing several rounds of alignment on top of the pre-trained model. Each round involved Supervised Fine-Tuning (SFT), Rejection Sampling (RS), and Direct Preference Optimization (DPO). **Data Freshness:** The pre-training data has a cutoff of December 2023.

Inference:

Engine: GGUF **Test Hardware :** RTX 4090

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.

Please report security vulnerabilities or NVIDIA AI Concerns here.