

SETUP LLM LOCALLY WITH LM STUDIO

Tutorial „Techniques for Using LLMs Effectively” at WAW ML 10



Content



- Overview
 - Technical requirements
 - LM Studio
- LM Studio Setup
 - Download & install LM Studio
 - Finding and downloading models
 - LM Studio local server

INTRODUCTION

Warning

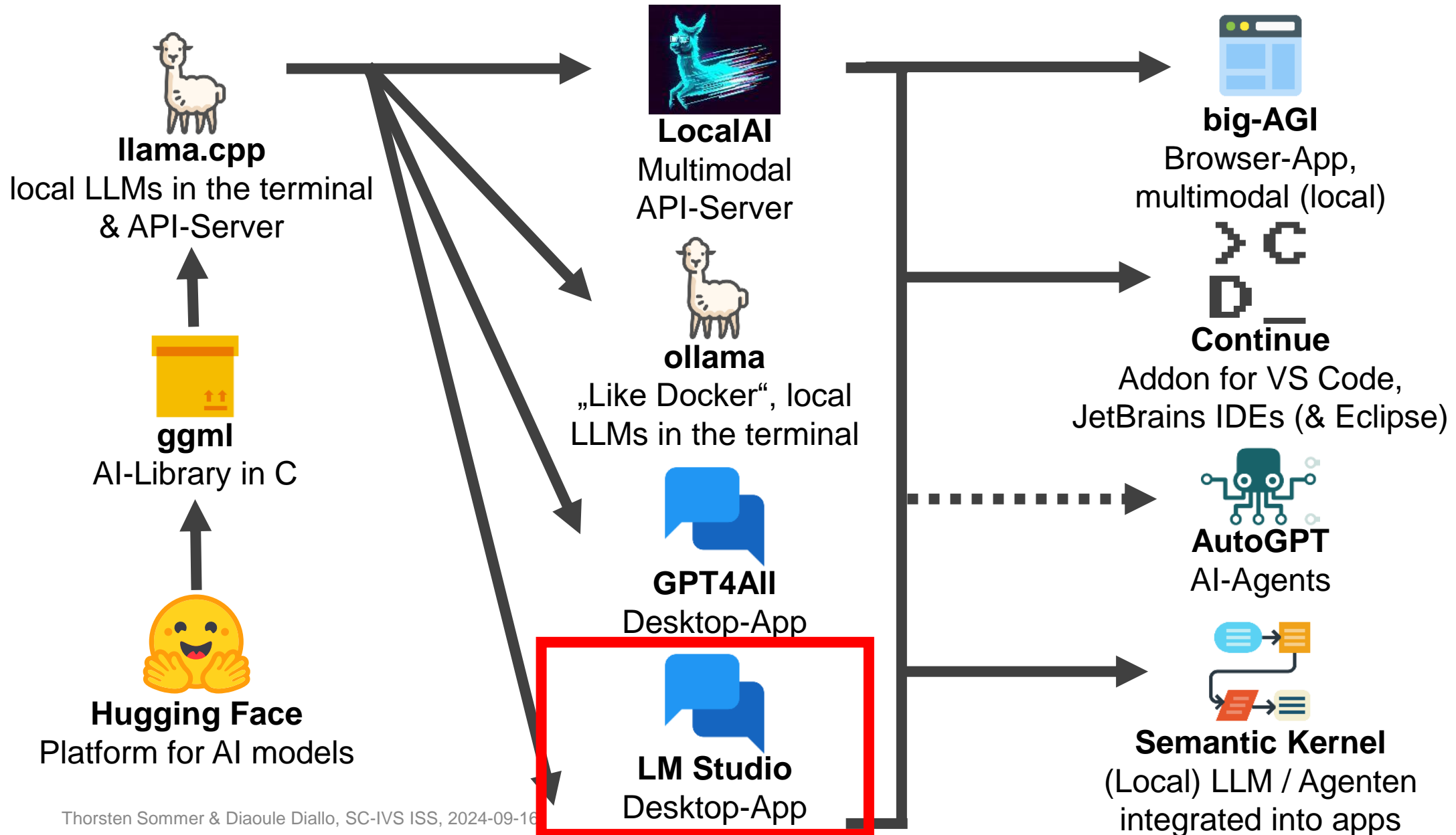
The background of the slide is a photograph of several wooden dominoes in various colors (tan, blue, red, white) arranged in a line on a grey surface. A hand is visible in the center, tipping one of the dominoes, which has caused a chain reaction where several dominoes have already fallen. This visual metaphor represents the risk of a cascading failure or system overload.

Please do not start any downloads during the event! We would overload the WIFI.

Please install before or after the WAW...

<https://pixabay.com/photos/balance-domino-business-risk-6815204/>

Overview of LLM Frameworks



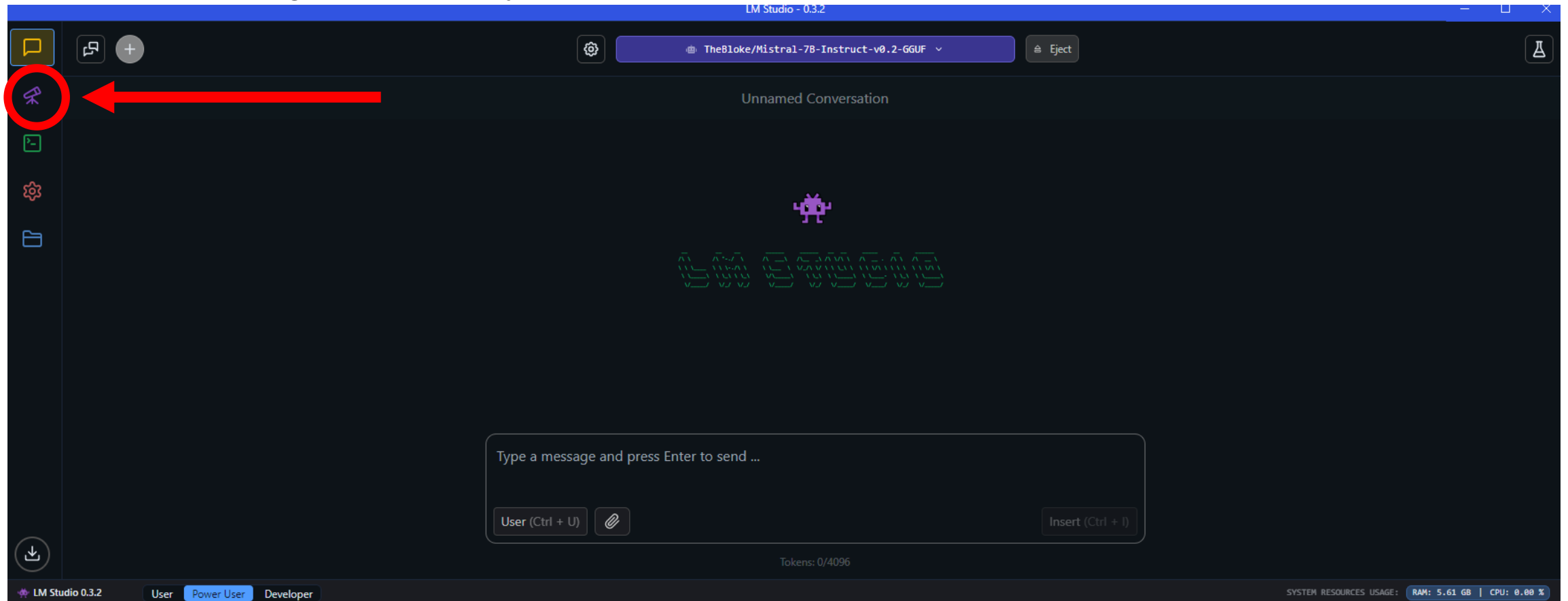
SETTING UP LM STUDIO

Download and install LM Studio



Download from <https://lmstudio.ai> (available for macOS, Linux and Windows), then run the setup.

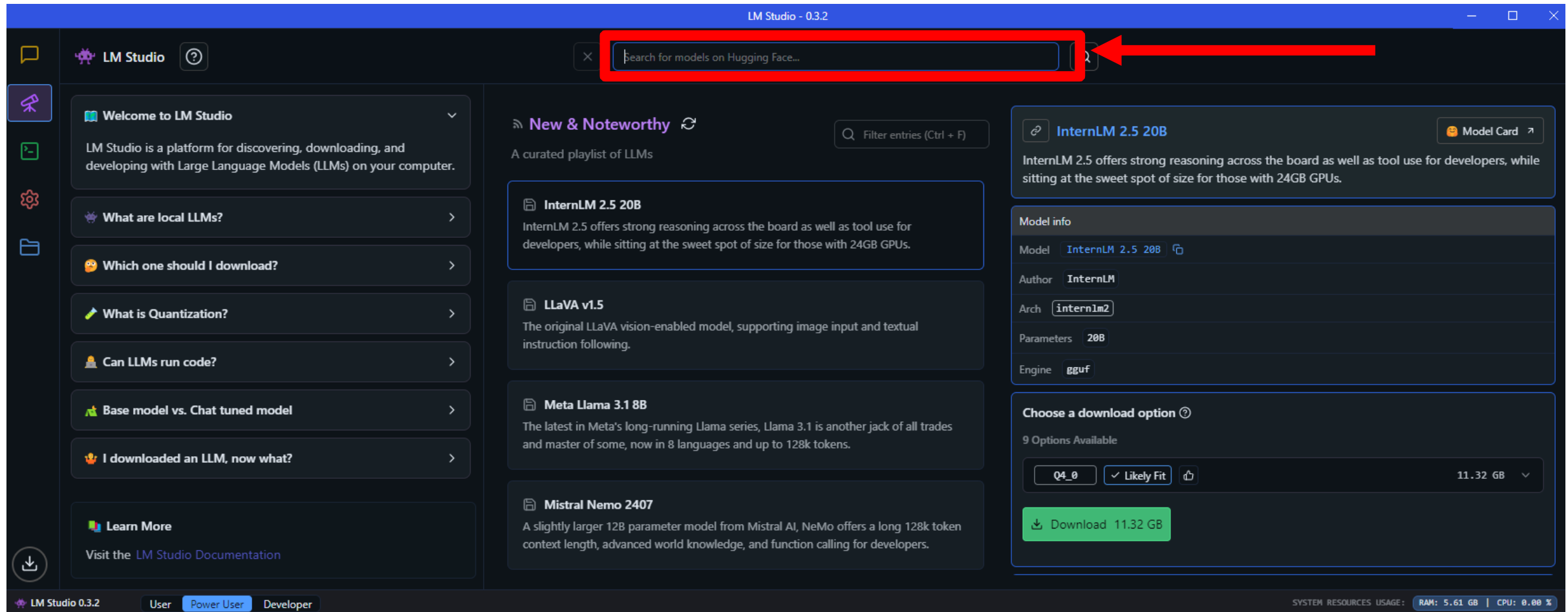
After installation, go to „Discovery“



Models: Setting up a model

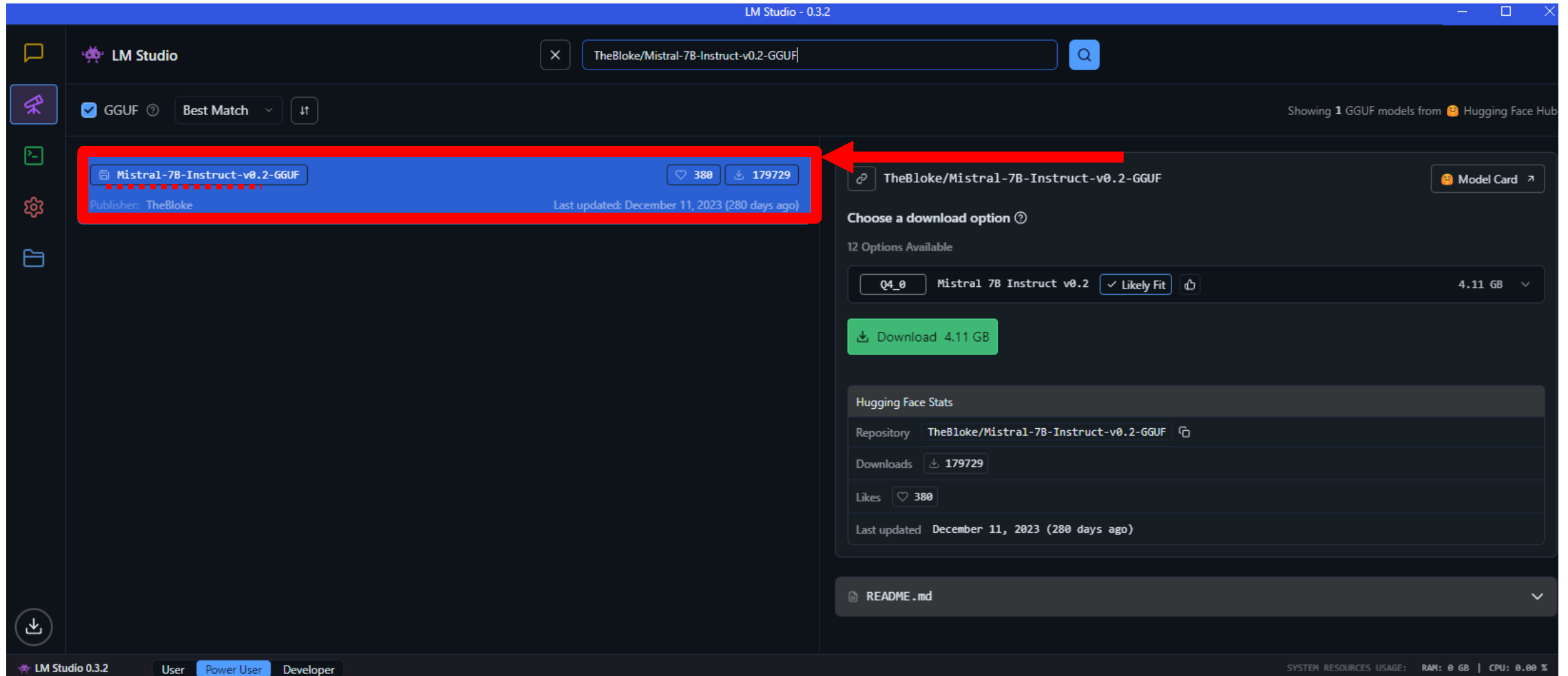


Enter the name of the model you want. **We will use “TheBloke/Mistral-7B-Instruct-v0.2-GGUF”.**



Models: Setting up a model

Select the model.

The screenshot shows the LM Studio application window. At the top, the search bar contains the text 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF'. Below the search bar, the 'GGUF' filter is selected, and the results are sorted by 'Best Match'. A red rectangle highlights the first search result, 'Mistral-7B-Instruct-v0.2-GGUF', which has 380 likes and 179,729 downloads. A red arrow points from this result to the right-hand panel. The right-hand panel displays the model's details, including the repository name 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF', a 'Model Card' link, and a section titled 'Choose a download option' with 12 options available. The selected option is 'Q4_0 Mistral 7B Instruct v0.2' with a 'Likely Fit' badge and a size of 4.11 GB. A green 'Download 4.11 GB' button is visible. Below this, the 'Hugging Face Stats' section shows the repository name, download count (179,729), likes (380), and last updated date (December 11, 2023). At the bottom, there is a 'README.md' section.

Models: Setting up a model



The list of variants (different quantizations) can be found at the right:

The screenshot shows the LM Studio 0.3.2 interface. On the left, a sidebar contains icons for chat, search, settings, and a file manager. The main area is divided into two panels. The left panel displays a search result for 'Mistral-7B-Instruct-v0.2-GGUF' by 'TheBloke', with 380 likes and 179,729 downloads. The right panel shows the 'Choose a download option' section for 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF'. A red rectangle highlights a list of 13 available options. The first option, 'Q4_0 Mistral 7B Instruct v0.2', is selected with a 'Likely Fit' badge and has a size of 4.11 GB. A red arrow points to this option from the right. Below the list is a green 'Download 4.11 GB' button. At the bottom, the 'Hugging Face Stats' section shows repository details, downloads, likes, and the last update date. The footer includes the LM Studio version, user type (Power User), and system resource usage (RAM: 0 GB, CPU: 0.00 %).

Models: Setting up a model



We are looking for the quantization **Q5_K_M**:

If you have 32GB of RAM, you can also install a bigger version, like Q8.

The screenshot shows the LM Studio interface with the search bar containing "TheBloke/Mistral-7B-Instruct-v0.2-GGUF". The left sidebar shows the "GGUF" filter selected. The main panel displays the model card for "TheBloke/Mistral-7B-Instruct-v0.2-GGUF" with 380 likes and 179,729 downloads. The "Choose a download option" section shows 12 options available. A red arrow points to the "Q5_K_M" option, which is highlighted with a red box. The system resources usage at the bottom shows RAM: 0 GB and CPU: 0.00 %.

Quantization	Model Name	Size (GB)
Q3_K_M	Mistral 7B Instruct v0.2	3.52 GB
Q3_K_L	Mistral 7B Instruct v0.2	3.82 GB
Q4_0	Mistral 7B Instruct v0.2	4.11 GB
Q4_K_S	Mistral 7B Instruct v0.2	4.14 GB
Q4_K_M	Mistral 7B Instruct v0.2	4.37 GB
Q5_0	Mistral 7B Instruct v0.2	5.00 GB
Q5_K_S	Mistral 7B Instruct v0.2	5.00 GB
Q5_K_M	Mistral 7B Instruct v0.2	5.13 GB
Q6_K	Mistral 7B Instruct v0.2	5.94 GB

Models: Setting up a model



We are looking for the quantization **Q5_K_M**:

If you have 32GB of RAM, you can also install a bigger version, like Q8.

The screenshot shows the LM Studio application window. At the top, the search bar contains 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF'. Below the search bar, the 'GGUF' filter is selected, and 'Best Match' is chosen. The main list on the left shows the selected model with 380 likes and 179,729 downloads. The right panel, titled 'Choose a download option', shows 12 options available. The 'Q5_K_M' option for 'Mistral 7B Instruct v0.2' is selected, marked as 'Likely Fit' with a size of 5.13 GB. A green 'Download' button is visible. Below this, the 'Hugging Face Stats' section shows the repository name, download count, likes, and last update date. A 'README.md' file is also listed at the bottom of the right panel. The bottom status bar shows 'LM Studio 0.3.2' and system resources usage.

Models: Setting up a model

Download the model:

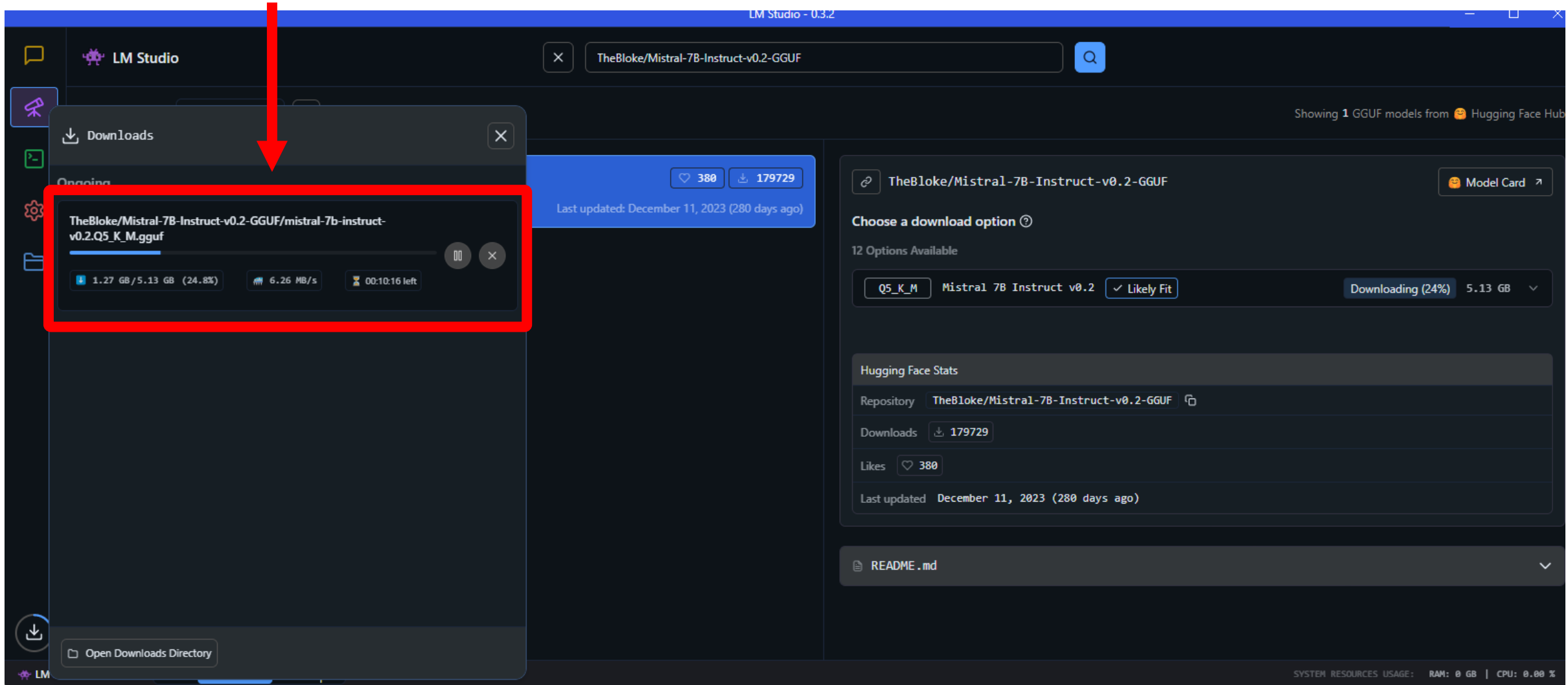


A screenshot of the LM Studio application window. The title bar reads 'LM Studio - 0.3.2'. The interface is dark-themed. At the top, there's a search bar containing 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF'. Below the search bar, there are filters for 'GGUF' and 'Best Match'. A list of models is shown, with 'Mistral-7B-Instruct-v0.2-GGUF' by 'TheBloke' highlighted. To the right of this model, there are buttons for '380' likes and '179729' downloads. Below the list, a detailed view of the selected model is shown. It includes the model name 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF', a 'Model Card' link, and a section titled 'Choose a download option'. Under this section, it says '12 Options Available' and shows a dropdown menu with 'Q5_K_M Mistral 7B Instruct v0.2' selected, with a 'Likely Fit' badge and '5.13 GB' size. A red arrow points from the search bar down to the 'Download 5.13 GB' button, which is highlighted with a red rectangle. Below the download button, there's a 'Hugging Face Stats' section showing repository details, downloads (179729), likes (380), and last updated date (December 11, 2023). At the bottom, there's a 'README .md' section. The bottom status bar shows 'LM Studio 0.3.2', user roles 'User', 'Power User', and 'Developer', and system resources usage: 'RAM: 0 GB | CPU: 0.00 %'.

Models: Setting up a model



Wait until the download has been completed:



The screenshot displays the LM Studio interface. A red arrow points to a download progress window for the model 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF/mistral-7b-instruct-v0.2.Q5_K_M.gguf'. The progress bar indicates 1.27 GB / 5.13 GB (24.8%) downloaded, with a download speed of 6.26 MB/s and 00:10:16 left. The background shows the model card for 'TheBloke/Mistral-7B-Instruct-v0.2-GGUF' with 179729 downloads and 380 likes. The system resources usage at the bottom right shows RAM: 0 GB and CPU: 0.00 %.

LM Studio Server Setup



Click on “Local Server”

LM Studio - 0.3.2

Models

My Models

You have 2 local models, taking up 10.05 GB of disk space.

Models Directory C:\Users\schu_p11\.cache\lm-studio\models

Arch	Params	Publisher	LLM	Quant	Size	Date Modified	Actions
llama	8B	TheBloke	Mistral-7B-Instruct-v0.2-GGUF / mistral-7b-instruct-v0.2.Q5_K_M.gguf	Q5_K_M	5.13 GB	September 16, 2024 (today)	
llama	8B	lmstudio-community	Meta-Llama-3.1-8B-Instruct-GGUF / Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf	Q4_K_M	4.92 GB	August 28, 2024 (20 days ago)	

Filter models... (Ctrl + F)

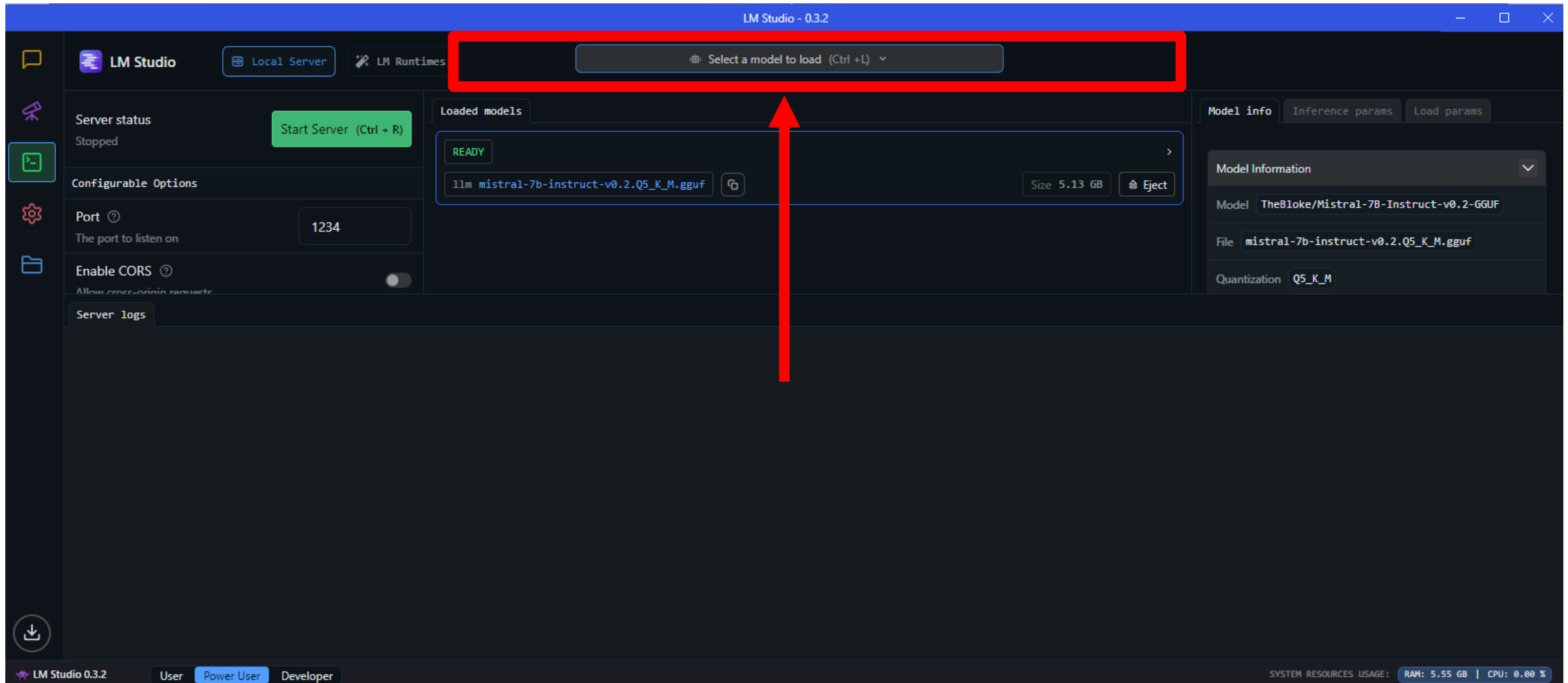
LM Studio 0.3.2

User Power User Developer

SYSTEM RESOURCES USAGE: RAM: 0 GB | CPU:

LM Studio Server Setup

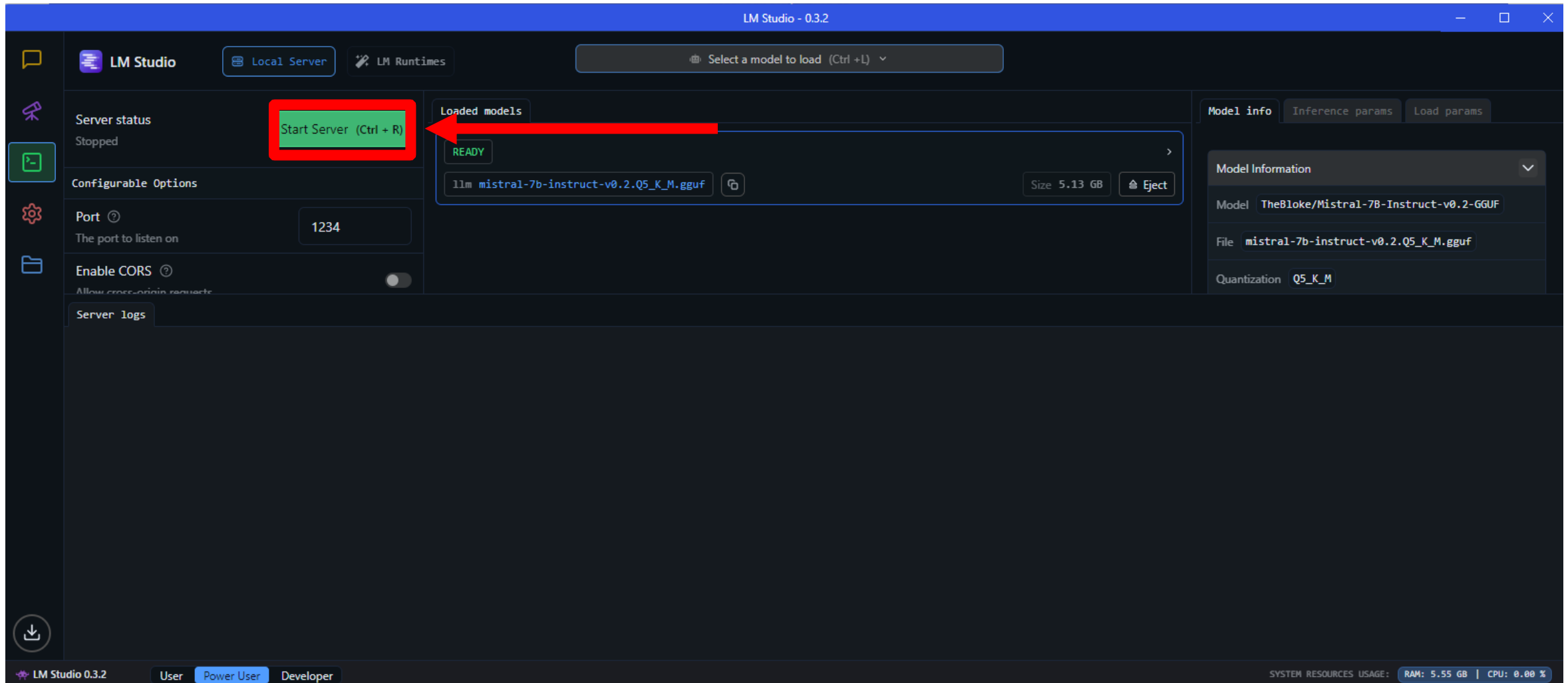
Select the appropriate model & wait until it has been loaded:



LM Studio Server Setup



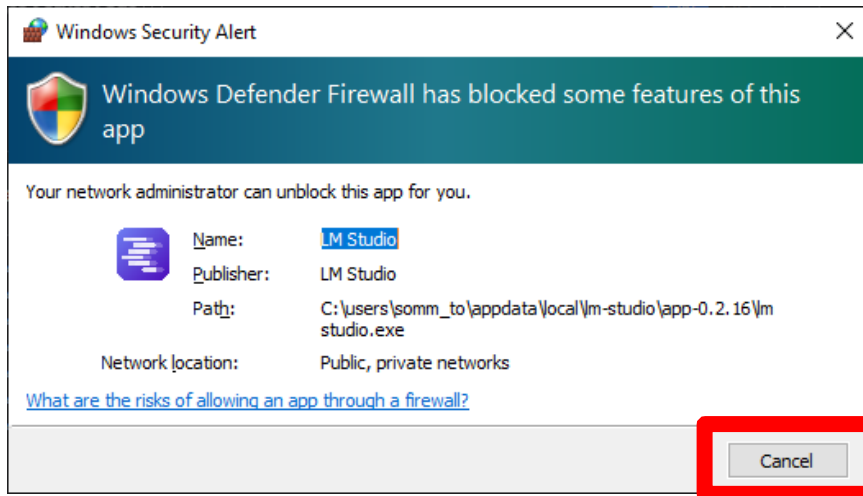
Start server:



LM Studio Server Setup



Close the warning from the Windows Firewall with Cancel:



Now you are ready!