
TheBloke / **Akins-3B-GGUF** 

♡ like 4

Text Generation

Transformers

GGUF

Norquinal/claude_multiround_chat_1k

English

stablelm

arxiv:2305.14314

License: cc-by-sa-4.0

⋮

Train

Deploy

Use in Transformers


Model card

Files

Community 1

Downloads last month

18



Text Generation

Inference API has been turned off for this model.


Finetuned from [acrastrt/Akins-3B](#)

Dataset used to train TheBloke/Akins-3B-GGUF

Norquinal/claude_multiround_chat_1k

Preview • Updated Aug 11 • 45 • 5

Edit model card



[Chat & support: TheBloke's Discord server](#)

[Want to contribute? TheBloke's Patreon page](#)

TheBloke's LLM work is generously supported by a grant from [andreessen horowitz \(a16z\)](#).

Akins 3B - GGUF

Model creator: [Bohan Du](#)

Original model: [Akins 3B](#)

Description

This repo contains GGUF format model files for [Bohan Du's Akins 3B](#).

These files were quantised using hardware kindly provided by [Massed Compute](#).

About GGUF

GGUF is a new format introduced by the llama.cpp team on August 21st 2023. It is a replacement for GGML, which is no longer supported by llama.cpp.

https://huggingface.co/TheBloke/Akins-3B-GGUF

1/7

Here is an incomplete list of clients and libraries that are known to support GGUF:

- [llama.cpp](#). The source project for GGUF. Offers a CLI and a server option.
- [text-generation-webui](#), the most widely used web UI, with many features and powerful extensions. Supports GPU acceleration.
- [KoboldCpp](#), a fully featured web UI, with GPU accel across all platforms and GPU architectures. Especially good for story telling.
- [LM Studio](#), an easy-to-use and powerful local GUI for Windows and macOS (Silicon), with GPU acceleration.
- [LoLLMS Web UI](#), a great web UI with many interesting and unique features, including a full model library for easy model selection.
- [Faraday.dev](#), an attractive and easy to use character-based chat GUI for Windows and macOS (both Silicon and Intel), with GPU acceleration.
- [ctransformers](#), a Python library with GPU accel, LangChain support, and OpenAI-compatible AI server.
- [llama-cpp-python](#), a Python library with GPU accel, LangChain support, and OpenAI-compatible API server.
- [candle](#), a Rust ML framework with a focus on performance, including GPU support, and ease of use.

Repositories available

- [GPTQ models for GPU inference, with multiple quantisation parameter options.](#)
- [2, 3, 4, 5, 6 and 8-bit GGUF models for CPU+GPU inference](#)
- [Bohan Du's original unquantised fp16 model in pytorch format, for GPU inference and for further conversions](#)

Prompt template: User-Assistant

```
USER: {prompt}
ASSISTANT:
```

Compatibility

These quantised GGUFv2 files are compatible with llama.cpp from August 27th onwards, as of commit [d0cee0d](#)

They are also compatible with many third party UIs and libraries - please see the list at the top of this README.

Explanation of quantisation methods

► [Click to see details](#)

Provided files

Name	Quant method	Bits	Size	Max RAM required	Use case
akins-3b-Q2_K.gguf	Q2_K	2	1.20 GB	3.70 GB	smallest, significant quality loss - not recommended for most purposes
akins-3b-Q3_K_S.gguf	Q3_K_S	3	1.25 GB	3.75 GB	very small, high quality loss
akins-3b-Q3_K_M.gguf	Q3_K_M	3	1.39 GB	3.89 GB	very small, high quality loss
akins-3b-Q3_K_L.gguf	Q3_K_L	3	1.51 GB	4.01 GB	small, substantial quality loss
akins-3b-Q4_0.gguf	Q4_0	4	1.61 GB	4.11 GB	legacy; small, very high quality loss - prefer using Q3_K_M
akins-3b-Q4_K_S.gguf	Q4_K_S	4	1.62 GB	4.12 GB	small, greater quality loss
akins-3b-Q4_K_M.gguf	Q4_K_M	4	1.71 GB	4.21 GB	medium, balanced quality - recommended
akins-3b-Q5_0.gguf	Q5_0	5	1.94 GB	4.44 GB	legacy; medium, balanced quality - prefer using Q4_K_M
akins-3b-Q5_K_S.gguf	Q5_K_S	5	1.94 GB	4.44 GB	large, low quality loss - recommended
akins-3b-Q5_K_M.gguf	Q5_K_M	5	1.99 GB	4.49 GB	large, very low quality loss - recommended
akins-3b-Q6_K.gguf	Q6_K	6	2.30 GB	4.80 GB	very large, extremely low quality loss
akins-3b-Q8_0.gguf	Q8_0	8	2.97 GB	5.47 GB	very large, extremely low quality loss - not recommended

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 download GGUF files

Note for manual downloaders: You almost never want to clone the entire repo! Multiple different quantisation formats are provided, and most users only want to pick and download a single file.

The following clients/libraries will automatically download models for you, providing a list of available models to choose from:

- LM Studio
- LoLLMS Web UI
- Faraday.dev

[🔗](#) In text-generation-webui

Under Download Model, you can enter the model repo: TheBloke/Akins-3B-GGUF and below it, a specific filename to download, such as: akins-3b.Q4_K_M.gguf.

Then click Download.

🔗 On the command line, including multiple files at once

I recommend using the `huggingface-hub` Python library:

```
pip3 install huggingface-hub
```

Then you can download any individual model file to the current directory, at high speed, with a command like this:

```
huggingface-cli download TheBloke/Akins-3B-GGUF akins-3b.Q4_K_M.gguf
```

► More advanced `huggingface-cli` download usage

🔗 Example `llama.cpp` command

Make sure you are using `llama.cpp` from commit [d0cee0d](#) or later.

```
./main -ngl 32 -m akins-3b.Q4_K_M.gguf --color -c 2048 --temp 0.7
```

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

Change `-c 2048` to the desired sequence length. For extended sequence models - eg 8K, 16K, 32K - the necessary RoPE scaling parameters are read from the GGUF file and set by `llama.cpp` automatically.

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

For other parameters and how to use them, please refer to [the llama.cpp documentation](#)

🔗 How to run in `text-generation-webui`

Further instructions can be found in the `text-generation-webui` documentation, here: [text-generation-webui/docs/04 - Model Tab.md](#).

🔗 How to run from Python code

You can use GGUF models from Python using the [llama-cpp-python](#) or [ctransformers](#) libraries.

🔗 How to load this model in Python code, using `ctransformers`

🔗 First install the package

Run one of the following commands, according to your system:

```
# Base ctransformers with no GPU acceleration
pip install ctransformers
# Or with CUDA GPU acceleration
pip install ctransformers[cuda]
# Or with AMD ROCm GPU acceleration (Linux only)
CT_HIPBLAS=1 pip install ctransformers --no-binary ctransformers
# Or with Metal GPU acceleration for macOS systems only
CT_METAL=1 pip install ctransformers --no-binary ctransformers
```

[Simple ctransformers example code](#)

```
to 0 if no GPU acceleration is available on your system.
JF", model_file="akins-3b.Q4_K_M.gguf", model_type="stablelm", gpu_
```

[How to use with LangChain](#)

Here are guides on using llama-cpp-python and ctransformers with LangChain:

- [LangChain + llama-cpp-python](#)
- [LangChain + ctransformers](#)

[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!

Thanks to Clay from gpus.llm-utils.org!

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: Aemon Algiz.

Patreon special mentions: Brandon Frisco, LangChain4j, Spiking Neurons AB, transmissions 11, Joseph William Delisle, Nitin Borwankar, Willem Michiel, Michael Dempsey, vamX, Jeffrey Morgan, zynix, jjj, Omer Bin Jawed, Sean

Connelly, jinyuan sun, Jeromy Smith, Shadi, Pawan Osman, Chadd, Elijah Stavena, Illia Dulskyi, Sebastain Graf, Stephen Murray, terasurfer, Edmond Seymore, Celu Ramasamy, Mandus, Alex, biorpg, Ajan Kanaga, Clay Pascal, Raven Klaugh, 阿明, K, ya boyyy, usrbinkat, Alicia Loh, John Villwock, ReadyPlayerEmma, Chris Smitley, Cap'n Zoog, fincy, GodLy, S_X, sidney chen, Cory Kujawski, OG, Mano Prime, AzureBlack, Pieter, Kalila, Spencer Kim, Tom X Nguyen, Stanislav Ovsianikov, Michael Levine, Andrey, Trailburnt, Vadim, Enrico Ros, Talal Aujan, Brandon Phillips, Jack West, Eugene Pentland, Michael Davis, Will Dee, webtim, Jonathan Leane, Alps Aficionado, Rooh Singh, Tiffany J. Kim, theTransient, Luke @flexchar, Elle, Caitlyn Gatomon, Ari Malik, subjectnull, Johann-Peter Hartmann, Trenton Dambrowitz, Imad Khwaja, Asp the Wyvern, Emad Mostaque, Rainer Wilmers, Alexandros Triantafyllidis, Nicholas, Pedro Madruga, SuperWojo, Harry Royden McLaughlin, James Bentley, Olakabola, David Ziegler, Ai Maven, Jeff Scroggin, Nikolai Manek, Deo Leter, Matthew Berman, Fen Risland, Ken Nordquist, Manuel Alberto Morcote, Luke Pendergrass, TL, Fred von Graf, Randy H, Dan Guido, NimbleBox.ai, Vitor Caleffi, Gabriel Tamborski, knownsqashed, Lone Striker, Erik Bjäreholt, John Detwiler, Leonard Tan, Iucharbius

Thank you to all my generous patrons and donaters!

And thank you again to a16z for their generous grant.

 **Original model card: Bohan Du's Akins 3B**



This is [StableLM 3B 4E1T](#)(Licensed under [CC BY-SA 4.0.](#)) instruction tuned on [Claude Multiround Chat 1K](#) for 2 epochs with QLoRA([2305.14314](#)).

Prompt template:

```
USER: {prompt}
ASSISTANT:
```

GPTQ quantizations available [here](#).



Company

- TOS
- Privacy
- About
- Jobs

Website

- Models
- Datasets
- Spaces
- Pricing

Docs

© Hugging Face