Gmail

Geraldo Checon <gchecon@gmail.com>

Get started with Llama 3.1

1 mensagem

AI at Meta <noreply@email.meta.com> 21 de outubro de 2024 às 11:25
Responder a: AI at Meta <noreply@email.meta.com>
Para: gchecon@gmail.com

You're all set to start building with Llama 3.1, Llama Guard 3 and Prompt Guard

The models listed below are now available to you as a commercial license holder. By downloading a model, you are agreeing to the terms and conditions of the License and Acceptable Use Policy and Meta's privacy policy.

MODELS AVAILABLE

With each model size, please find:

1. Pretrained weights: These are base weights that can be finetuned, domain adapted with full flexibility
2. Instruct weights: These weights are for the model that have been fine-tuned and aligned to follow instructions. They can be used as-is in chat applications or further finetuned and aligned for specific use cases

Pretrained:

- Meta-Llama-3.1-8B
- Meta-Llama-3.1-70B
- Meta-Llama-3.1-405B
- Meta-Llama-3.1-405B-MP16
- Meta-Llama-3.1-405B-FP8

Fine-tuned:

- Meta-Llama-3.1-8B-Instruct
- Meta-Llama-3.1-70B-Instruct
- Meta-Llama-3.1-405B-Instruct
- Meta-Llama-3.1-405B-Instruct-MP16
- Meta-Llama-3.1-405B-Instruct-FP8

Llama-Guard-3-8B

Llama-Guard-3-8B-INT8

Prompt-Guard-86M

NOTE 405B:

- Model requires significant storage and computational resources, occupying approximately 750GB of disk storage space and necessitating two nodes on MP16 for inferencing.
- We are releasing multiple versions of the 405B model to accommodate its large size and facilitate multiple deployment options: MP16 (Model Parallel 16) is the full version of BF16 weights. These weights can only be served on multiple nodes using pipelined parallel inference. At minimum it would need 2 nodes of 8 GPUs to serve.

- MP8 (Model Parallel 8) is also the full version of BF16 weights, but can be served on a single node with 8 GPUs by using dynamic FP8 (floating point 8) quantization. We are providing reference code for it. You can download these weights and experiment with different quantization techniques outside of what we are providing.
- FP8 (Floating Point 8) is a quantized version of the weights. These weights can be served on a single node with 8 GPUs by using the static FP quantization. We have provided reference code for it as well.

HOW TO DOWNLOAD THE MODEL

1. Visit the Llama Models repository for the model on GitHub and follow the instructions in the README to choose and download the model using Llama CLI. Pass the custom URL below when prompted to start the download. (Clicking on the URL itself does not access the model):

   https://llama3-1.llamameta.net/*?Policy=eyJTdGF0ZW1lbnQiOlt7InVuaXF1ZV9oYXNoIjoia2toeDk5eXo5cjB2bnJyaGR3bGJhYXV3IiwiUmVzb3VyY2UiOiJodHRwczpcL1wvbGxhbWEzLTEubGxhbWFtZXRhLm5ldFwvKiIsIkNvbmRpdGlvbiI6eyJEYXRlTGVzc1RoYW4iOnsiQVdTOkVwb2NoVGltZSI6MTcyOTY5MzQ5OX19fV19&Signature=cp3Kze6ItgwFdB6bYQ03NflxOyvJclYvjKAubfIvS7Yy-MtIOT2FHSEtfp2xmJxoJEAGfyy02b6Ppz6Sgn%7E5oW5cdQR8SrpekocroUBxrwSkliqGPfwj2%7E5UI%7EiKxsLQ5jrS3MGgVaTlbJCikMQEeI%7EphhdRmOi6tMoQL1HD1-325UpFwTmc6lycgyWH1Z6Grq26savN6iL8wpdaEfIcz9B%7EC0i42Uo6OiLmOrHScctOaPxH7hbldJlu1GKnUWGI2FQK36Qu13hY2u%7E-38d-PCva79zo4qRDeZNrVU%7E1Zq1FUBZsznVhOTncPOefn2DjP64yhhYbN6HW5uelMu43NA__&Key-Pair-Id=K15QRJLYKIFSLZ&Download-Request-ID=3794225714151550

2. Specify models to download.

Please save copies of the unique custom URLs provided above, they will remain valid for 24 hours to download each model up to 5 times, and requests can be submitted multiple times. An email with the download instructions will also be sent to the email address you used to request the models.

Now you're ready to start building with Llama 3.1, Llama Guard 3 and Prompt Guard.

HELPFUL TIPS:

Please read the instructions in the GitHub repo and use the provided code examples to understand how to best interact with the models. In particular, for the fine-tuned models you must use appropriate formatting and correct system/instruction tokens to get the best results from the model.

You can find additional information about how to responsibly deploy Llama models in our Responsible Use Guide.

IF YOU NEED TO REPORT ISSUES:

If you or any Llama user becomes aware of any violation of our license or acceptable use policies —or any bug or issues with Llama that could lead to any such violations—please report it through one of the following means:

- Reporting issues with the model: https://github.com/meta-llama/llama-models/issues
- Giving feedback about potentially problematic output generated by the model: http://developers.facebook.com/llama_output_feedback
- Reporting bugs and security concerns: facebook.com/whitehat/info
- Reporting violations of the Acceptable Use Policy: LlamaUseReport@meta.com

Subscribe to get the latest updates on Llama 3.1 and AI at Meta.

Meta GenAI Team