You are all set!

Request ID: 432021479926556

Requested models:

- Llama 3.1 8B
- Llama 3.1 70B
- Llama 3.1 405B
- Llama Guard 3 8B
- 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, Acceptable Use Policy and Meta's privacy policy.

How to download the model

Visit the Llama repository in GitHub where instructions can be found in the Llama README

1 Install the Llama CLI

In your preferred environment run the command below

```
pip install llama-stack
```

2 Find models list

See latest available models by running the following command and determine the model ID you wish to download:

```
llama model list
```

If you want older versions of models, run the command below to show all the available Llama models:

```
llama model list --show-all
```

3 Select a model

Select a desired model by running:

```
llama model download --source meta --model-id  MODEL_ID
```

4 Specify custom URL

Llama 3.1: 405B, 70B & 8B

When the script asks for your unique custom URL, please paste the URL below

```
https://llama3-1.llamameta.net/*?Policy=eyJTdGF0ZW1lbnQiOl...
```

Please save copies of the unique custom URLs provided above, they will remain valid for **48 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.

Available models

Available models for download include:

- Pretrained:
  - Llama-3.1-8B
  - Llama-3.1-70B
  - Llama-3.1-405B-MP16
  - Llama-3.1-405B-FP8
- Fine-tuned:
  - Llama-3.1-8B-Instruct
  - Llama-3.1-70B-Instruct
  - Llama-3.1-405B-Instruct
  - Llama-3.1-405B-Instruct-MP16
  - Llama-3.1-405B-Instruct-FP8
- Llama Guard:
  - Llama-Guard-3-8B

  - Llama-Guard-3-8B-INT8
  - Llama-Guard-2-8B
  - Llama-Guard-8B
  - Prompt-Guard-86M

Note for 405B:

- 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.
- 405B model requires significant storage and computational resources, occupying approximately 750GB of disk storage space and necessitating two nodes on MP16 for inferencing.

Recommended tools

Code Shield

A system-level approach to safeguard tools, Code Shield adds support for inference-time filtering of insecure code produced by LLMs. This offers mitigation of insecure code suggestions risk, code interpreter abuse prevention, and secure command execution.
Now available on Github

Cybersecurity Eval

The first and most comprehensive set of open source cybersecurity safety evals for LLMs. These benchmarks are based on industry guidance and standards (e.g. CWE & MITRE ATT&CK) and built in collaboration with our security subject matter experts.
Now available on Github

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.

Review our Documentation to start building

Open Documentation

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
- Giving feedback about potentially problematic output generated by the model
- Reporting bugs and security concerns
- Reporting violations of the Acceptable Use Policy: LlamaUseReport@meta.com

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

Meta

Documentation

Overview

Models

Getting the Models

Running Llama

How-To Guides

Integration Guides

Community Support

Community ⌄

Resources ⌄

Trust & Safety ⌄