r/LocalLLaMA Search Reddit

r/LocalLLaMA

LocalLLaMA

Join

r/LocalLLaMA

Posts Wiki

TABLE OF CONTENTS

r/LocalLLaMA

LocalLLaMA is a subreddit to discuss about LLaMA, the large language model created by Meta AI. It was created to foster a community around LLaMA similar to communities dedicated to open source like Stable Diffusion.

LLaMA FAQ

Q: What is LLaMA?

A: LLaMA (Large Language Model Meta AI) is a foundational large language model designed primarily for researchers. Like other large language models, LLaMA works by taking a sequence of words as an input and predicts a next word to recursively generate text. The data used to train the model is collected from various sources, mostly from the Web.

A: No. The original LLaMA models are not finetuned for question answering. They should be prompted so that the expected answer is the natural continuation of the prompt. Nonetheless, it is possible to chat with them in a way similar to ChatGPT but not near the same quality. With the possibility of new LoRAs and finetuning, however, bringing LLaMA up to ChatGPT quality is gradually becoming closer.

Q: What is Alpaca? Is it the same as LLaMA?

A: That refers to the Stanford Alpaca project, an effort to build an instruction-following LLaMA model from the standard 7B LLaMA model. It has been shown to produce results similar to OpenAI's text-davinci-003 off of a dataset that cost less than $500 and with training less than $100. Stanford released the full dataset, data generation code, and finetuning code on their GitHub page.

Q: What is Vicuna? Koala? OASST LLaMA? How are they different?

A: Those are the names of finetuned versions of LLaMA models, and they can provide substantially better quality than the original LLaMA models. By some evaulations, finetuned options like Vicuna reach near parity with ChatGPT when compared with certain test questions. However, keep in mind that none of these models can reach the quality of GPT-4, and the comparisons with ChatGPT are not always accurate. Models like Vicuna and OASST LLaMA (Open Assistant) also have similar restrictions to ChatGPT.

Q: How can I start using LLaMA now?

A: Follow either the simplified install guide or the original guide. Both continue to be updated with the latest information as necessary, including any fixes or extra tips. Feel free to ask questions in the comment section of the original guide for additional assistance, or you can check the GitHub pages for the repository you're using.

Getting started tips

After you've followed the guide and you have LLaMA running on your computer, ideally with text-generation-webui, you may be left wondering where to go from there. This section will cover some basic tips.

Models

For writing stories, the standard 30B and 65B LLaMA models are the best. The creative gap between 13B and 30B is huge, and 7B cannot compare at all. However, storywriting with untuned LLaMA models may require a lot more effort than finetuned versions. For easier storywriting with less effort, use a model like GPT4 x Alpaca, which can produce elegant language with little prompting.

For chatting, all models can be good, although results on 7B may be greatly subpar. For systems that are not very high end, 13B is the best tradeoff between performance and quality, and it is excellent in providing assistant answers when used with an informative character card or models

For all other tasks, like essay writing or other similar requests, Vicuna is excellent as it was trained using real ChatGPT data from ShareGPT. OASST LLaMA is a good alternative. GPT4 x Alpaca is capable of writing essays, but it is not as good at following instructions and may require more effort to reach a desired result.

Note: It's not recommended to use the standard 7B LLaMA model for anything. Use it with Alpaca LoRA or use a finetuned version of it instead.

Parameters

Whether you're using the web UI or llama.cpp, you'll want to experiment with the parameters to see what best suits the type of content you're looking to generate. Here are some example parameters you can use on their own or as a base for testing:

	Precise	**Creative**	**Sphinx**
Good for:	*factual responses and straightforward assistant answers*	*chatting, storywriting, and interesting assistant answers*	*varied storywriting* ***(on 30B/65B)*** *and unconventional chatting*
temperature	0.7	0.72	1.99
repetition penalty	1.176 (1/0.85)	1.1	1.15
top_k	40	0	30
top_p	0.1	0.73	0.18

It's recommended to use precise for mostly everything besides chatting and generating stories.

Prompting

Prompts matter. Using the right prompt in line with the idea you're trying to generate is needed for the best results. Prompting strategies are slightly different based on the model you're using.

Standard LLaMA

Knockoff Alpaca

You can use the Alpaca prompt for standard LLaMA, and it can work with varying degrees of success. If you're not trying to chat or create stories, this alone can be used for all other generations:

> ### Instruction:
>
> (instruction)
>
> ### Response:
>
> (your cursor should be here when generating)

This can be prone to hallucinations, especially near the end, but it's helpful when trying to force the model to give you a certain response. This knockoff Alpaca can then be combined with other prompting methods for better outputs.

Storywriting

Some variation of:

> Write a 1000 word story...

usually works well when trying to generate a long story. It won't actually generate a 1000 word story in most cases, but it can help with story lengths. For example:

> Write a 1000 word fantasy story about...
>
> Write a 1000 word horror story about...

and so on. This method works best when combined with knockoff Alpaca, and using the two together allows you to give extra instructions. For example, if you want a longer build-up to a main event, you can write something like:

> Write a 1000 word story about...(etc.) Write a long build-up to the protagonist setting off at sea.

and you should generally see that show up in the response. Or if you want metaphors:

> Write a 1000 word story about...(etc.) Use metaphors when writing.
>
> Write a 1000 word story about...(etc.) Use metaphors and figurative language when writing.

Another simple and alternative storywriting trick is to prime the response. This can be more helpful when using settings more likely to cause hallucinations. For instance, if you're trying to write a WW2 story, you can start a line with the first sentence or paragraph of that story:

> Write a 1000 word dark story about a soldier fighting in the trenches during WW2.
>
> *I was down on my luck, but then again, I was already down in the trenches.*

If you're using knockoff Alpaca, the primer goes below "Response", and your cursor should be directly in front of the period or right below the line when generating. Think of where you want your story to start. Cursor placement is important.

CAI-Chat

UI, if you want an experience similar to ChatGPT, you can use a character card modified to act like it. Here is one example:

```
{

    "char_name": "LLaMA",

    "char_persona": "LLaMA's primary function is to interact with users through nat

    "char_greeting": "Hello there! How can I help you today? Do you have any questi

    "world_scenario": "",

    "example_dialogue": "{{user}}: Why is the sky blue?\n{{char}}: The blue color c

}
```

Copy that and save it as LLaMA.json inside text-generation-webui/characters. If you provide a PNG or JPEG image with the same name inside the same folder, the character will use that image. For users on llama.cpp, you can instead use a script to emulate ChatGPT. Here's an example showing a similar character with llama.cpp:

```
#!/bin/bash
MODEL="./models/30B/ggml-model-q4_0.bin"
user_name="${USER_NAME:-User}"
assistant_name="${AI_NAME:-LLaMA}"
echo user_name: $user_name
echo assistant_name: $assistant_name
# Change number of threads(-t) and context length (-c) to what fits your hardware
GEN_OPTIONS="-c 2048 --temp 0.7 --top_k 40 --top_p 0.5 --repeat_last_n 500 --repeat
./main -m "$MODEL" -t 18 -n 4096 $GEN_OPTIONS --color -i -r "${user_name}:" -p \ "
LLaMA's primary function is to interact with users through natural language process
${assistant_name}: Hello there! How can I help you today? Do you have any questions
${user_name}: Why is the sky blue?
${assistant_name}: The blue color of the sky is due to a phenomenon called Rayleigh
${user_name}: Interesting, can you give me the equations for Einstein's theory of r
${assistant_name}: Yes, I can provide you with some of the equations used in Einste
${user_name}: People seem to tend to have a shallow amount of contextualization whe
${assistant_name}: Yes, what you're describing is a common phenomenon in communicat
${user_name}:" "$@"
```

For different chatting personalities or types, you can create or download character cards for what you want out of LLaMA. If you want an assistant that provides rote answers like Wikipedia, then all you would have to do is copy text from Wikipedia articles and use that for your example chats.

Update: This section mostly applies to Alpaca LoRA and Alpaca Native. It was written before finetuned versions like GPT4 x Alpaca were created.

The default Alpaca format to always use is:

```
Below is an instruction that describes a task. Write a response that appropriatel

### Instruction:

Tell me about alpacas.

### Response:

(your cursor should be on this line)
```

Alpaca isn't the right option for creating stories or chatting, so that won't be covered too much here. However, for everything else, there is one effortless way to get greatly improved and longer results. You can turn this:

Tell me about Edgar Allan Poe.

> Edgar Allan Poe was an American writer, poet, and critic who is widely regarded as a central figure of Romanticism in the United States and of the short story genre. He was best known for his tales of mystery and macabre, including The Raven, The Fall of the House of Usher, and The Tell-Tale Heart.

into this:

Tell me about Edgar Allan Poe.

> Edgar Allan Poe was one of the most influential writers of the 19th century. He is best known for his dark, mysterious stories and poems, which have been widely read and adapted into many different media over the years. Poe's work explored themes of death, madness, love, and revenge, often in a Gothic setting. His writing style was characterized by its vivid imagery, rich language, and deep symbolism.
>
> Poe had a difficult life from a young age. After being orphaned at the age of three, he moved around frequently with his foster family and struggled to find stability. He attended several colleges but never graduated. He eventually found success as a writer, publishing numerous short stories and poems. However, he also suffered from depression and alcoholism, and died at the age of 40. Despite this, his legacy lives on through his works, which continue to inspire people today.

with a single added sentence. Dozens of combinations have been tested over numerous hours to come up with the easiest and most effective way of improving results for all tasks. All you have to do is prepend this special line before your actual instruction:

> In a long and detailed response, tell me in at least two paragraphs: "instruction"

instructions, but what's provided above is a catchall for when you don't know what to use.

```
"tell me in at least three paragraphs" doesn't work as effectively.

"tell me in a very long essay" doesn't work as effectively.

"In a verbose and extra detailed response" doesn't work as effectively.
```

Variations like those have already been tried, so it's recommended to simply use the original line instead. Keep in mind that you don't always want to use this line. For responses with very short, factual answers, do not prepend the line and instead use the direct instruction. For example:

> "If I put a small bowl of water on top of a door then walk through the door, what will happen?"

is a question that you do not want to use the special line for.

When it comes to trying to chat with Alpaca, you can try the suggestion about changing the names and context, but it's not recommended to chat with Alpaca. It's also not recommended to try to create stories with Alpaca but to instead use a standard LLaMA model with the knockoff Alpaca for more creative results. However, if you're very limited in hardware capabilities and cannot use one of the larger models for creating stories, then Alpaca Native 4-bit would be the right choice to use.

Other models

Vicuna

This model can give results that are very close to ChatGPT, and it is excellent at following instructions. It may be the best option if you want a ChatGPT clone, but it may not be the right choice for creativity.

The default Vicuna format is:

```
A chat between a curious human and an artificial intelligence assistant. The assi…

### Human: Tell me about vicuñas.

### Assistant: (your cursor can be on this line)
```

The default Vicuna v1.1 format is:

```
A chat between a curious user and an artificial intelligence assistant. The assis…

User: Tell me about vicuñas.

Assistant: (your cursor can be on this line)
```

The default OASST LLaMA format is:

```
<|prompter|>What is an assistant, and what's the history behind their role in soc

(your cursor can be on this line)
```

Other wiki pages

When it comes to inference of LLaMA, text-generation-webui and llama.cpp are the two main ways to interacting with it. The web UI is currently the ideal choice.

For other projects made by members of the community, see this list here. This page also includes various datasets.

For models, see this list here. It includes models that can be used with the web UI and llama.cpp.

Last revised by Civil_Collection7267 - 1 month ago