**Llama 2**

Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. This is the repository for the 7B pretrained model. Links to other models can be found in the index at the bottom.

**Model Details**

*Note: Use of this model is governed by the Meta license. In order to download the model weights and tokenizer, please visit the*[*website*](https://ai.meta.com/resources/models-and-libraries/llama-downloads/)*and accept our License before requesting access here.*

Meta developed and publicly released the Llama 2 family of large language models (LLMs), a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.

**Model Developers** Meta

**Variations** Llama 2 comes in a range of parameter sizes — 7B, 13B, and 70B — as well as pretrained and fine-tuned variations.

**Input** Models input text only.

**Output** Models generate text only.

**Model Architecture** Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety.

|  | **Training Data** | **Params** | **Content Length** | **GQA** | **Tokens** | **LR** |
| --- | --- | --- | --- | --- | --- | --- |
| Llama 2 | *A new mix of publicly available online data* | 7B | 4k | ✗ | 2.0T | 3.0 x 10-4 |
| Llama 2 | *A new mix of publicly available online data* | 13B | 4k | ✗ | 2.0T | 3.0 x 10-4 |
| Llama 2 | *A new mix of publicly available online data* | 70B | 4k | ✔ | 2.0T | 1.5 x 10-4 |

*Llama 2 family of models.* Token counts refer to pretraining data only. All models are trained with a global batch-size of 4M tokens. Bigger models - 70B -- use Grouped-Query Attention (GQA) for improved inference scalability.

**Model Dates** Llama 2 was trained between January 2023 and July 2023.

**Status** This is a static model trained on an offline dataset. Future versions of the tuned models will be released as we improve model safety with community feedback.

**License** A custom commercial license is available at: <https://ai.meta.com/resources/models-and-libraries/llama-downloads/>

**Research Paper** ["Llama-2: Open Foundation and Fine-tuned Chat Models"](https://huggingface.co/meta-llama/Llama-2-7b/blob/main/arxiv.org/abs/2307.09288)

**Intended Use**

**Intended Use Cases** Llama 2 is intended for commercial and research use in English. Tuned models are intended for assistant-like chat, whereas pretrained models can be adapted for a variety of natural language generation tasks.

To get the expected features and performance for the chat versions, a specific formatting needs to be followed, including the INST and <<SYS>> tags, BOS and EOS tokens, and the whitespaces and breaklines in between (we recommend calling strip() on inputs to avoid double-spaces). See our reference code in github for details: [chat\_completion](https://github.com/facebookresearch/llama/blob/main/llama/generation.py" \l "L212).

**Out-of-scope Uses** Use in any manner that violates applicable laws or regulations (including trade compliance laws).Use in languages other than English. Use in any other way that is prohibited by the Acceptable Use Policy and Licensing Agreement for Llama 2.

**Hardware and Software**

**Training Factors** We used custom training libraries, Meta's Research Super Cluster, and production clusters for pretraining. Fine-tuning, annotation, and evaluation were also performed on third-party cloud compute.

**Carbon Footprint** Pretraining utilized a cumulative 3.3M GPU hours of computation on hardware of type A100-80GB (TDP of 350-400W). Estimated total emissions were 539 tCO2eq, 100% of which were offset by Meta’s sustainability program.

|  | **Time (GPU hours)** | **Power Consumption (W)** | **Carbon Emitted(tCO2eq)** |
| --- | --- | --- | --- |
| Llama 2 7B | 184320 | 400 | 31.22 |
| Llama 2 13B | 368640 | 400 | 62.44 |
| Llama 2 70B | 1720320 | 400 | 291.42 |
| Total | 3311616 |  | 539.00 |

**CO2 emissions during pretraining.** Time: total GPU time required for training each model. Power Consumption: peak power capacity per GPU device for the GPUs used adjusted for power usage efficiency. 100% of the emissions are directly offset by Meta's sustainability program, and because we are openly releasing these models, the pretraining costs do not need to be incurred by others.

**Training Data**

**Overview** Llama 2 was pretrained on 2 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over one million new human-annotated examples. Neither the pretraining nor the fine-tuning datasets include Meta user data.

**Data Freshness** The pretraining data has a cutoff of September 2022, but some tuning data is more recent, up to July 2023.

**Evaluation Results**

In this section, we report the results for the Llama 1 and Llama 2 models on standard academic benchmarks.For all the evaluations, we use our internal evaluations library.

| **Model** | **Size** | **Code** | **Commonsense Reasoning** | **World Knowledge** | **Reading Comprehension** | **Math** | **MMLU** | **BBH** | **AGI Eval** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Llama 1 | 7B | 14.1 | 60.8 | 46.2 | 58.5 | 6.95 | 35.1 | 30.3 | 23.9 |
| Llama 1 | 13B | 18.9 | 66.1 | 52.6 | 62.3 | 10.9 | 46.9 | 37.0 | 33.9 |
| Llama 1 | 33B | 26.0 | 70.0 | 58.4 | 67.6 | 21.4 | 57.8 | 39.8 | 41.7 |
| Llama 1 | 65B | 30.7 | 70.7 | 60.5 | 68.6 | 30.8 | 63.4 | 43.5 | 47.6 |
| Llama 2 | 7B | 16.8 | 63.9 | 48.9 | 61.3 | 14.6 | 45.3 | 32.6 | 29.3 |
| Llama 2 | 13B | 24.5 | 66.9 | 55.4 | 65.8 | 28.7 | 54.8 | 39.4 | 39.1 |
| Llama 2 | 70B | **37.5** | **71.9** | **63.6** | **69.4** | **35.2** | **68.9** | **51.2** | **54.2** |

**Overall performance on grouped academic benchmarks.** *Code:* We report the average pass@1 scores of our models on HumanEval and MBPP. *Commonsense Reasoning:* We report the average of PIQA, SIQA, HellaSwag, WinoGrande, ARC easy and challenge, OpenBookQA, and CommonsenseQA. We report 7-shot results for CommonSenseQA and 0-shot results for all other benchmarks. *World Knowledge:* We evaluate the 5-shot performance on NaturalQuestions and TriviaQA and report the average. *Reading Comprehension:* For reading comprehension, we report the 0-shot average on SQuAD, QuAC, and BoolQ. *MATH:* We report the average of the GSM8K (8 shot) and MATH (4 shot) benchmarks at top 1.

|  |  | **TruthfulQA** | **Toxigen** |
| --- | --- | --- | --- |
| Llama 1 | 7B | 27.42 | 23.00 |
| Llama 1 | 13B | 41.74 | 23.08 |
| Llama 1 | 33B | 44.19 | 22.57 |
| Llama 1 | 65B | 48.71 | 21.77 |
| Llama 2 | 7B | 33.29 | **21.25** |
| Llama 2 | 13B | 41.86 | 26.10 |
| Llama 2 | 70B | **50.18** | 24.60 |

**Evaluation of pretrained LLMs on automatic safety benchmarks.** For TruthfulQA, we present the percentage of generations that are both truthful and informative (the higher the better). For ToxiGen, we present the percentage of toxic generations (the smaller the better).

|  |  | **TruthfulQA** | **Toxigen** |
| --- | --- | --- | --- |
| Llama-2-Chat | 7B | 57.04 | **0.00** |
| Llama-2-Chat | 13B | 62.18 | **0.00** |
| Llama-2-Chat | 70B | **64.14** | 0.01 |

**Evaluation of fine-tuned LLMs on different safety datasets.** Same metric definitions as above.

**Ethical Considerations and Limitations**

Llama 2 is a new technology that carries risks with use. Testing conducted to date has been in English, and has not covered, nor could it cover all scenarios. For these reasons, as with all LLMs, Llama 2’s potential outputs cannot be predicted in advance, and the model may in some instances produce inaccurate, biased or other objectionable responses to user prompts. Therefore, before deploying any applications of Llama 2, developers should perform safety testing and tuning tailored to their specific applications of the model.

Please see the Responsible Use Guide available at [https://ai.meta.com/llama/responsible-use-guide/](https://ai.meta.com/llama/responsible-use-guide)

**Reporting Issues**

Please report any software “bug,” or other problems with the models through one of the following means:

* Reporting issues with the model: [github.com/facebookresearch/llama](http://github.com/facebookresearch/llama)
* Reporting problematic content generated by the model: [developers.facebook.com/llama\_output\_feedback](http://developers.facebook.com/llama_output_feedback)
* Reporting bugs and security concerns: [facebook.com/whitehat/info](http://facebook.com/whitehat/info)

**Llama Model Index**

| **Model** | **Llama2** | **Llama2-hf** | **Llama2-chat** | **Llama2-chat-hf** |
| --- | --- | --- | --- | --- |
| 7B | [Link](https://huggingface.co/meta-llama/Llama-2-7b) | [Link](https://huggingface.co/meta-llama/Llama-2-7b-hf) | [Link](https://huggingface.co/meta-llama/Llama-2-7b-chat) | [Link](https://huggingface.co/meta-llama/Llama-2-7b-chat-hf) |
| 13B | [Link](https://huggingface.co/meta-llama/Llama-2-13b) | [Link](https://huggingface.co/meta-llama/Llama-2-13b-hf) | [Link](https://huggingface.co/meta-llama/Llama-2-13b-chat) | [Link](https://huggingface.co/meta-llama/Llama-2-13b-chat-hf) |
| 70B | [Link](https://huggingface.co/meta-llama/Llama-2-70b) | [Link](https://huggingface.co/meta-llama/Llama-2-70b-hf) | [Link](https://huggingface.co/meta-llama/Llama-2-70b-chat) | [Link](https://huggingface.co/meta-llama/Llama-2-70b-chat-hf) |