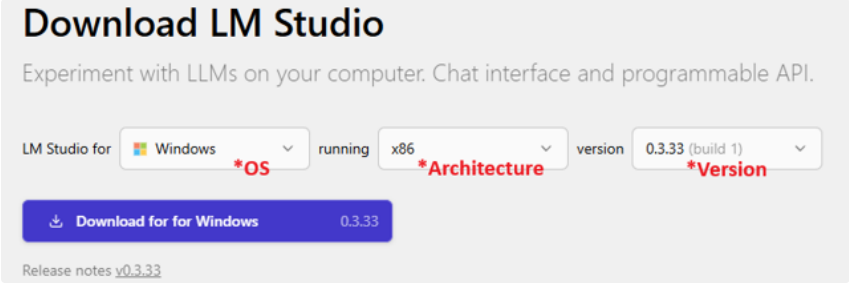
How To Download and Configure LM Studio

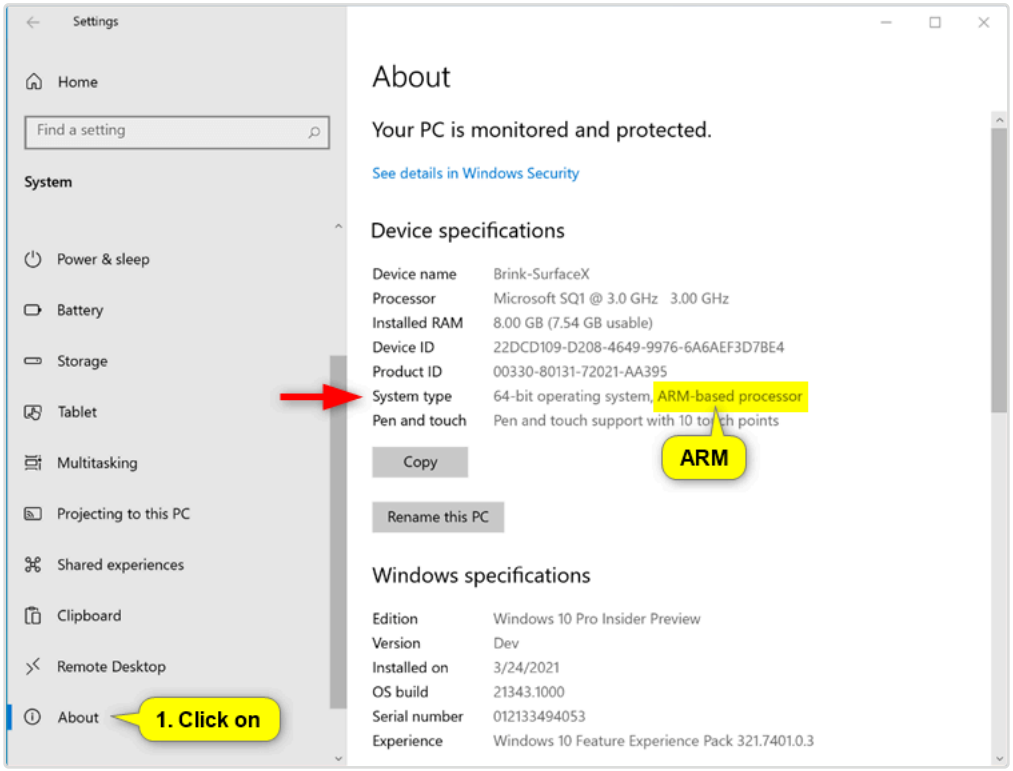
This is a step by step guide on how to download LM Studio and configure it for your personal needs so that you can easily use it as you would an online chat bot.

Download LM Studio Installer from browser

1. Go to the following link: Download LM Studio - Mac, Linux, Windows
2. Choose OS, architecture, and Version.



*Note: to verify the correct architecture on Windows, go to (Settings > System > About) and under “Device specifications” find “System type”. Unless it says ARM based processor choose the x86 architecture option

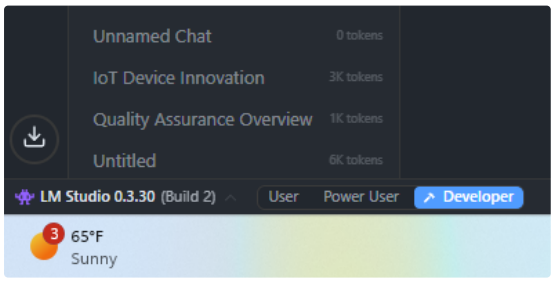


3. Click Download
4. Open Installer file “LM-Studio-0.3.33-1-x64.exe”

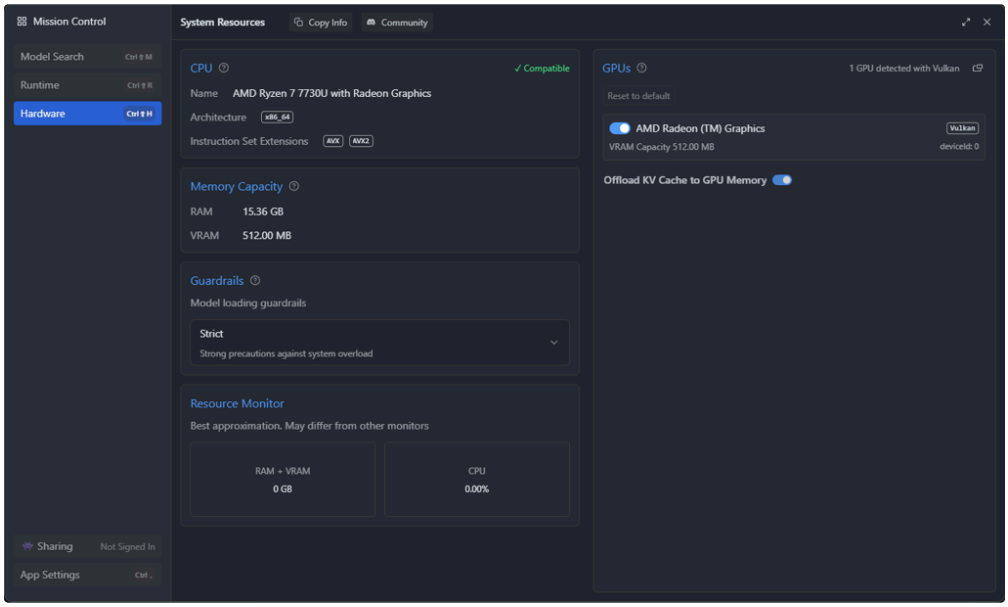
5. If you choose to download for “Anyone who uses this computer (all users)” it will be saved in “C:\Program Files\LM Studio”
   If you choose to download “Only for me (username)” it will be saved in ”C:\Users\username\AppData\Local\Programs\LM Studio”
6. Confirm Download
7. Run LM Studio either from start menu, or from .exe file if needed
   “C:\Users\username\AppData\Local\Programs\LM Studio\LM Studio.exe”

Assessing Hardware

1. Toggle on developer mode in bottom right hand corner



2. Click on the search icon in the left hand panel, this opens “Mission Control” pop up
3. Click on “Hardware” in “Mission Control” pop up



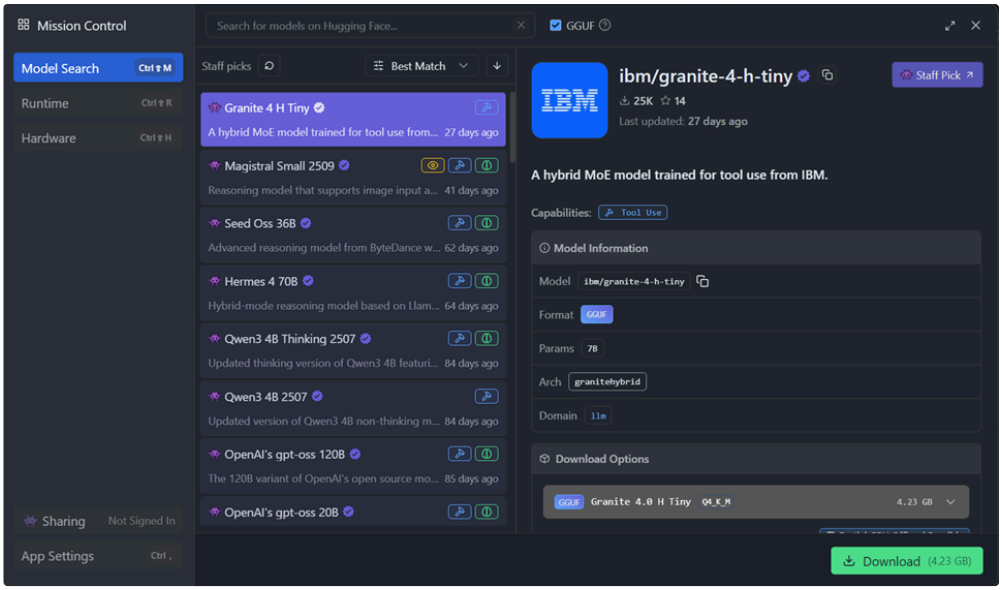
4. Your RAM + VRAM gives you the total capacity your computer has to run a model.
   *Note: here is where the trade off comes in. AI Computers used by companies like OpenAI, have more VRAM allowing for quicker processing and responses. However, if your computer does not have enough VRAM for the specified model, it simply distributes the workload to your RAM, which will be a bit slower.
5. With your (RAM + VRAM), you can now assess what models are compatible on your computer by using the following chart.

Table values refer to (RAM + VRAM) Requirements in (GB), ex. 3.3 = 3.3 GB of (RAM + VRAM)

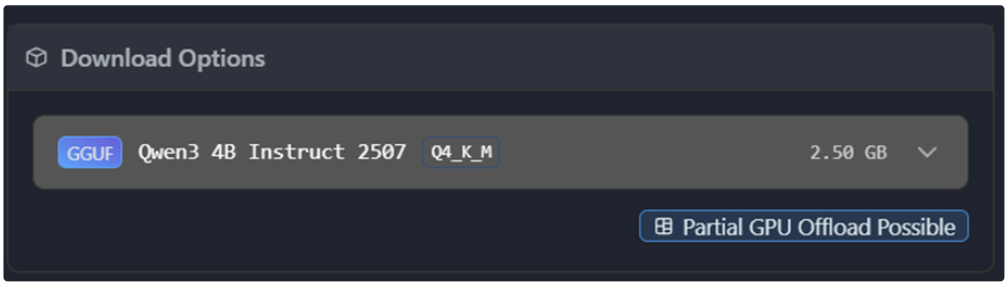
LLM Size	Q8	Q6	Q5	Q4	Q3	Q2	Q1
3B	3.3	2.5	2.1	1.7	1.3	0.9	0.6
7B	7.7	5.8	4.8	3.9	2.9	1.9	1.3
8B	8.8	6.6	5.5	4.4	3.3	2.2	1.5
9B	9.9	7.4	6.2	5.0	3.7	2.5	1.7
12B	13.2	9.9	8.3	6.6	5.0	3.3	2.2
13B	14.3	10.7	8.9	7.2	5.4	3.6	2.4
14B	15.4	11.6	9.6	7.7	5.8	3.9	2.6
21B	23.1	17.3	14.4	11.6	8.7	5.8	3.9
22B	24.2	18.2	15.1	12.1	9.1	6.1	4.1
27B	29.7	22.3	18.6	14.9	11.2	7.4	5.0
33B	36.3	27.2	22.7	18.2	13.6	9.1	6.1
65B	71.5	53.6	44.7	35.8	26.8	17.9	11.9
70B	77.0	57.8	48.1	38.5	28.9	19.3	12.8
74B	81.4	61.1	50.9	40.7	30.5	20.4	13.6
105B	115.5	86.6	72.2	57.8	43.3	28.9	19.3
123B	135.3	101.5	84.6	67.7	50.7	33.8	22.6
205B	225.5	169.1	141.0	112.8	84.6	56.4	37.6
405B	445.5	334.1	278.4	222.8	167.1	111.4	74.3

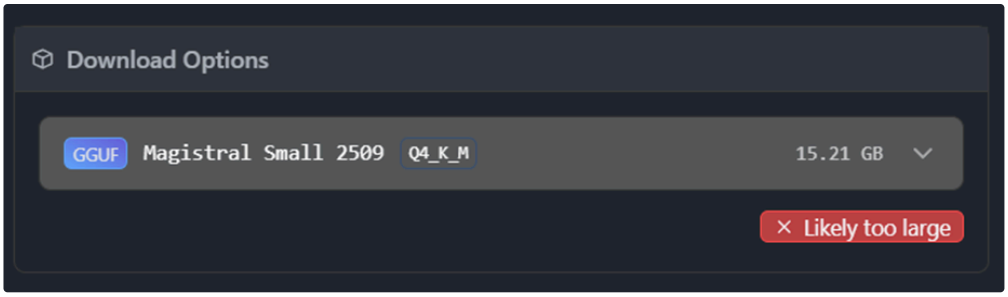
Choose and download a model

1. Toggle on developer mode in bottom right hand corner
2. Click magnifying glass “Discover” from left hand panel
3. Once you see a pop-up window Click “Model Search” from left hand panel

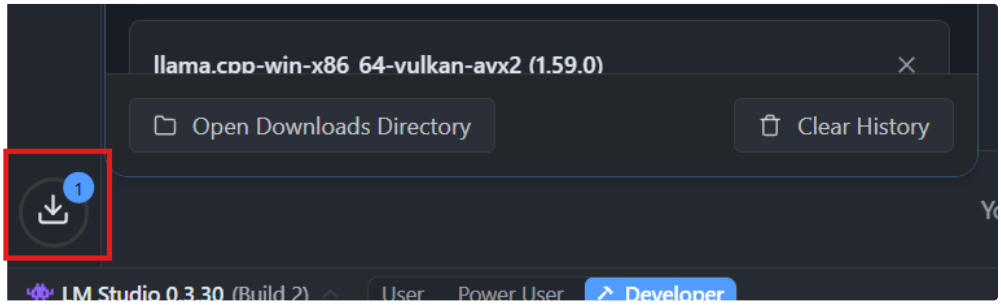


4. Toggle the GGUF checkbox to the right of the search bar
5. Now without typing any keyword in the search bar you will automatically see staff picked models by LM studio, recognized by the purple gamer icon next to it.
6. Alternatively you can find a GGUF model of your choosing on Hugging face (🤗 Hugging Face – The AI community building the future.) such as “Qwen/Qwen3-0.6B-GGUF” and then search the model in the LM Studio “Model Search”. You can use the number of likes and downloads to verify it is the same one.
7. Once you click on a model you will see a preview of the model’s information on the right. Here you can also see the download options available which may show different quantization versions of the model
8. Depending on your hardware and the size of the models some “Download Options” may have warning labels next to them notifying you that the model may only use partial GPU processing or that the model is entirely too large for your machine





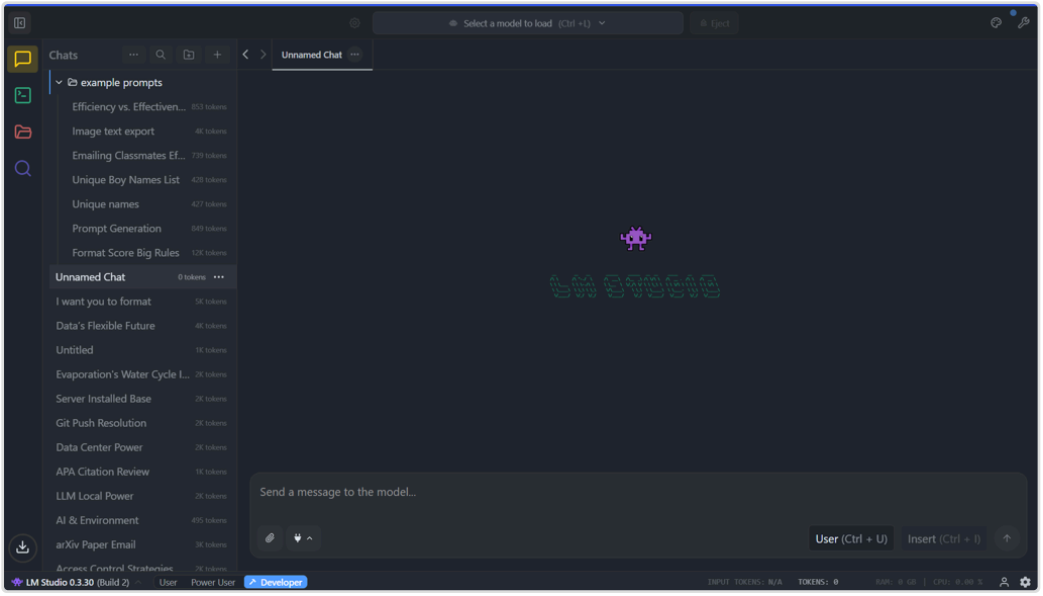
9. Choose your model and Click “Download”
10. You can see the download progress in the “Downloads” section in the bottom right corner.



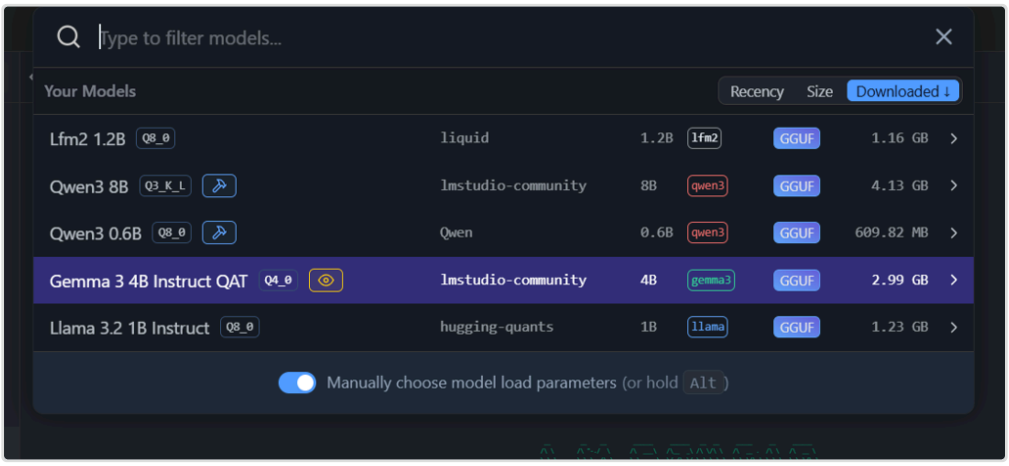
11. Once your download is done you are ready to give it a try through LM Studio’s chatbot interface!

Using a Model

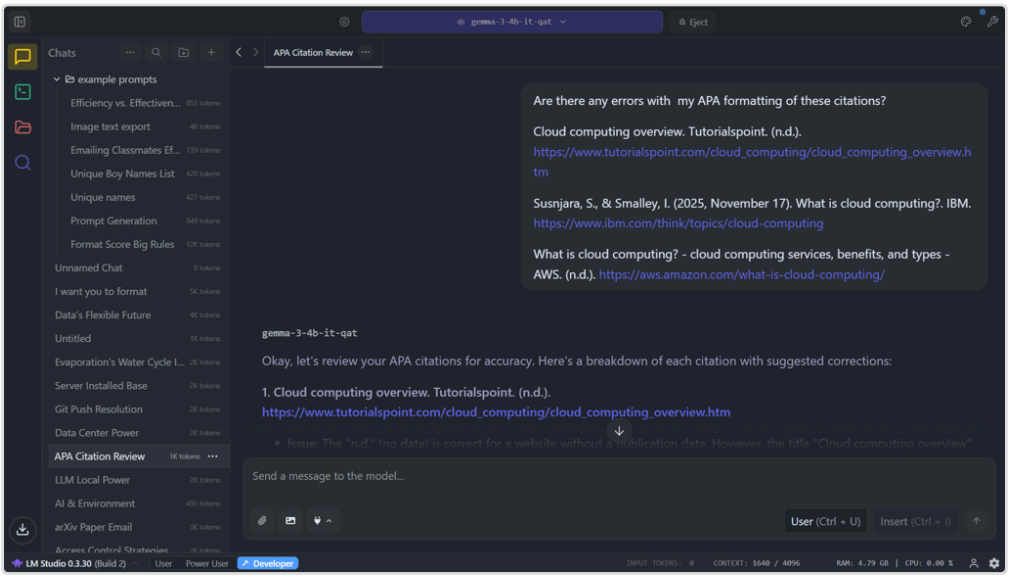
1. Click on the “Chat” icon in the top left corner or switch to “User” mode to open up the chatbot interface.
2. On the left panel you can see a history of chats which you can search through using keywords and even organize into folders.
3. Across the middle of your screen is your chat area where you can type prompts in at the bottom, and see your responses up above just like majority of online chatbot interfaces.



4. To start chatting, first load a model by clicking “Select a model to load (Ctrl +L)” at the top of the screen.
5. Choose a model from the drop down list of downloaded models, you might see a page open up with Model Configurations settings you can modify, otherwise you can leave the settings as they are and click “Load Model” in the bottom right.



6. Once the model is fully loaded you can type a prompt in at the bottom and start chatting. *Note: the responses you get back from are mostly dependent on the model you chose, try out different models to find out which ones work best for you. Model Suggestions for everyday use



Model Configurations

Change Context Length

- Use Gear Icon to the left of your selected model at the top of your screen

System prompt

- Use Wrench icon at top right corner of the screen to pull up model settings on the right panel of the screen

- Click Context on top left of the panel to enter a system prompt to be used throughout your chat

Model settings

- Temperature means the randomness of the answers
- Toggle Response length to set a limit in tokens of the models response
- Toggle All switch next to wrench icon within the right hand panel to show context overflow settings, choosing between the options of Rolling window, Truncate Middle, and Stop at limit

Model Suggestions for everyday use

*Note: highly recommend “gemma-3-4b-it-qat” as a good mid range model, it's my personal favorite

LLM	Publisher	Params	Quant	Size	HugginFace Link
gemma-3-4b-it-qat	lmstudio-community	4B	Q4_0	3.21 GB	gemma-3-4B-it-qat
llama-3.2-1b-instruct	hugging-quants	1B	Q8_0	1.31 GB	Llama-3.2-1B-Instruct
qwen3-8b	lmstudio-community	8B	Q3_K_L	4.43 GB	Qwen3-8B
More to come, still in progress					