› Run a ChatGPT-Like LLM on You
›

Updated: March 29, 2023 5:12 pm

›er of AI players in the market right now, including ChatGPT, Google Bard, Bing AI Chat, and many more. However, all of them require you to have an internet connection to interact with the AI. What if you want to install a similar Large Language Model (LLM) on your computer and use it locally? An AI chatbot that you can use privately and without internet connectivity. Well, with the new Alpaca model released by Stanford, you can come close to that reality. Yeah, you can run a ChatGPT-like language model on your PC offline. So on that note, let's go ahead and learn how to use an LLM locally without the internet.

Run a ChatGPT-Like LLM Locally Without Internet (Private and Secure)

In this article, I have mentioned everything about how to run a ChatGPT-like LLM on a local PC without the internet. You can expand the table below and learn about the steps in detail.

Table of Contents

What is Alpaca and LLaMA?

effective it is. With just 7 billion parameters, Alpaca is as good as OpenAI's text-davinci-003 model. And you can run it on your local computer without requiring an internet connection. That's pretty cool, right?

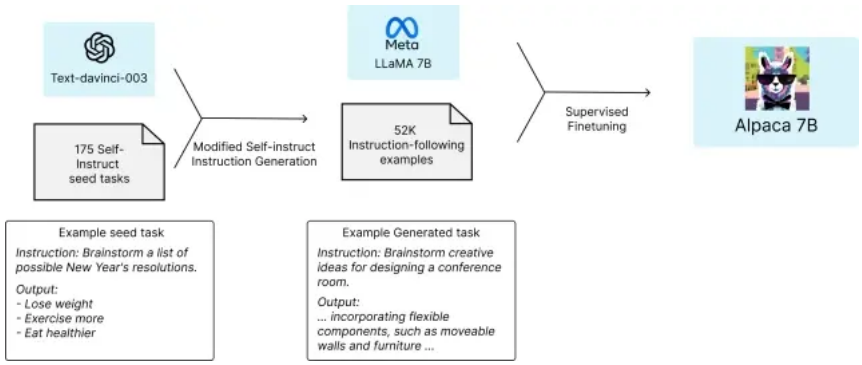


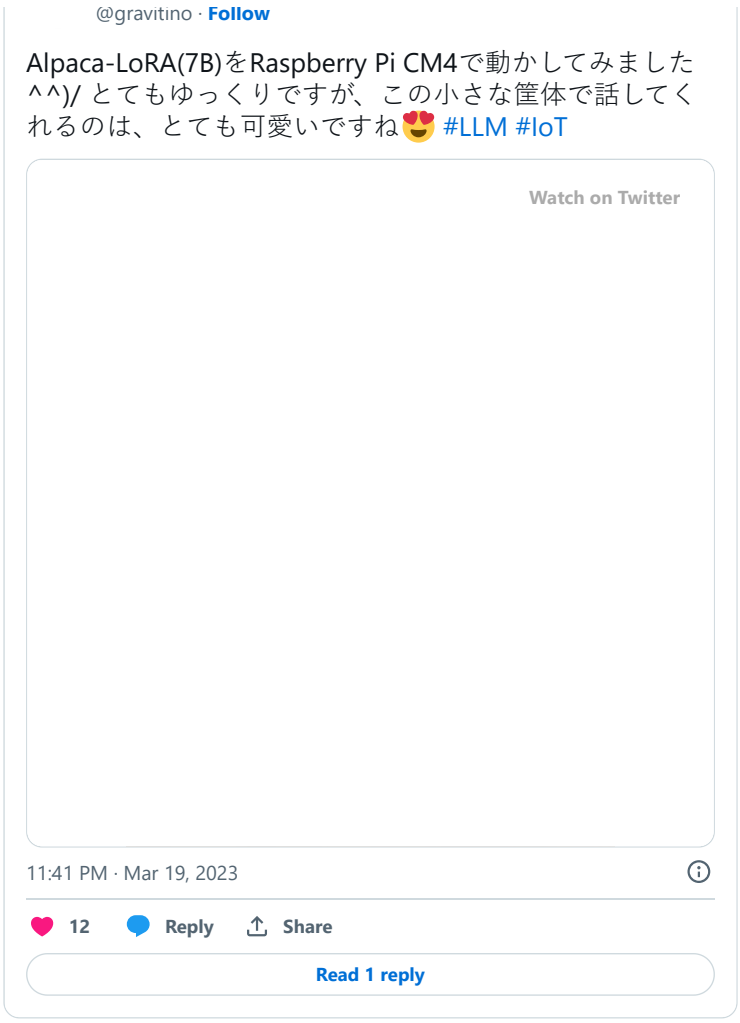
Image: Stanford.edu

But how was it trained? Surprisingly, Alpaca is **fine-tuned on LLaMa**, Meta's large language model which recently leaked online. And to train this language model, scientists used OpenAI's "text-davinci-003" model to generate 52K high-quality self-instruction data. With this dataset, they fine-tuned the LLaMA model using HuggingFace's training framework and released the **Alpaca 7B**. You can also use Meta's LLaMA model, but in my testing, Stanford's Alpaca LLM performed much better and it's also quite fast.

What Kind of Hardware Do You Need to Run Alpaca?

You can use Alpaca 7B on any decent machine. I installed Alpaca 7B on my entry-level PC and it worked quite well. To give you some idea, my PC is powered by a 10th-Gen Intel i3 processor with 256GB of SSD and 8GB of RAM. For GPU, I am using Nvidia's entry-level GeForce GT 730 GPU with 2GB of VRAM.

AI
How to Turn off Chat History in ChatGPT
NEWS
Apple Plans to Introduce an AI-Health Coach and More Healthy
AI
How to Make Money with Chat(Easy Ways)
AI
How to Fix ChatGPT "Error in Bc Stream" (9 Methods)
AI
How to Use ChatGPT to Write E That Impress



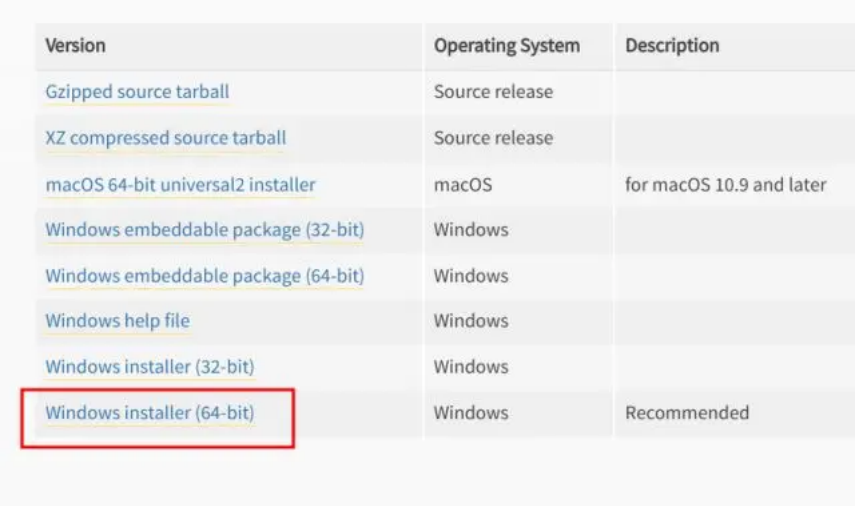
Even **without a dedicated GPU**, you can run Alpaca locally. However, the response time will be slow. Apart from that, there are users who have been able to run Alpaca even on a tiny computer like Raspberry Pi 4. So you can infer that the Alpaca language model can very well run on entry-level computers as well.

Set Up the Software Environment to Run Alpaca and LLaMA

Windows

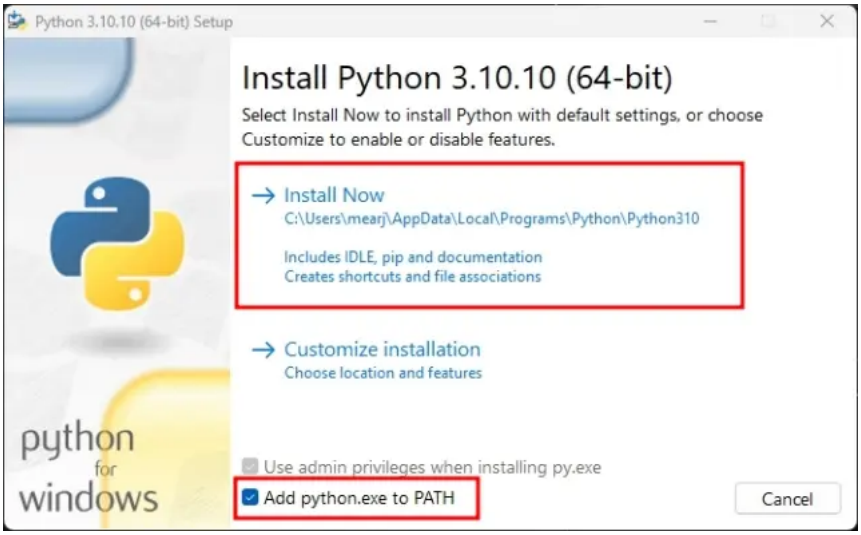
On Windows, you need to install Python, Node.js, and C++ to get started with using a large language model offline on your computer. Here is how to go about it.

1. First, download **Python 3.10** (or below) from here. Scroll down and click on "Windows installer (64-bit)" to download the setup file.

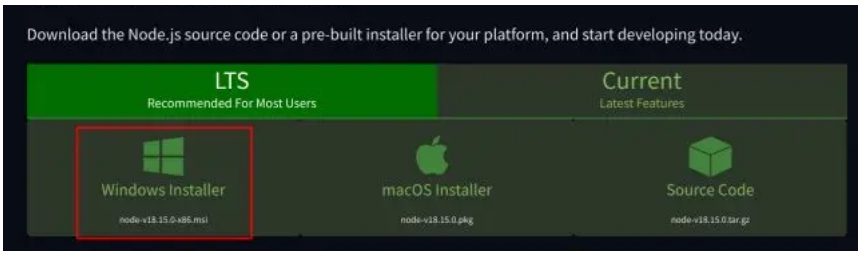


Version	Operating System	Description
Gzipped source tarball	Source release	
XZ compressed source tarball	Source release	
macOS 64-bit universal2 installer	macOS	for macOS 10.9 and later
Windows embeddable package (32-bit)	Windows	
Windows embeddable package (64-bit)	Windows	
Windows help file	Windows	
Windows installer (32-bit)	Windows	
Windows installer (64-bit)	Windows	Recommended

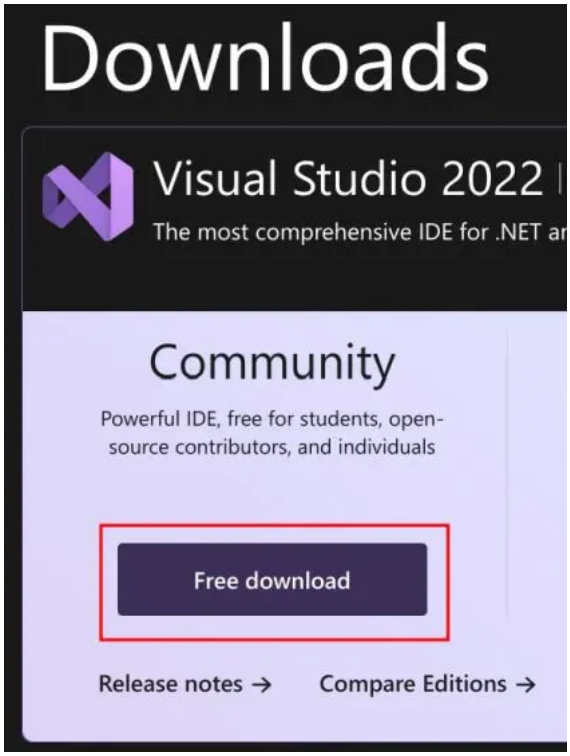
2. Launch the setup file and enable the checkbox next to "**Add Python.exe to PATH**." Now, install Python with all default settings.



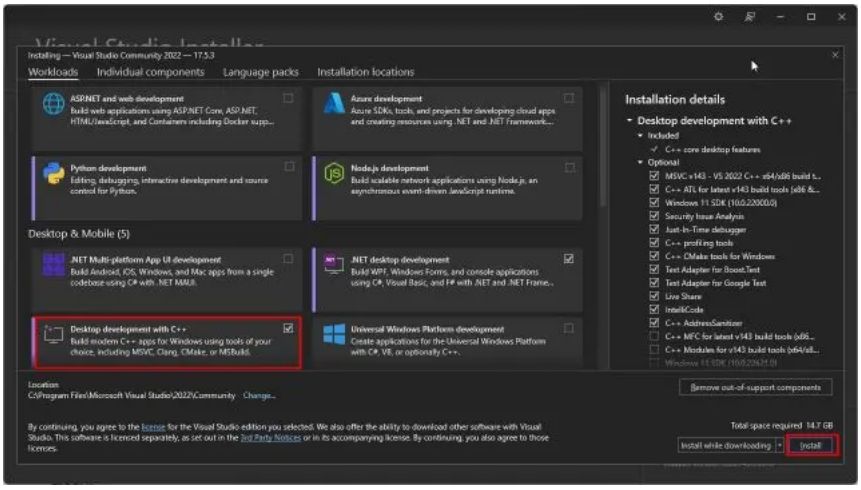
3. After that, install **Node.js** version 18.0 (or above) from here. Keep everything default while installing the program.



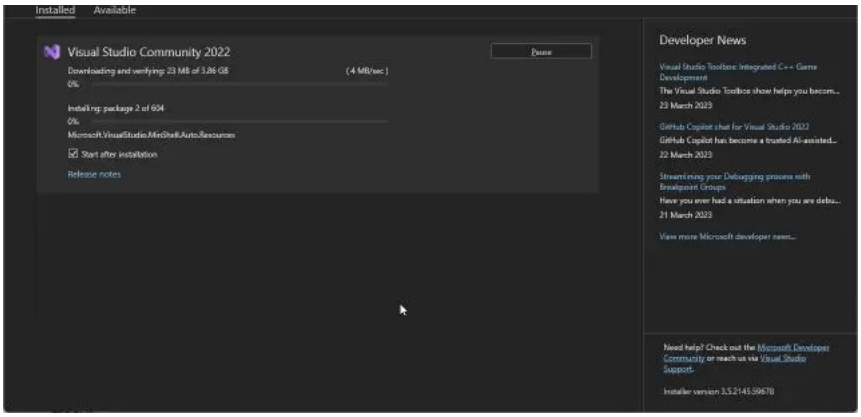
4. Finally, download the Visual Studio "Community" edition from this link for free.



5. Launch the Visual Studio 2022 setup file, and it will initially download some files. After that, a new window will launch. Here, make sure "**Desktop development with C++**" is enabled.

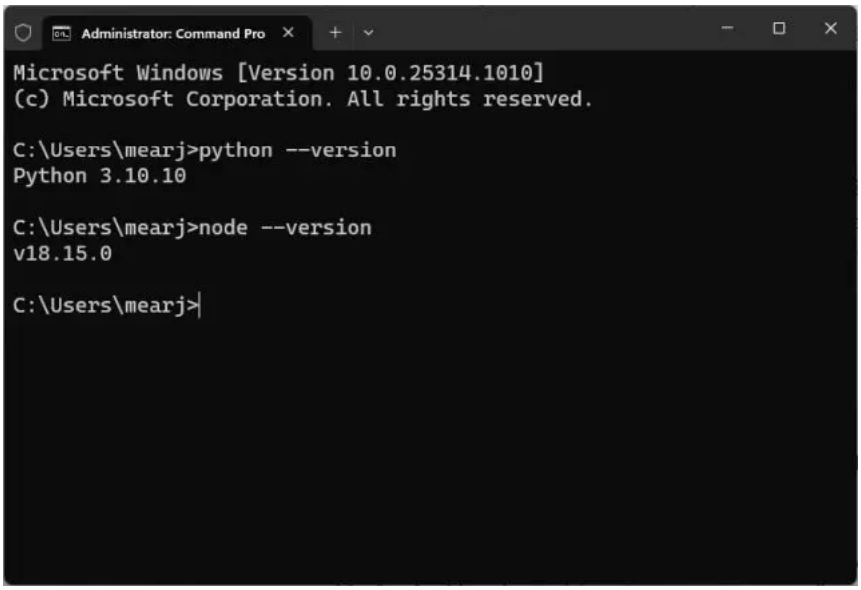


6. Finally, click "**Install**" and wait until it completes the installation.



7. I recommend restarting your computer once everything is installed. Next, open "**Command Prompt**" and run the below commands to check if Python and Node.js are installed successfully. Both should return the version number. You are now good to go.

```
python --version
node --version
```

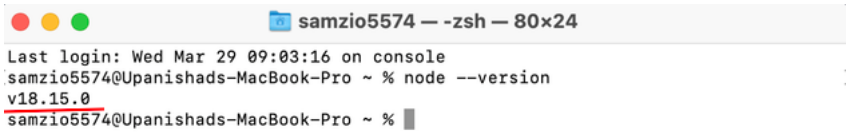


Apple macOS

Python generally comes pre-installed on macOS, so you only need to install Node.js (version 18.0 or above). Here is how you can do it:

2. Next, open the Terminal and run the below command to **check if Node.js is installed properly**. If you get a version number in return, you are good to go.

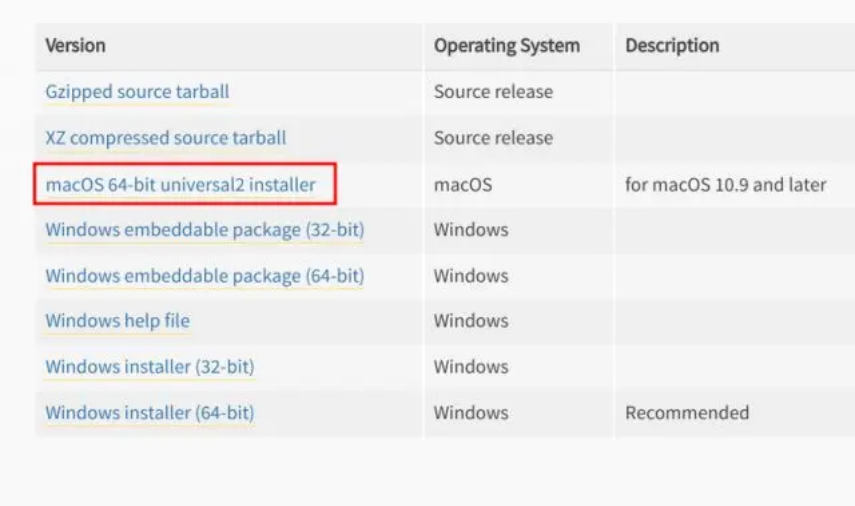
```
node --version
```



3. Next, check the Python version by running the below command. It should be **Python 3.10 or below**.

```
python3 --version
```

4. If you don't get output or you happen to have the latest Python version, download Python 3.10 (or below) from here. Scroll down and click on "**macOS 64-bit universal2 installer**" to download Python. Now, install it on your Mac.



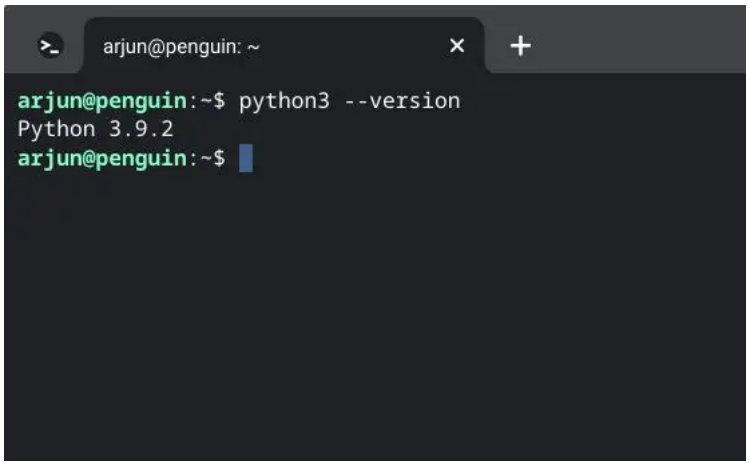
Version	Operating System	Description
Gzipped source tarball	Source release	
XZ compressed source tarball	Source release	
macOS 64-bit universal2 installer	macOS	for macOS 10.9 and later
Windows embeddable package (32-bit)	Windows	
Windows embeddable package (64-bit)	Windows	
Windows help file	Windows	
Windows installer (32-bit)	Windows	
Windows installer (64-bit)	Windows	Recommended

Linux and ChromeOS

On Linux and ChromeOS, you need to set up Python and Node.js before you run offline Alpaca and LLaMA models. Here are the steps to follow.

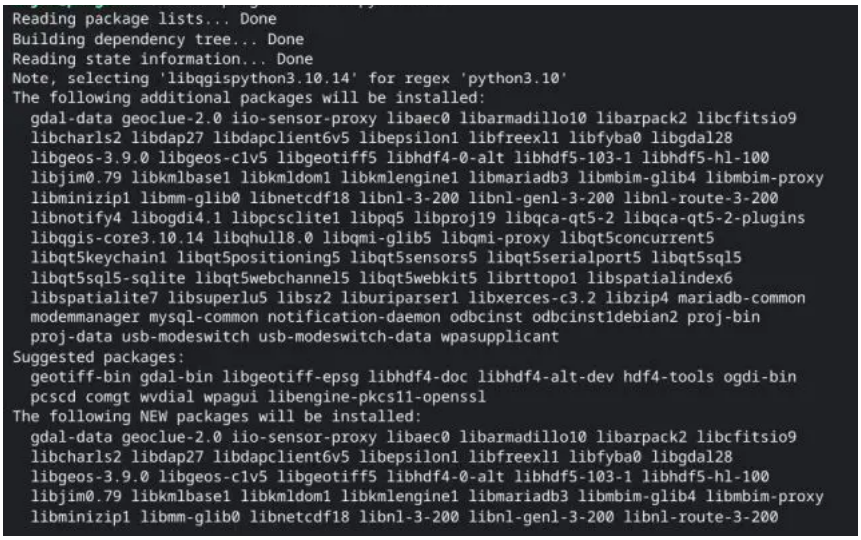
1. Open the Terminal and run the below command to check the Python version. If it's **Python 3.10 or below**, you are all set.

```
python3 --version
```



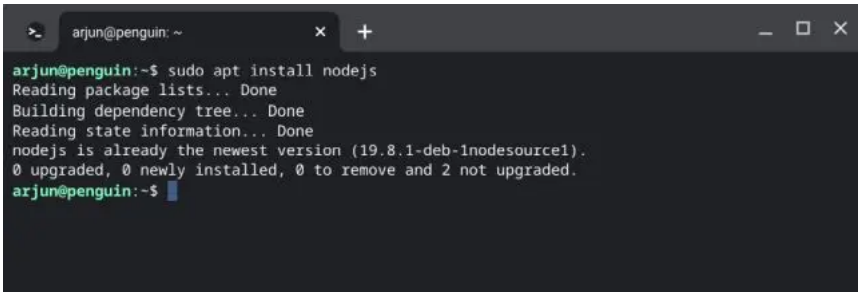
2. In case you have a **higher version**, you can use the below commands to install Python 3.10 on Linux and ChromeOS.

```
sudo apt install software-properties-common
sudo add-apt-repository ppa:deadsnakes/ppa
sudo apt-get update
sudo apt-get install python3.10
```



3. After Python, **install Node.js** by running the below command.

```
sudo apt install nodejs
```



4. After the installation, run the below command to check the Node.js version. It should be **18.0 or higher**.

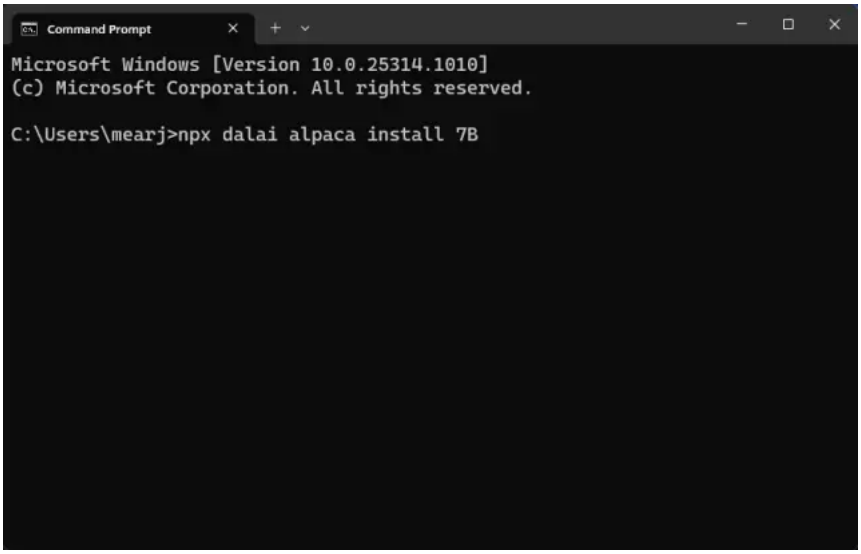

```
node --version
```

Install Alpaca and LLaMA Models On Your Computer

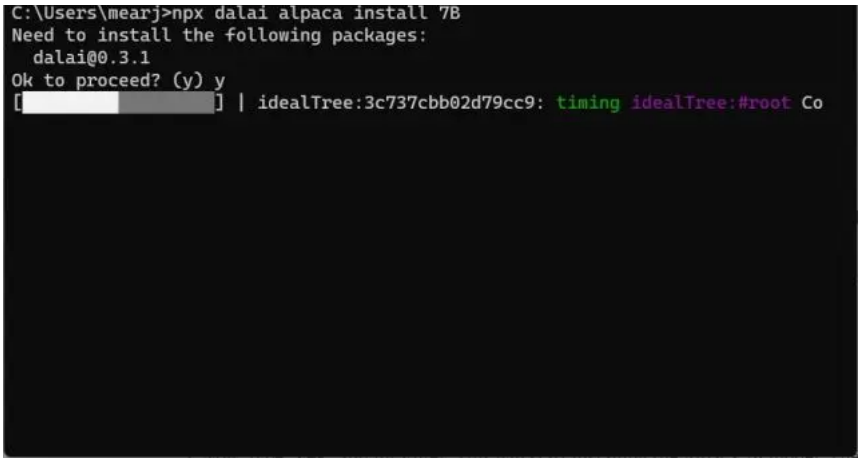
Once you have set up Python and Node.js, it's time to install and run a ChatGPT-like LLM on your PC. Make sure the Terminal detects both `python` and `node` commands before you proceed.

1. Open the Terminal (in my case, Command Prompt) and run the below command to **install the Alpaca 7B LLM model** (around 4.2GB disk space required). If you want to install the Alpaca 13B model, replace `7B` with `13B`. The larger model needs 8.1GB of space.

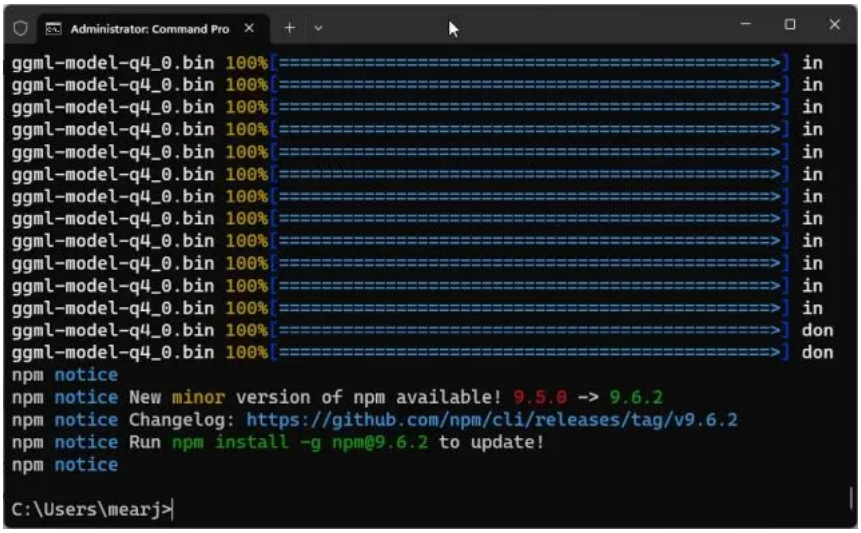
```
npx dalai alpaca install 7B
```



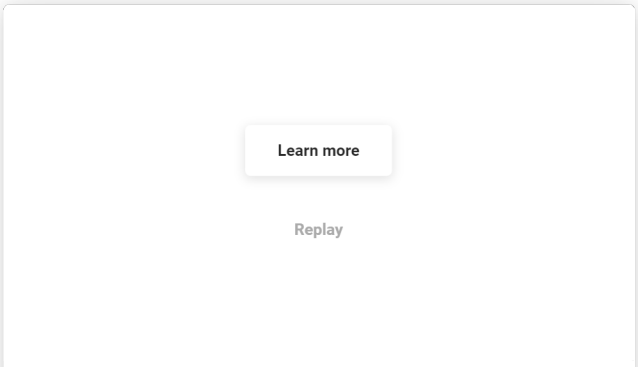
2. Now, type "**y**" and hit Enter. This will start installing the Alpaca 7B model. The whole process will take 20 to 30 minutes, depending on your internet connectivity and model size.



3. After the **installation is complete**, you will see a screen like this.



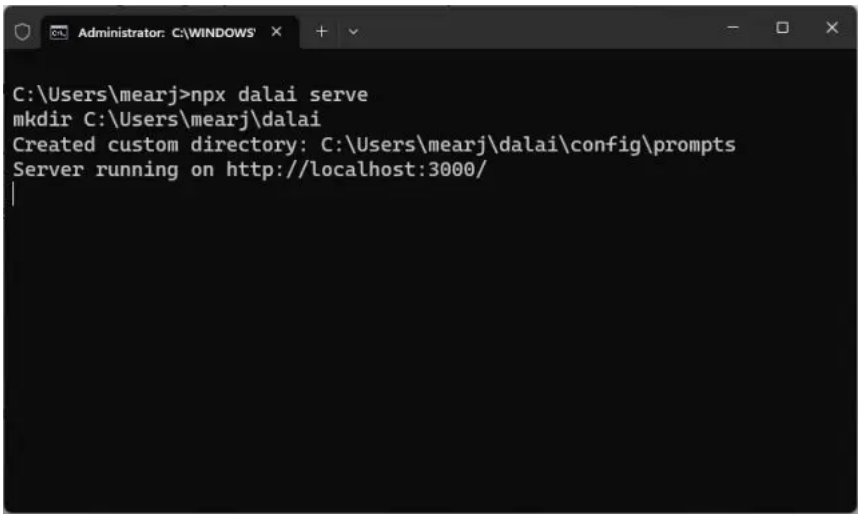
4. You can choose to **install LLaMA models** as well or move to the next step to test the Alpaca model instantly. Remember, LLaMA is much larger in size. Its 7B model takes up to 31GB of space. To install it, run the below command. You can replace `7B` with `13B`, `30B`, and `65B`. The largest model takes up to 432GB of space.




```
npx dalai llama install 7B
```

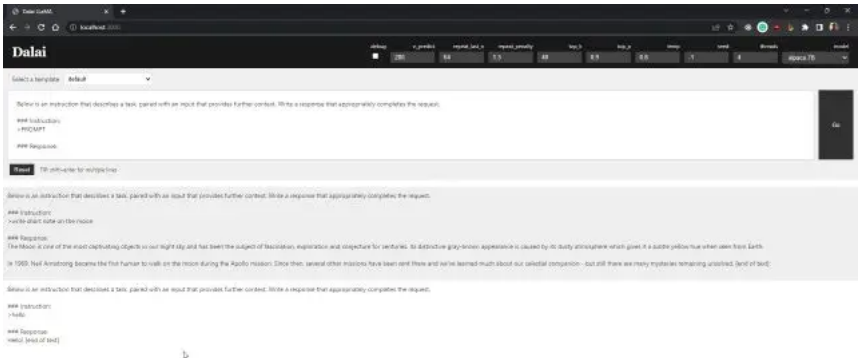
5. Finally, run the below command, and it will **start the webserver** instantly.

```
npx dalai serve
```

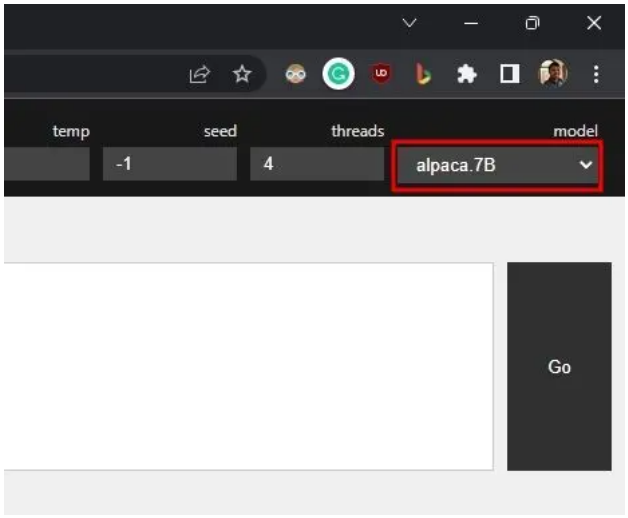


6. **Use a web browser** on your PC and open the below address. This will take you to the web UI where you can test Alpaca and LLaMA models locally and without the internet.

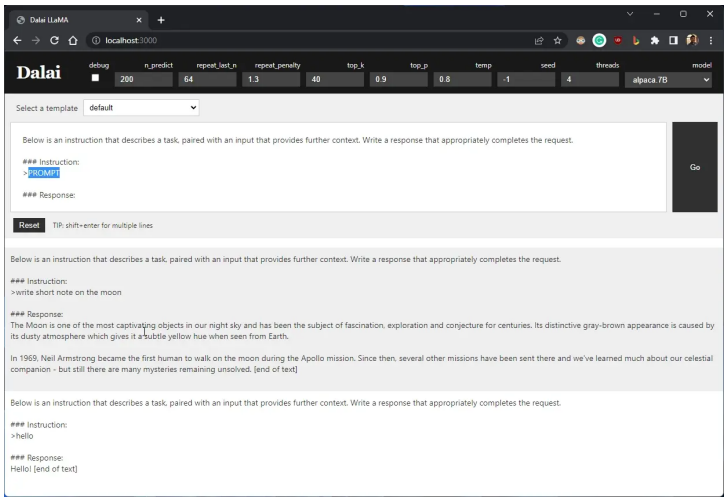
```
http://localhost:3000
```



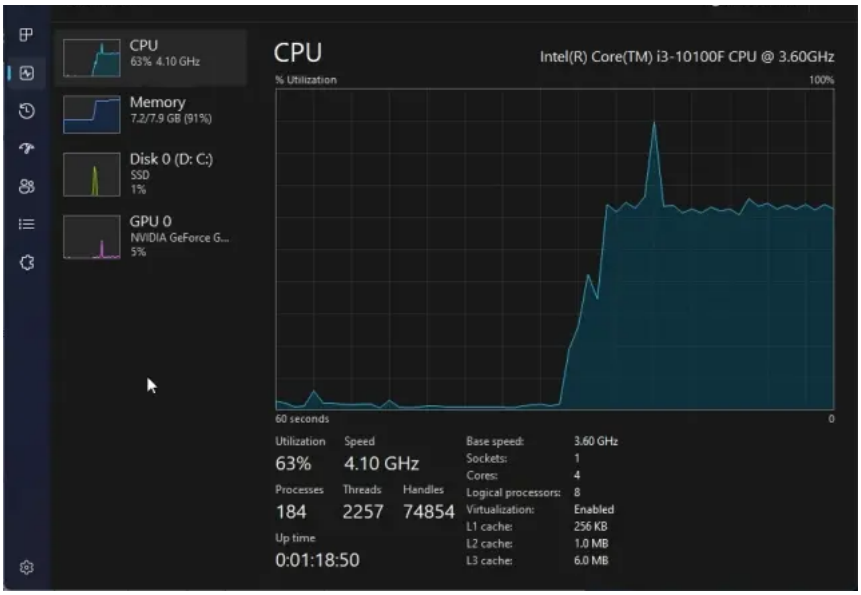
model, this is my default.



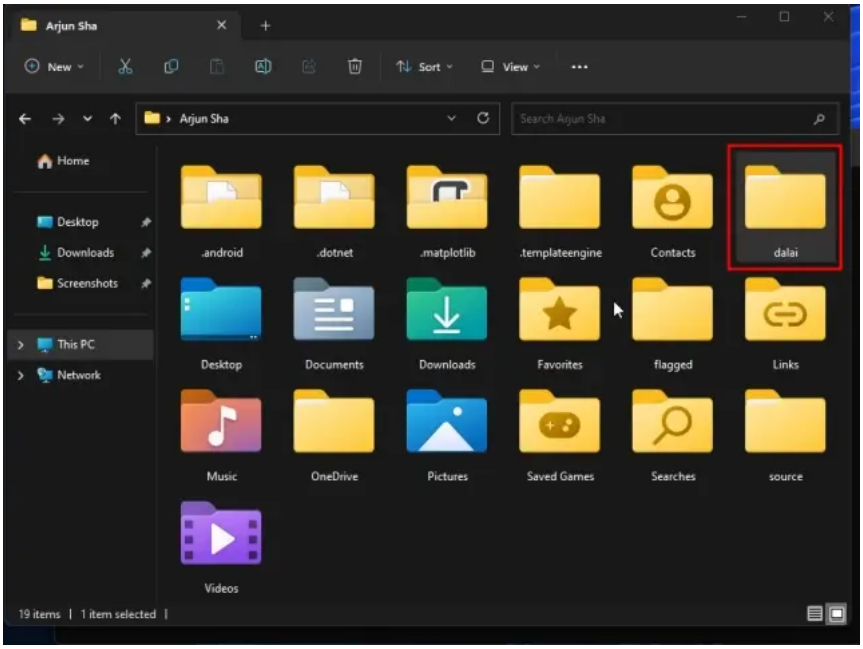
8. You can now **start using this ChatGPT-like language model on your PC** without internet connectivity. Replace “PROMPT” with your query and click on “Go”.



9. Here is what the **resource usage looks like** while running the local Alpaca LLM server on my Windows PC.



10. In case you want to **delete the downloaded models** to free up disk space, open your user profile directory. Here, the “dalai” folder has all the files, including the model. Deleting the “dalai” folder will free up space immediately.



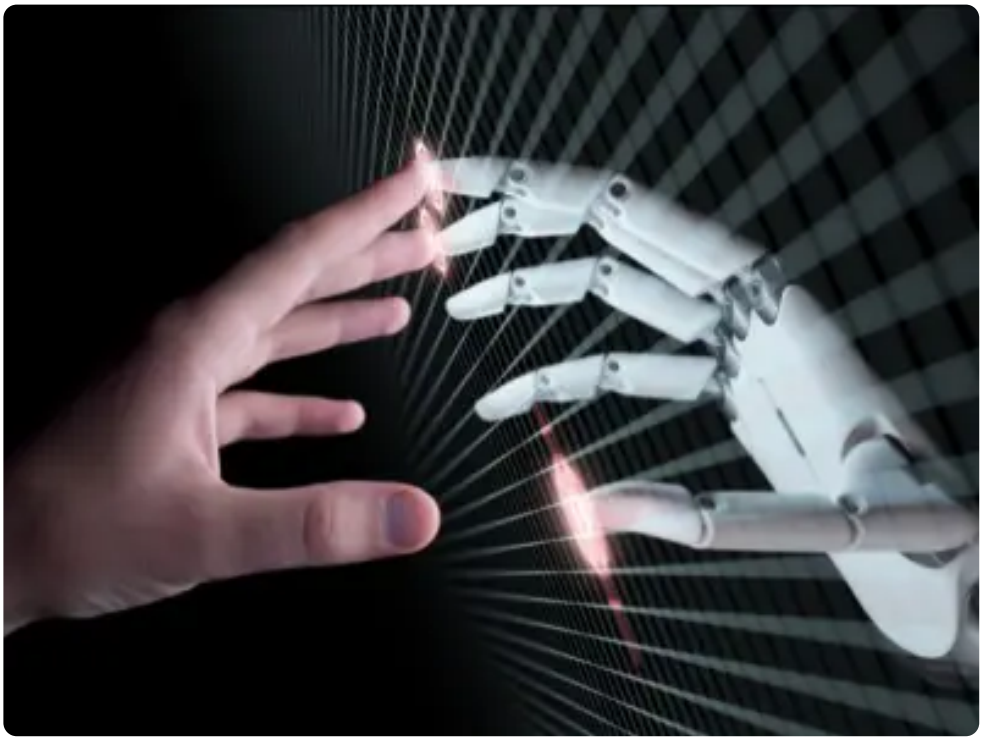
Use a ChatGPT-Like Service Privately and Completely Offline

So this is how you can run a ChatGPT-like LLM on your local PC and get decent results as well. As time goes by, new and highly-efficient LLM models will be available in the future which can be run on smartphones to small-board computers like Raspberry Pi. Anyway, that is all from us. If you want to use ChatGPT 4 for free, head to our linked article for some amazing resources. And in case you want to train an AI chatbot based on your own documents, we have an in-depth guide ready for you. Finally, if you are facing any problems, let us know in the comment section below.

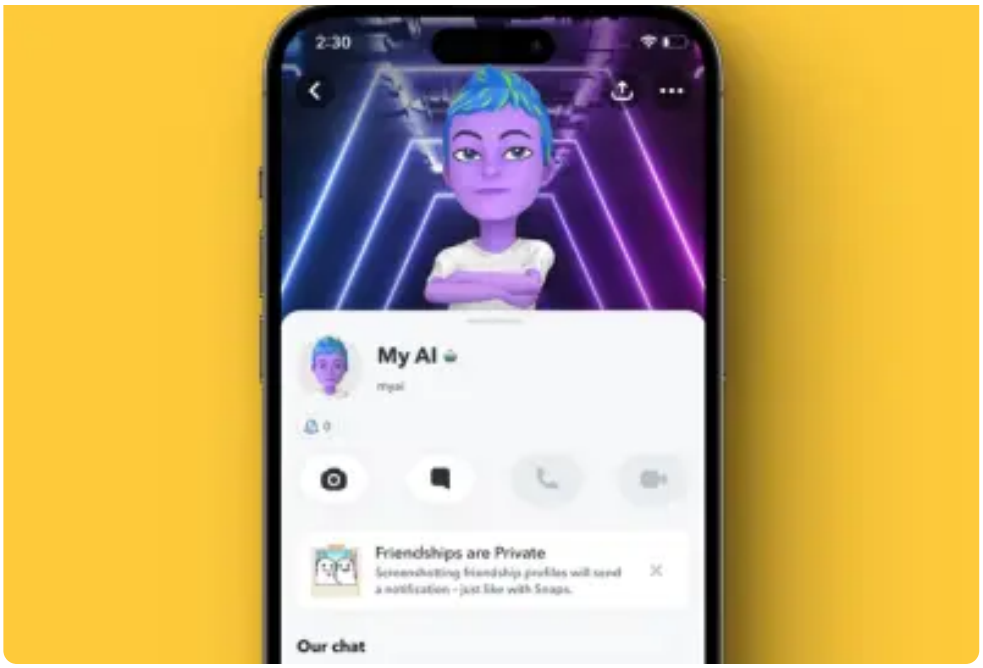
TAGS AI Alpaca chatGPT featured LLaMa

8 Comments

RECOMMENDED ARTICLES

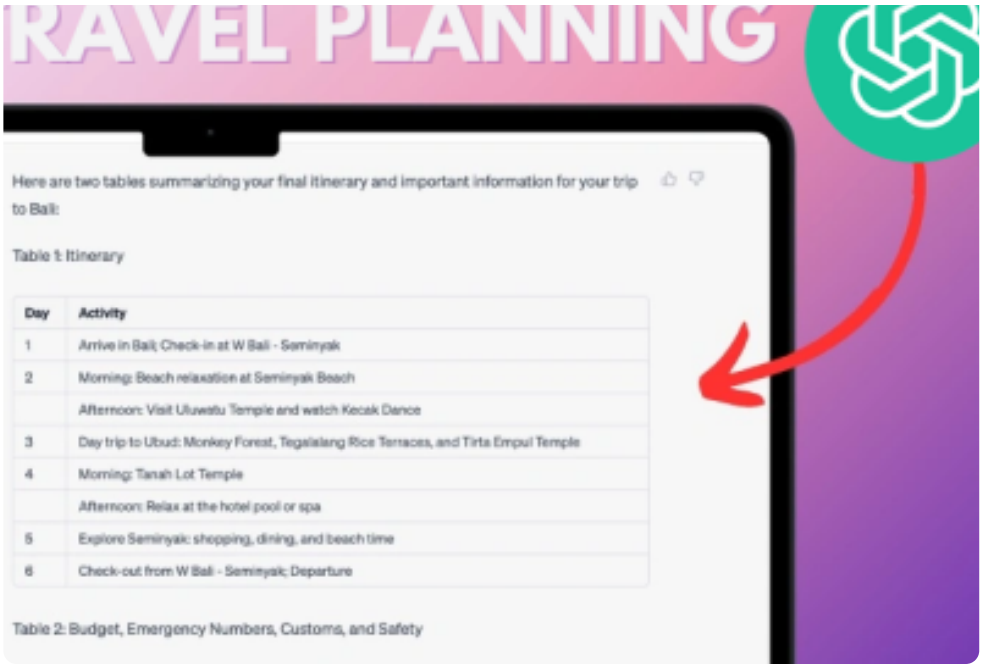


18 Examples of AI You're Using in Daily Life in 2023



How to Get Rid of Snapchat “My AI” Bot from Chat Feed

This Is How Much the OnePlus Pad Will Cost You in India



How to Use ChatGPT for Travel Planning

Apple Saket Store Is Now Open; Here's My Experience!

Minecraft Legends Review: Fun Strategy Spin-off But with Flaws

8 COMMENTS

max Apr 26, 2023 at 1:28 am

One more vote and please to add to Fahim and Jake, please provide code to run custom data!

Reply

Alex Apr 25, 2023 at 4:55 am

The chatbot doesn't seem to work. I reinstalled it multiple times but got the following error on CMD:

```
Server running on http://localhost:3000/
> query: { method: 'installed', models: [] }
modelsPath C:\Users\alexy\dalai\alpaca\models
{ modelFolders: [ '13B', '30B' ] }
exists 30B
modelsPath C:\Users\alexy\dalai\llama\models
{ modelFolders: [ '7B' ] }
> query: {
seed: -1,
threads: 4,
n_predict: 200,
top_k: 40,
top_p: 0.9,
temp: 0.8,
repeat_last_n: 64,
repeat_penalty: 1.3,
debug: false,
models: [ 'alpaca.30B' ],
prompt: 'Below is an instruction that describes a task, paired with an input that provides further
context. Write a response that appropriately completes the request.\n' +
'\n' +
'### Instruction:\n' +
'5+5 = \n' +
'\n' +
'### Response:\n',
id: 'TS-1682378437315-60919'
}
C:\Users\alexy\AppData\Local\npm-
cache\_npx\3c737cbb02d79cc9\node_modules\dalai\index.js:219
let [Core, Model] = req.model.split(".")
^

TypeError: Cannot read properties of undefined (reading 'split')
at Dalai.query (C:\Users\alexy\AppData\Local\npm-
cache\_npx\3c737cbb02d79cc9\node_modules\dalai\index.js:219:35)
```

at Socket.emitUntyped (C:\Users\alexy\AppData\Local\npm-cache_npx\3c737cbb02d79cc9\node_modules\socket.io\dist\typed-events.js:69:22)
at C:\Users\alexy\AppData\Local\npm-cache_npx\3c737cbb02d79cc9\node_modules\socket.io\dist\socket.js:703:39
at process.processTicksAndRejections (node:internal/process/task_queues:77:11)

Node.js v18.16.0

Does anyone who where the issues are and what I have to do to fix them?

C:\Users\alexy>python –version
Python 3.10.10

C:\Users\alexy>node –version
v18.16.0

Reply

John Apr 21, 2023 at 5:26 am

npx dalai llama install 7B – always hangs downloading the model, I was able to download the model manually in the web browser “https://agi.gpt4.org/llama/LLaMA/7B/consolidated.00.pth” still running the webserver I was not given the dropdown select for LLaMa model likely because it didn’t complete the installation.

Reply

Saral Apr 23, 2023 at 8:48 am

There’s another way to run. Search Llama.cpp (compiled exe files) and then search ggml-vicuna-7b-4bit.bin. Download both.

create a bat file with following codes, and run it.

———————
title llama.cpp
:start
main -i –interactive-first -r “### Human:” –temp 0 -c 2048 -n -1 –ignore-eos –repeat_penalty 1.2 –instruct -m ggml-vicuna-7b-4bit.bin
pause
goto start
———————

if there’s a high CPU usage, use the switch -t 2 in above code. You can change 2 to other number.

Reply

max Apr 26, 2023 at 1:29 am

Thank you, I am doing this with the larger models now, where I had persistent fails before!

Reply

Jake Apr 1, 2023 at 10:54 pm

Echoing what Fahim said, I’d like to run the Alpaca LLM against my custom dataset offline. Can you advise? Thanks!

Reply

Addy Mar 31, 2023 at 3:22 pm

npx dalai alpaca install 7B command stops itself while running, idk why its happening. HELP?

Reply

Fahim Masud Choudhury Mar 30, 2023 at 10:08 pm

Hi Arjun! Thanks for the excellent write-up.

I was wondering if it’s possible to train the private and offline models with a custom dataset. Similar to this article here: https://beebom.com/how-train-ai-chatbot-custom-knowledge-base-chatgpt-api/amp/ but not sending data to OpenAI API.

I’ll appreciate any feedback, thanks!

Reply

LEAVE A REPLY

Comment:

Your Name

Name:*

Your Email *

Email:*

Post Comment

REVIEWS

BenQ PD2706UA Review: A Perfect 4K Productivity Monitor

9.2

The BenQ PD2706UA monitor is here, and it comes with all the bells and whistles that productivity users would appreciate. 4K resolution, factory-cal colors, a 27-inch panel, an ergonomic stand that can be adjusted easily, and more. It has many [...]

Minecraft Legends Review: Fun Strategy Spin-off But with Flaws

7.5

Minecraft Legends is a game that piqued my interest at its original reveal last year. But, I will admit that I did not actively follow the game well until w closer to its official release. After all, my love [...]

MSI Titan GT77 HX 13V Review: Desktop-Grade Performance for the Price of a Car

8.6

Last year, MSI launched the Titan GT77 with the Intel Core i9-12900HX and the RTX 3080 Ti Laptop GPU, and it was the most powerful gaming lapt face of the planet. It was the heaviest of heavy hitters [...]