Search models, datasets, users...

TheBloke/Llama-2-7B-Chat-GGML

like 336

Text Generation | Transformers | PyTorch | English | llama | facebook | meta | llama-2 | text-generation-inference | License: other

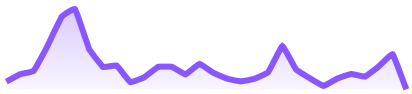
Train

Deploy

Use in Transformers

Model card | Files | Community 20

Downloads last month
7,967



Hosted inference API

Text Generation

Inference API has been turned off for this model.

Spaces using TheBloke/Llama-2-7B-Chat-GGML 39

mikeee/llama2-7b-chat-ggml | r3gm/ConversaDocs | johnwick123forevr/WizardLM7b | mikeee/nousresearch-nous-hermes-llama2-13b-ggml | Artples/LLaMA-2-CHAT | Artples/llama-2-7b-chat | mikeee/langchain-llama2-7b-chat-uncensored-ggml | 0xSynapse/LlamaGPT | TotoB12/llama2-7b-chat-ggml | maknee/kani-llama-v2-ggml | mikeee/llama2-7b-chat-uncensored-ggml | rphrp1985/llama2-7b-chat-uncensored-ggml | memef4rmer/llama2-7b-chat-uncensored-ggml | BojanSimoski/SocialMovezVeggieAssistant | lavanjv/vec-digichat | ndn1954/pdfchatbot | Jafta/llama2-7b-chat-ggml | XiangJinYu/Your_Cat | ndn1954/chatwithpdf | gordonchan/h2oo + 19 Spaces

Edit model card

Chat & support: my new Discord server

Want to contribute? TheBloke's Patreon page

Meta's Llama 2 7b Chat GGML

These files are GGML format model files for Meta's Llama 2 7b Chat.

GGML files are for CPU + GPU inference using llama.cpp and libraries and UIs which support this format, such as:

- KoboldCpp, a powerful GGML web UI with full GPU acceleration out of the box. Especially good for story telling.
- LoLLMS Web UI, a great web UI with GPU acceleration via the c_transformers backend.
- LM Studio, a fully featured local GUI. Supports full GPU accel on macOS. Also supports Windows, without GPU accel.
- text-generation-webui, the most popular web UI. Requires extra steps to enable GPU accel via llama.cpp backend.
- ctransformers, a Python library with LangChain support and OpenAI-compatible AI server.
- llama-cpp-python, a Python library with OpenAI-compatible API server.

Repositories available

- GPTQ models for GPU inference, with multiple quantisation parameter options.
- 2, 3, 4, 5, 6 and 8-bit GGML models for CPU+GPU inference
- Unquantised fp16 model in pytorch format, for GPU inference and for further conversions

Prompt template: Llama-2-Chat

```
[INST] <<SYS>>
You are a helpful, respectful and honest assistant. Always answer as helpful

If a question does not make any sense, or is not factually coherent, explain
<</SYS>>

{prompt} [/INST]
```

Compatibility

Original llama.cpp quant methods: `q4_0, q4_1, q5_0, q5_1, q8_0`

These are guaranteed to be compatible with any UIs, tools and libraries released since late May. They may be phased out soon, as they are largely superseded by the new k-quant methods.

New k-quant methods: `q2_K, q3_K_S, q3_K_M, q3_K_L, q4_K_S, q4_K_M, q5_K_S, q6_K`

These new quantisation methods are compatible with llama.cpp as of June 6th, commit 2d43387.

They are now also compatible with recent releases of text-generation-webui, KoboldCpp, llama-cpp-python, ctransformers, rustformers and most others. For compatibility with other tools and libraries, please check their documentation.

Explanation of the new k-quant methods

▶ Click to see details

Provided files

Name	Quant method	Bits	Size	Max RAM required	Use case
llama-2-7b-chat.ggmlv3.q2_K.bin	q2_K	2	2.87 GB	5.37 GB	New k-quant method. Uses GGML_TYPE_Q4_K for the attention.vw and feed_forward.w2 tensors, GGML_TYPE_Q2_K for the other tensors.
llama-2-7b-chat.ggmlv3.q3_K_L.bin	q3_K_L	3	3.60 GB	6.10 GB	New k-quant method. Uses GGML_TYPE_Q5_K for the attention.wv, attention.wo, and feed_forward.w2 tensors, else GGML_TYPE_Q3_K
llama-2-7b-chat.ggmlv3.q3_K_M.bin	q3_K_M	3	3.28 GB	5.78 GB	New k-quant method. Uses GGML_TYPE_Q4_K for the attention.wv, attention.wo, and feed_forward.w2 tensors, else GGML_TYPE_Q3_K
llama-2-7b-chat.ggmlv3.q3_K_S.bin	q3_K_S	3	2.95 GB	5.45 GB	New k-quant method. Uses GGML_TYPE_Q3_K for all tensors
llama-2-7b-chat.ggmlv3.q4_0.bin	q4_0	4	3.79 GB	6.29 GB	Original quant method, 4-bit.
llama-2-7b-chat.ggmlv3.q4_1.bin	q4_1	4	4.21 GB	6.71 GB	Original quant method, 4-bit. Higher accuracy than q4_0 but not as high as q5_0. However has quicker inference than q5 models.
llama-2-7b-chat.ggmlv3.q4_K_M.bin	q4_K_M	4	4.08 GB	6.58 GB	New k-quant method. Uses GGML_TYPE_Q6_K for half of the attention.wv and feed_forward.w2 tensors, else GGML_TYPE_Q4_K
llama-2-7b-chat.ggmlv3.q4_K_S.bin	q4_K_S	4	3.83 GB	6.33 GB	New k-quant method. Uses GGML_TYPE_Q4_K for all tensors
llama-2-7b-chat.ggmlv3.q5_0.bin	q5_0	5	4.63 GB	7.13 GB	Original quant method, 5-bit. Higher accuracy, higher resource usage and

Name	Quant method	Bits	Size	Max RAM required	Use case
					slower inference.
llama-2-7b-chat.ggmlv3.q5_1.bin	q5_1	5	5.06 GB	7.56 GB	Original quant method, 5-bit. Even higher accuracy, resource usage and slower inference.
llama-2-7b-chat.ggmlv3.q5_K_M.bin	q5_K_M	5	4.78 GB	7.28 GB	New k-quant method. Uses GGML_TYPE_Q6_K for half of the attention.wv and feed_forward.w2 tensors, else GGML_TYPE_Q5_K
llama-2-7b-chat.ggmlv3.q5_K_S.bin	q5_K_S	5	4.65 GB	7.15 GB	New k-quant method. Uses GGML_TYPE_Q5_K for all tensors
llama-2-7b-chat.ggmlv3.q6_K.bin	q6_K	6	5.53 GB	8.03 GB	New k-quant method. Uses GGML_TYPE_Q8_K for all tensors - 6-bit quantization
llama-2-7b-chat.ggmlv3.q8_0.bin	q8_0	8	7.16 GB	9.66 GB	Original quant method, 8-bit. Almost indistinguishable from float16. High resource use and slow. Not recommended for most users.

Note: the above RAM figures assume no GPU offloading. If layers are offloaded to the GPU, this will reduce RAM usage and use VRAM instead.

How to run in `llama.cpp`

I use the following command line; adjust for your tastes and needs:

```
./main -t 10 -ngl 32 -m llama-2-7b-chat.ggmlv3.q4_0.bin --color -c 4096 --te
```

Change `-t 10` to the number of physical CPU cores you have. For example if your system has 8 cores/16 threads, use `-t 8`.

Change `-ngl 32` to the number of layers to offload to GPU. Remove it if you don't have GPU acceleration.

If you want to have a chat-style conversation, replace the `-p <PROMPT>` argument with `-i -ins`

How to run in `text-generation-webui`

Further instructions here: text-generation-webui/docs/llama.cpp-models.md.

Discord

For further support, and discussions on these models and AI in general, join us at:

TheBloke AI's Discord server

Thanks, and how to contribute.

Thanks to the chirper.ai team!

I've had a lot of people ask if they can contribute. I enjoy providing models and helping people, and would love to be able to spend even more time doing it, as well as expanding into new projects like fine tuning/training.

If you're able and willing to contribute it will be most gratefully received and will help me to keep providing more models, and to start work on new AI projects.

Donaters will get priority support on any and all AI/LLM/model questions and requests, access to a private Discord room, plus other benefits.

- Patreon: https://patreon.com/TheBlokeAI
- Ko-Fi: https://ko-fi.com/TheBlokeAI

Special thanks to: Luke from CarbonQuill, Aemon Algiz.

Patreon special mentions: Space Cruiser, Nikolai Manek, Sam, Chris McCloskey, Rishabh Srivastava, Kalila, Spiking Neurons AB, Khalefa Al-Ahmad, WelcomeToTheClub, Chadd, Lone Striker, Viktor Bowallius, Edmond Seymore, Ai Maven, Chris Smitley, Dave, Alexandros Triantafyllidis, Luke @flexchar, Elle, ya boyyy, Talal Aujan, Alex , Jonathan Leane, Deep Realms, Randy H, subjectnull, Preetika Verma, Joseph William Delisle, Michael Levine, chris gileta, K, Oscar Rangel, LangChain4j, Trenton Dambrowitz, Eugene Pentland, Johann-Peter Hartmann, Femi Adebogun, Illia Dulskyi, senxiiz, Daniel P. Andersen, Sean Connelly, Artur Olbinski, RoA, Mano Prime, Derek Yates, Raven Klaugh, David Flickinger, Willem Michiel, Pieter, Willian Hasse, vamX, Luke Pendergrass, webtim, Ghost , Rainer Wilmers, Nathan LeClaire, Will Dee, Cory Kujawski, John Detwiler, Fred von Graf, biorpg, Iucharbius , Imad Khwaja, Pierre Kircher, terasurfer , Asp the Wyvern, John Villwock, theTransient, zynix , Gabriel Tamborski, Fen Risland, Gabriel Puliatti, Matthew Berman, Pyrater, SuperWojo, Stephen Murray, Karl Bernard, Ajan Kanaga, Greatston Gnanesh, Junyu Yang.

Thank you to all my generous patrons and donaters!

Original model card: Meta's Llama 2 7b Chat

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 fine-tuned model, optimized for dialogue use cases and converted for the Hugging Face Transformers format. 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 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	3.0×10^{-4}
Llama 2	*A new mix of publicly available online data*	13B	4k	✗	2.0T	3.0×10^{-4}
Llama 2	*A new mix of publicly available online data*	70B	4k	✔	2.0T	1.5×10^{-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/

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`.

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(tCO_2eq)
Llama 2 7B	184320	400	31.22
Llama 2 13B	368640	400	62.44
Llama 2 70B	1720320	400	291.42

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted(tCO_2eq)
Total	3311616		539.00

CO_2 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	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
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**

		TruthfulQA	Toxigen
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/

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
- Reporting problematic content generated by the model: developers.facebook.com/llama_output_feedback
- Reporting bugs and security concerns: facebook.com/whitehat/info

Llama Model Index

Model	Llama2	Llama2-hf	Llama2-chat	Llama2-chat-hf
7B	Link	Link	Link	Link
13B	Link	Link	Link	Link
70B	Link	Link	Link	Link

Company

TOS

Privacy

About

Jobs

Website

Models

Datasets

Spaces

Pricing

Docs