

Lab 04

Installing LMStudio



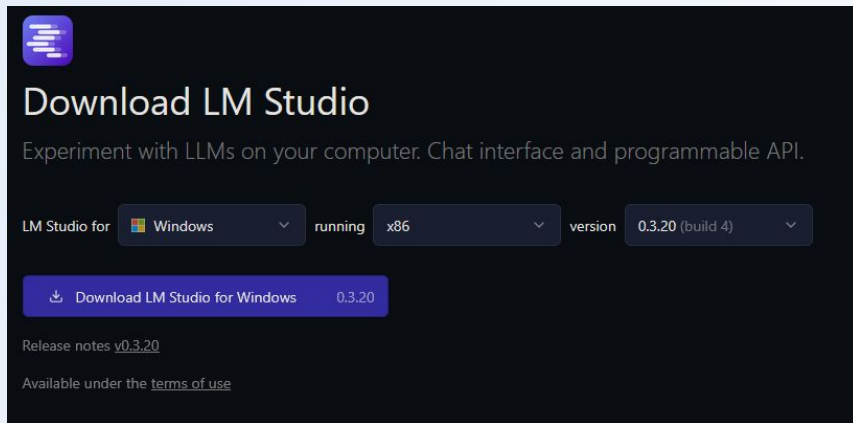
Lab Objectives

- Learn how to install LMStudio on your personal laptop
- Configure the various parameters necessary to leverage built-in hardware on your laptop
- Configure a OpenAPI compatible server using the built-in tools
- Run a test query against your own model



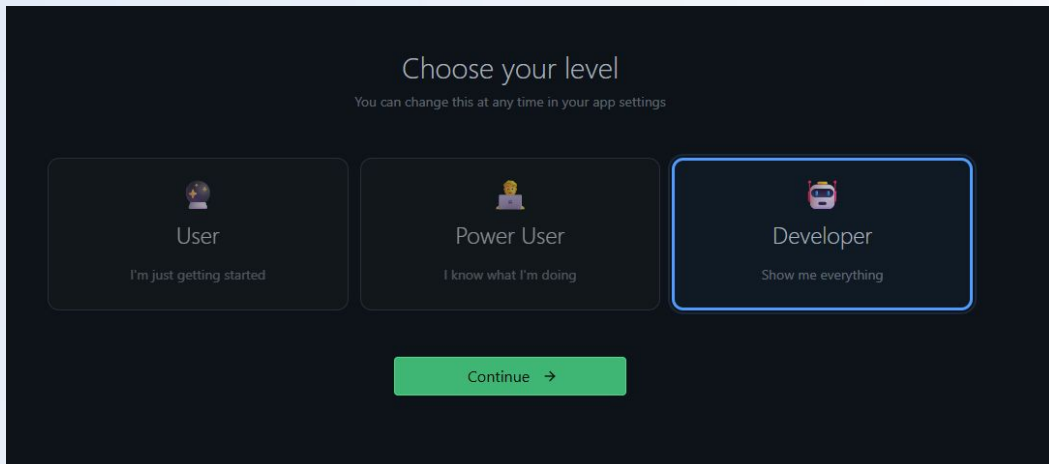
LMStudio.ai

That is both the website and the title of this slide. Download the appropriate executable for your host. On Windows, LMStudio will automatically install itself, launch, and make a desktop shortcut with little intervention.

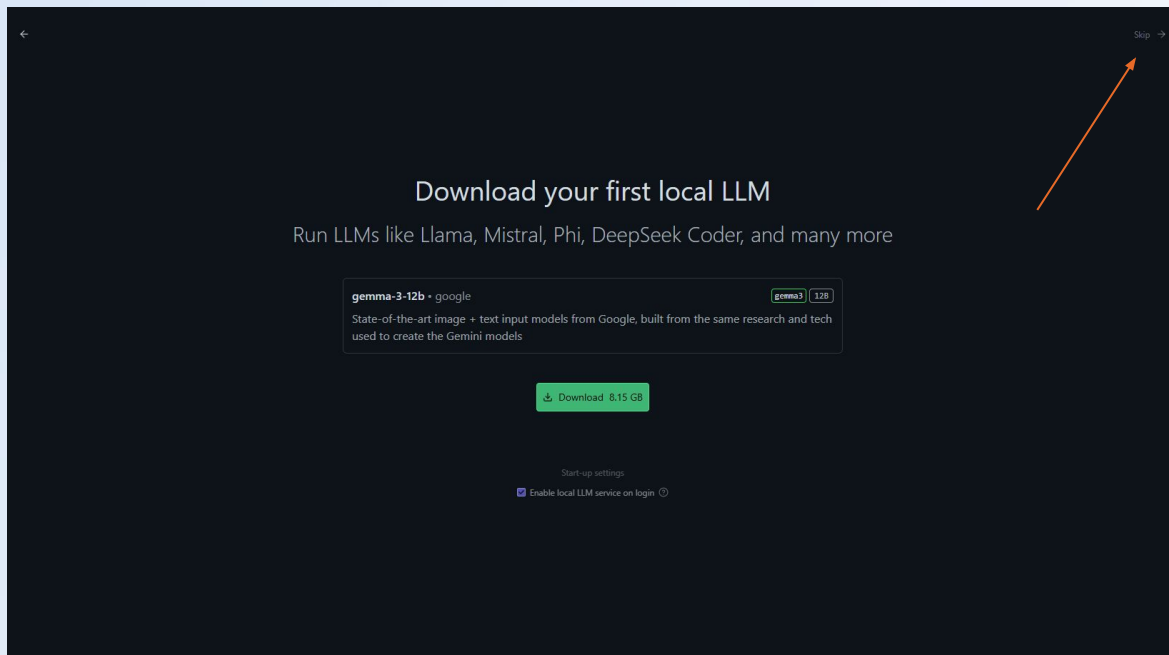


Setup

Once you install the program and launch it, start clicking through until you get to this screen. Make sure to select `Developer`.

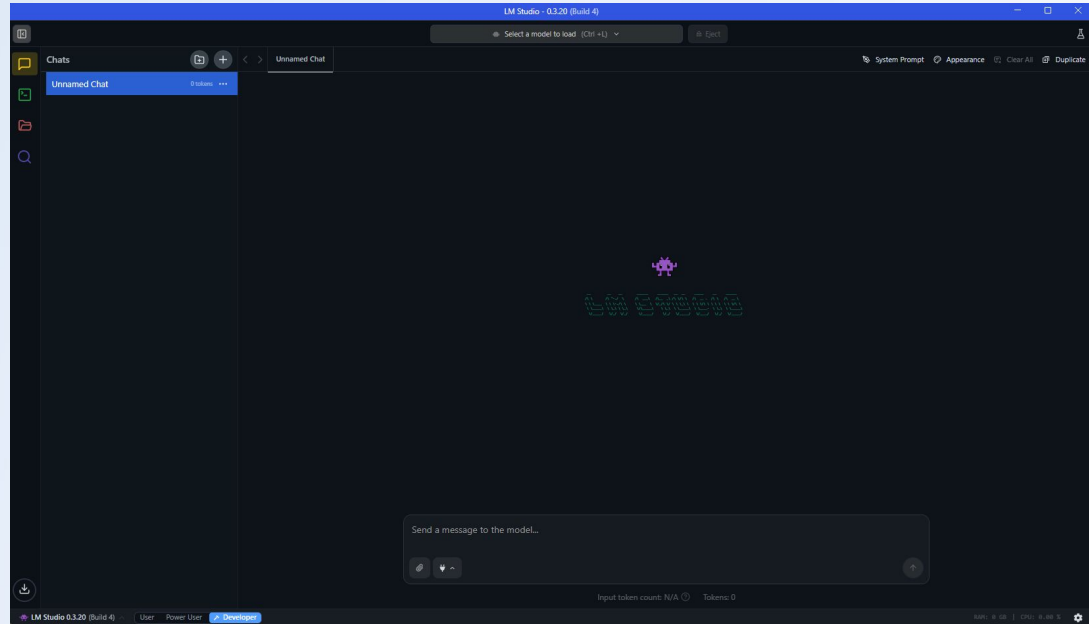


Skip downloading a model for now



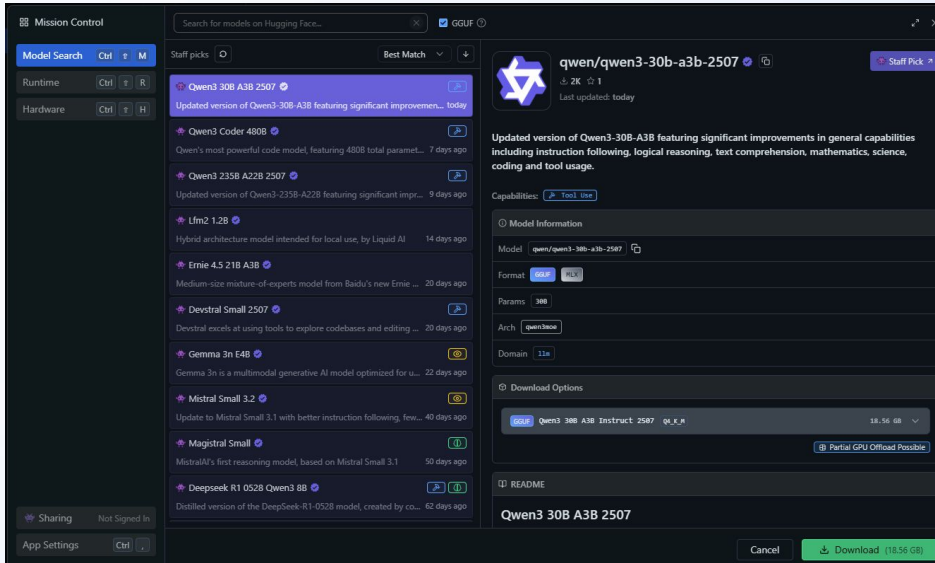
Home Page

The default home page is now a chat box. There is no model loaded, so this is currently not functional.



This is essentially a dedicated search page for models. Here, you can choose the quantization of the specific model you are interested in and download that.

Search Page

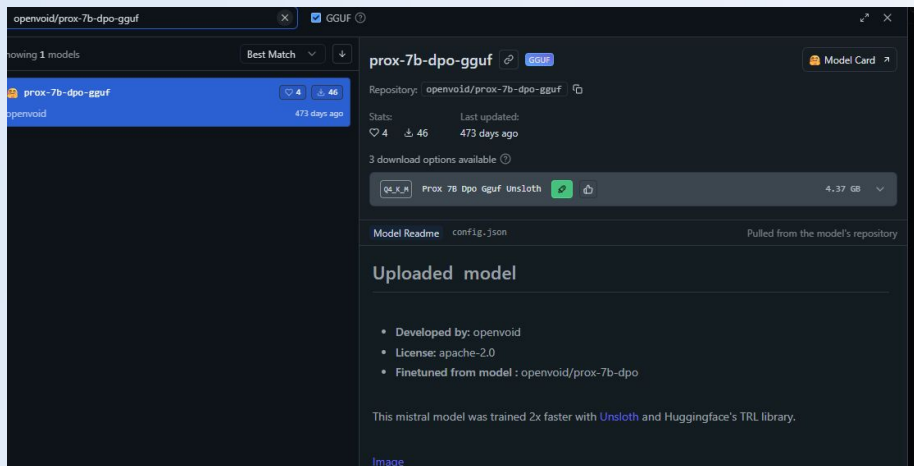


Downloading a model

Search the model “openvoid/prox-7b-dpo-gguf”. This is the model we will be using in the class.

Notice towards the bottom, there is a little helpful popup explaining the difference between these files.

For this lab, you should have already copied the Q4_K_M model discussed previously to:
`C:\Users\%USER\.cache\lm-studio\models`



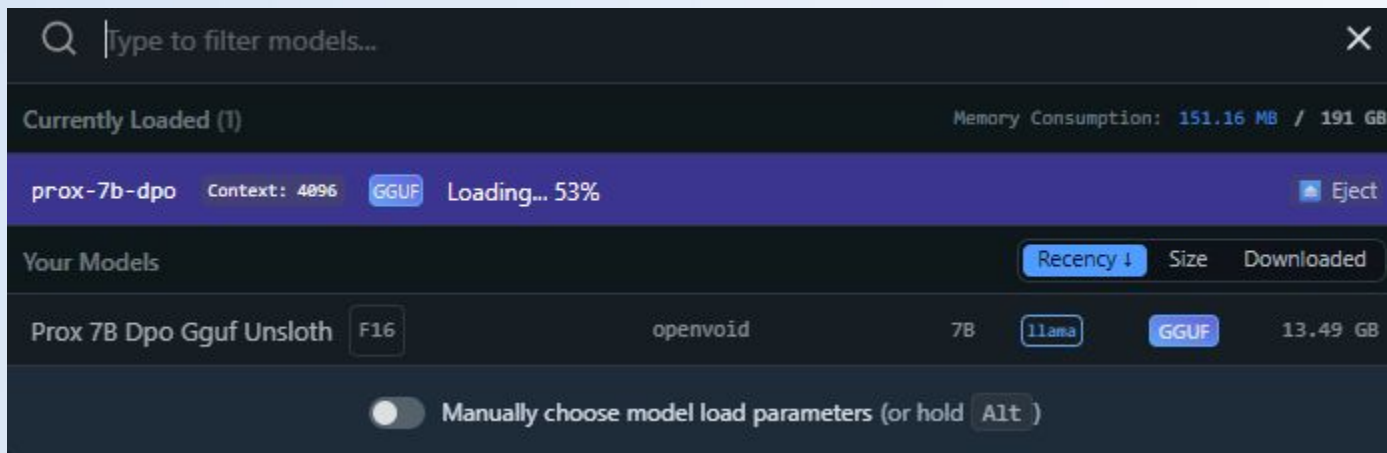
Helpful Hints by LMStudio

What are the various options?

Quantization is a technique to compress the model weights and reduce the model size. The lower the quantization level, the smaller the model size. Lower quantization levels may reduce the model accuracy since the model weights are compressed to fit the smaller size.

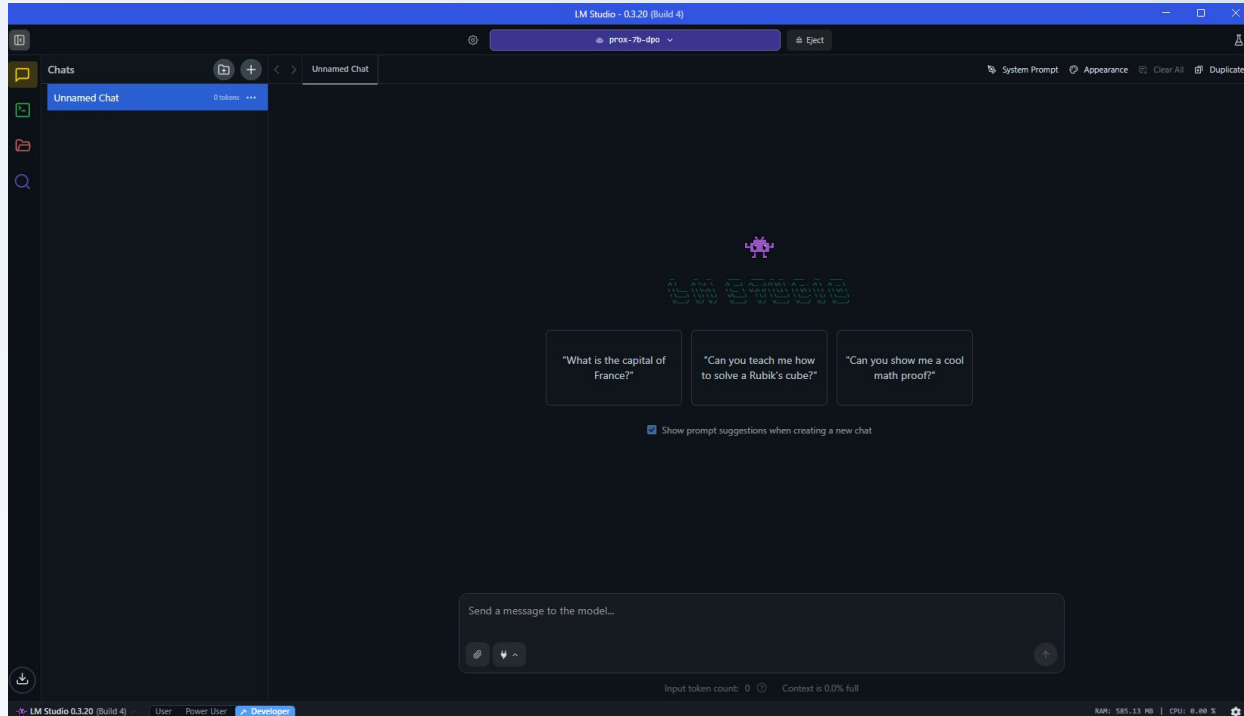
Loading a Model

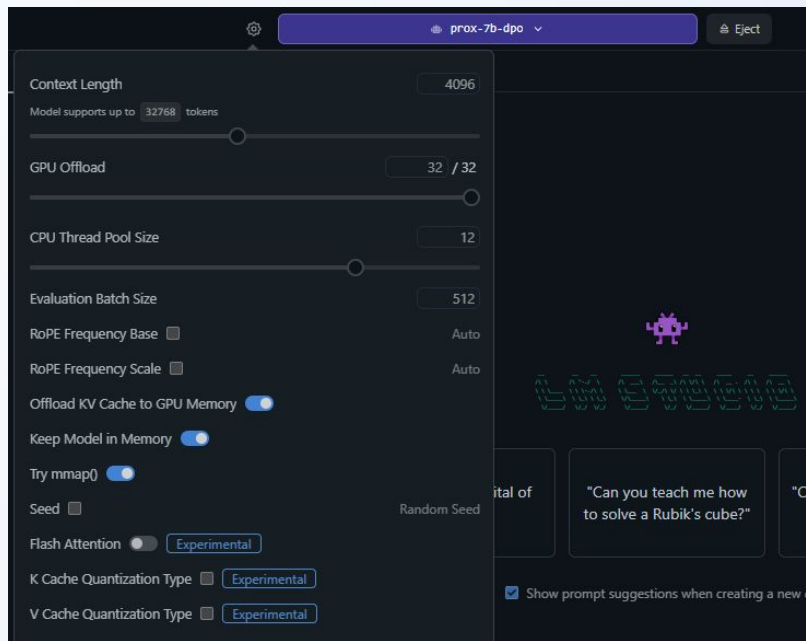
To Load a model, Simply select the model you've just downloaded or copied over. LMStudio will automatically load the model.



Talking with the Devil - The Chat Interface

Try making your own query!

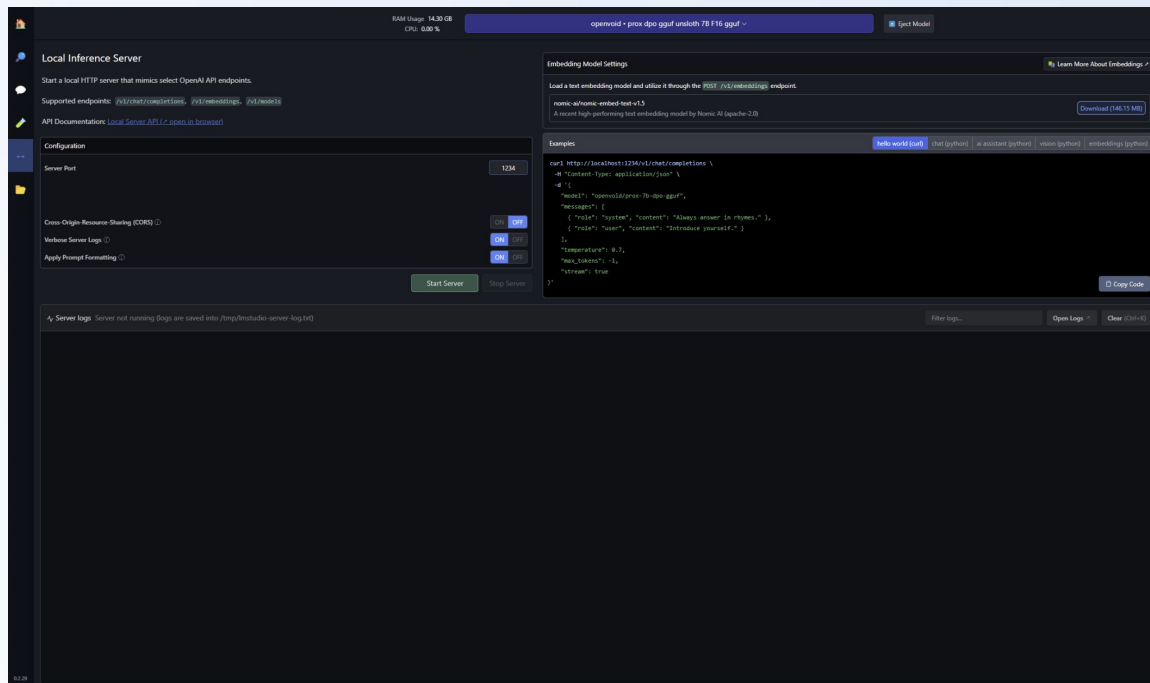




LMStudio helpfully provides little helping prompts for each configuration option to explain what they do.

Configuration Options

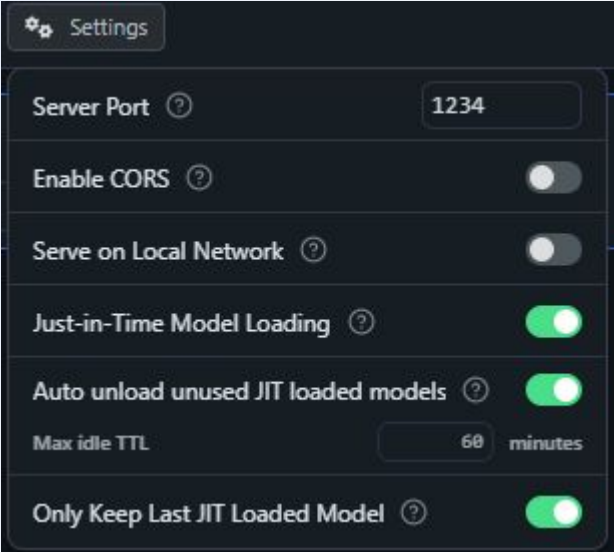
Local Server Configuration Page



This is an overview of the server configuration page. On the right side, you have the settings panel with multiple tabs for various parameters. On the main middle panel, you have a list of actively loaded models. LMStudio uses an API that mirrors ChatGPT/OpenAI's API.

Additional Server Settings

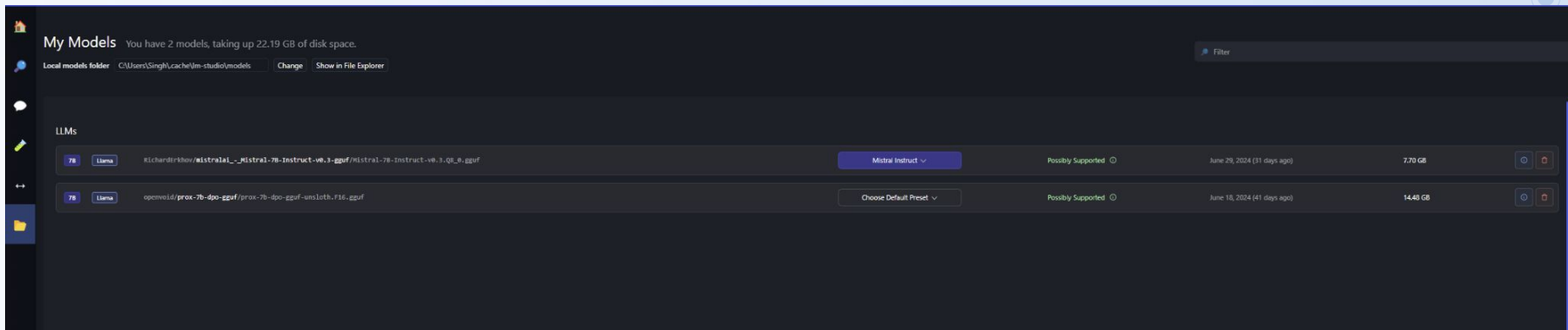
Next to the server start/stop slider, there is a small settings tab with some more settings. An important setting to disable here is Just-in-Time model loading - this can cause severe delay on responses with larger models.



The screenshot shows a dark-themed settings panel with a 'Settings' header and a gear icon. It contains several configuration options:

- Server Port**: A text input field containing the value '1234'.
- Enable CORS**: A toggle switch that is currently turned off (grey).
- Serve on Local Network**: A toggle switch that is currently turned off (grey).
- Just-in-Time Model Loading**: A toggle switch that is currently turned on (green).
- Auto unload unused JIT loaded models**: A toggle switch that is currently turned on (green).
- Max idle TTL**: A text input field containing '60' followed by the unit 'minutes'.
- Only Keep Last JIT Loaded Model**: A toggle switch that is currently turned on (green).

Local Model Browser



Lab End

