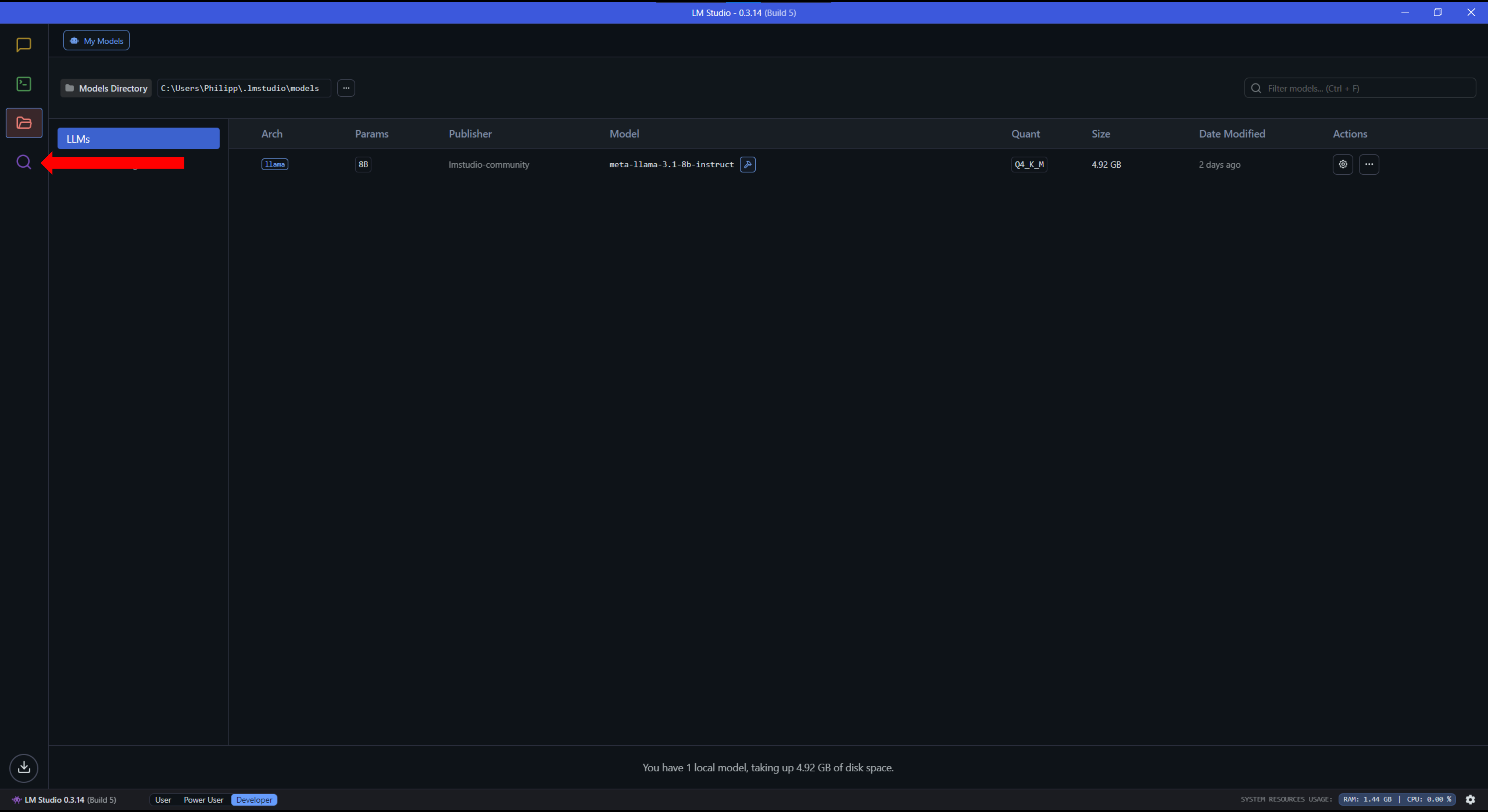
Setup

- Download LM Studio
  - https://lmstudio.ai/
- Download a *small* model
  - E.G., *meta-llma-3.1-8b-instruct*

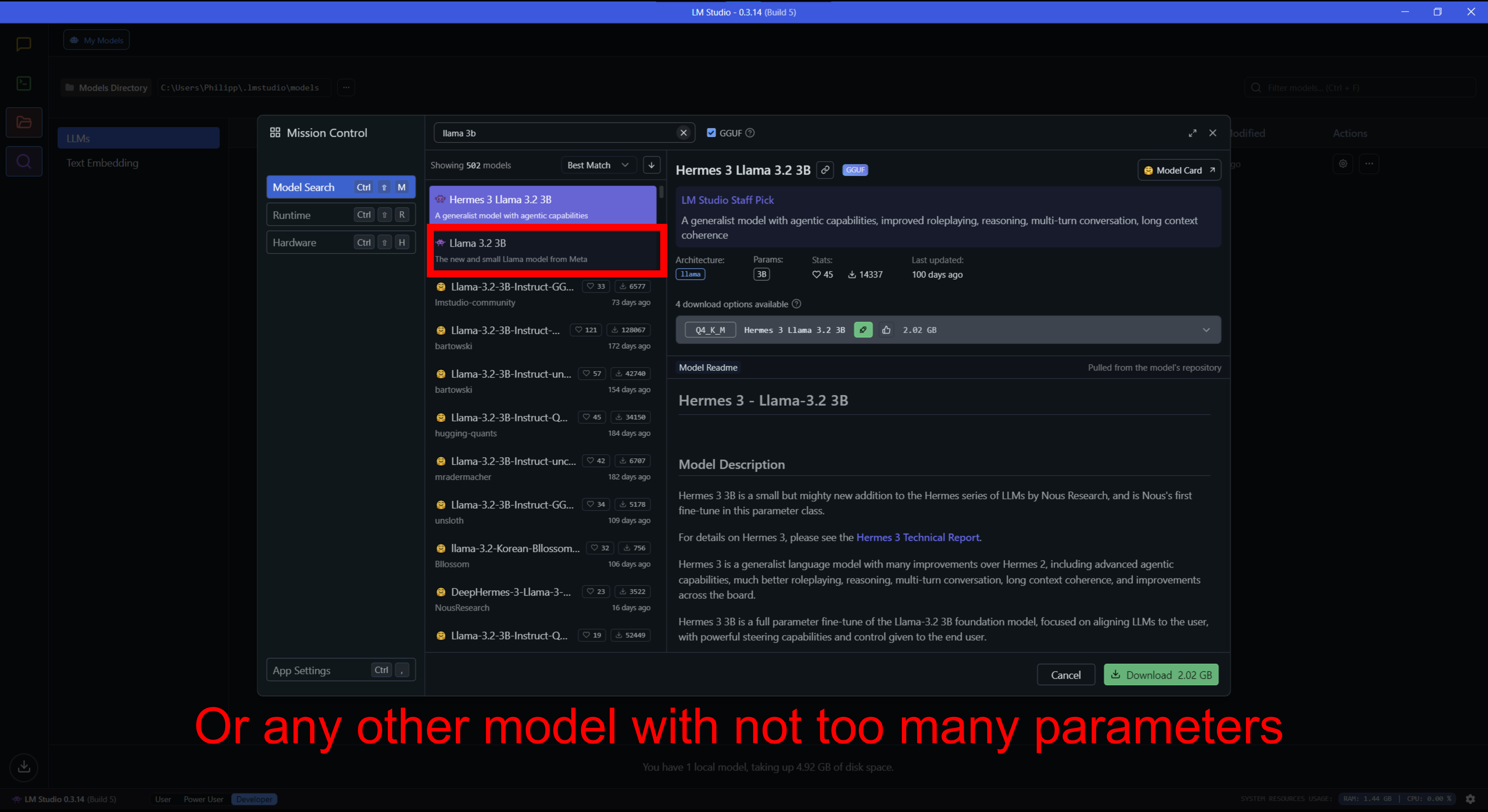

LM Studio - 0.3.14 (Build 5)
My Models
Models Directory
LLMs
Filter models... (Ctrl + F)
LM Studio 0.3.14 (Build 5)
User
Power User
Developer

C:\Users\Philipp\.lmstudio\models

Arch	Params	Publisher	Model	Quant	Size	Date Modified	Actions
llama	8B	lmstudio-community	meta-llama-3.1-8b-instruct	Q4_K_M	4.92 GB	2 days ago	

You have 1 local model, taking up 4.92 GB of disk space.

SYSTEM RESOURCES USAGE: RAM: 1.44 GB | CPU: 0.00 %



C:\Users\Philipp\.lmstudio\models

llama 3b

☑ GGUF

Showing 502 models

Model	Likes	Downloads
Hermes 3 Llama 3.2 3B A generalist model with agentic capabilities		
Llama 3.2 3B The new and small Llama model from Meta		
Llama-3.2-3B-Instruct-GG... lmstudio-community · 73 days ago	33	6577
Llama-3.2-3B-Instruct-... bartowski · 172 days ago	121	128067
Llama-3.2-3B-Instruct-un... bartowski · 154 days ago	57	42740
Llama-3.2-3B-Instruct-Q... hugging-quants · 184 days ago	45	34150
Llama-3.2-3B-Instruct-unc... mradermacher · 182 days ago	42	6707
Llama-3.2-3B-Instruct-GG... unsloth · 109 days ago	34	5178
llama-3.2-Korean-Bllossom... Bllossom · 106 days ago	32	756
DeepHermes-3-Llama-3-... NousResearch · 16 days ago	23	3522
Llama-3.2-3B-Instruct-Q...	19	52449

Hermes 3 Llama 3.2 3B GGUF

LM Studio Staff Pick

A generalist model with agentic capabilities, improved roleplaying, reasoning, multi-turn conversation, long context coherence

Architecture: llama

Params: 3B

Stats: ♡ 45 ⤓ 14337

Last updated: 100 days ago

4 download options available

Q4_K_M Hermes 3 Llama 3.2 3B 2.02 GB

Pulled from the model's repository

Hermes 3 - Llama-3.2 3B

Model Description

Hermes 3 3B is a small but mighty new addition to the Hermes series of LLMs by Nous Research, and is Nous's first fine-tune in this parameter class.

For details on Hermes 3, please see the Hermes 3 Technical Report.

Hermes 3 is a generalist language model with many improvements over Hermes 2, including advanced agentic capabilities, much better roleplaying, reasoning, multi-turn conversation, long context coherence, and improvements across the board.

Hermes 3 3B is a full parameter fine-tune of the Llama-3.2 3B foundation model, focused on aligning LLMs to the user, with powerful steering capabilities and control given to the end user.

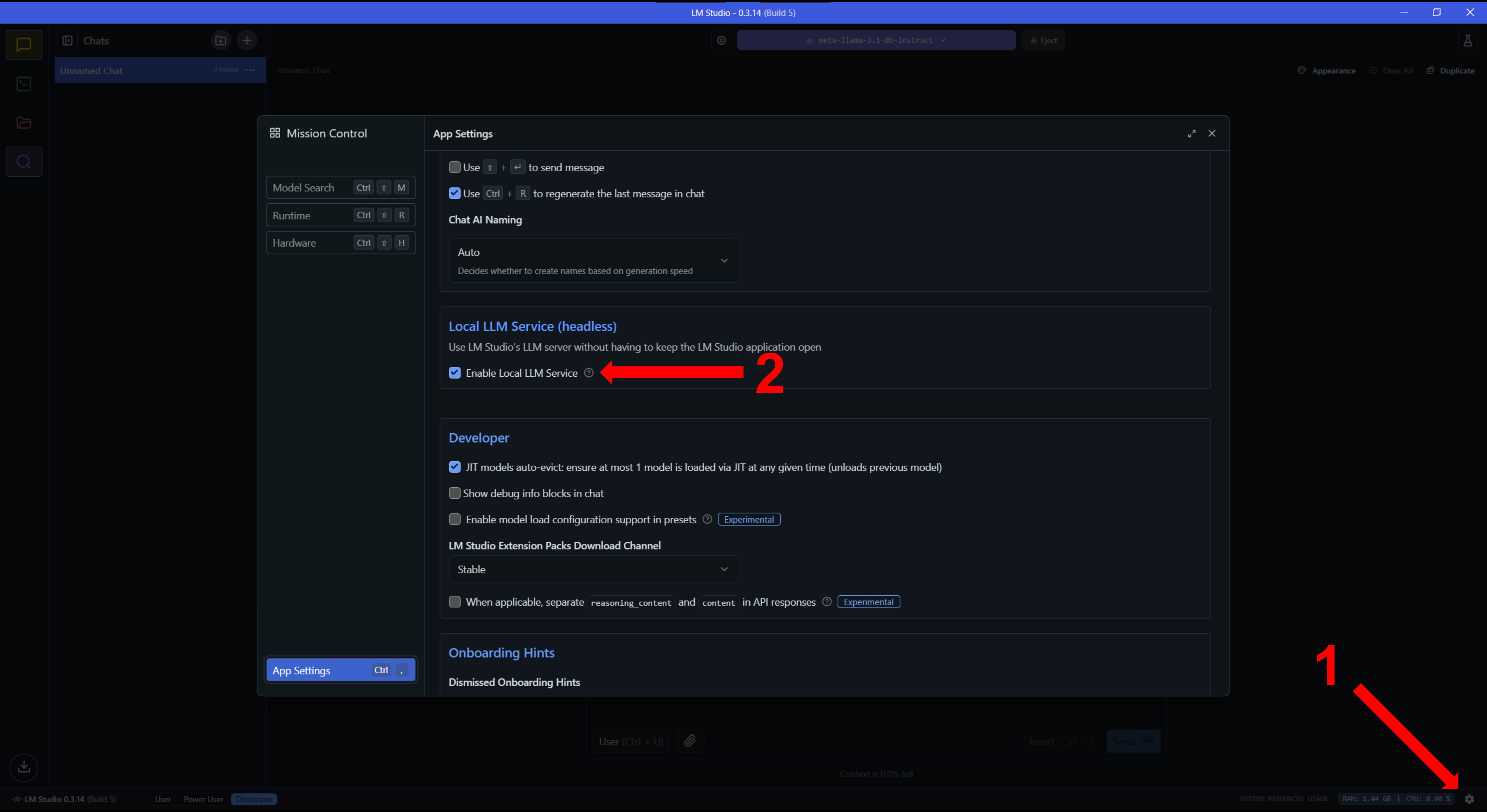
You have 1 local model, taking up 4.92 GB of disk space.

LM Studio 0.3.14 (Build 5) User Power User Developer

SYSTEM RESOURCES USAGE: RAM: 1.44 GB | CPU: 0.00 %

Or any other model with not too many parameters

Set up a server!



App Settings

□ Use ⇧ + ↵ to send message

☑ Use Ctrl + R to regenerate the last message in chat

Chat AI Naming

Auto
Decides whether to create names based on generation speed

Local LLM Service (headless)

Use LM Studio's LLM server without having to keep the LM Studio application open

☑ Enable Local LLM Service

Developer

☑ JIT models auto-evict: ensure at most 1 model is loaded via JIT at any given time (unloads previous model)

□ Show debug info blocks in chat

□ Enable model load configuration support in presets Experimental

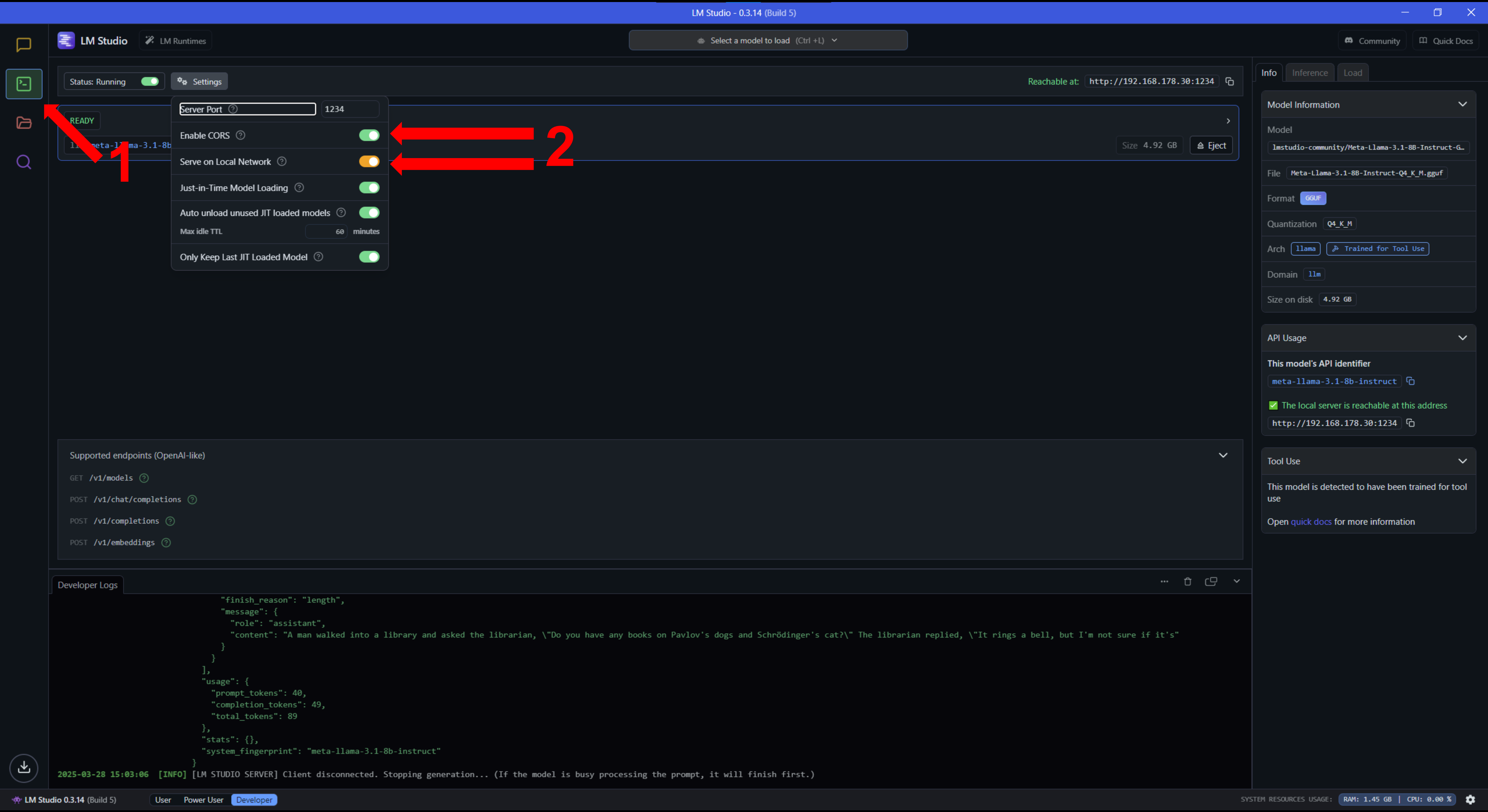
LM Studio Extension Packs Download Channel

Stable

□ When applicable, separate `reasoning_content` and `content` in API responses Experimental

Onboarding Hints

Dismissed Onboarding Hints



Status: Running

Reachable at: http://192.168.178.30:1234

Server Port: 1234

Enable CORS: on

Serve on Local Network: on

Just-in-Time Model Loading: on

Auto unload unused JIT loaded models: on

Max idle TTL: 60 minutes

Only Keep Last JIT Loaded Model: on

READY

llm meta-llama-3.1-8b

Size 4.92 GB

Model Information

Model: lmstudio-community/Meta-Llama-3.1-8B-Instruct-G…

File: Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf

Format: GGUF

Quantization: Q4_K_M

Arch: llama, Trained for Tool Use

Domain: llm

Size on disk: 4.92 GB

API Usage

This model's API identifier: meta-llama-3.1-8b-instruct

✅ The local server is reachable at this address: http://192.168.178.30:1234

Tool Use

This model is detected to have been trained for tool use

Open quick docs for more information

Supported endpoints (OpenAI-like)

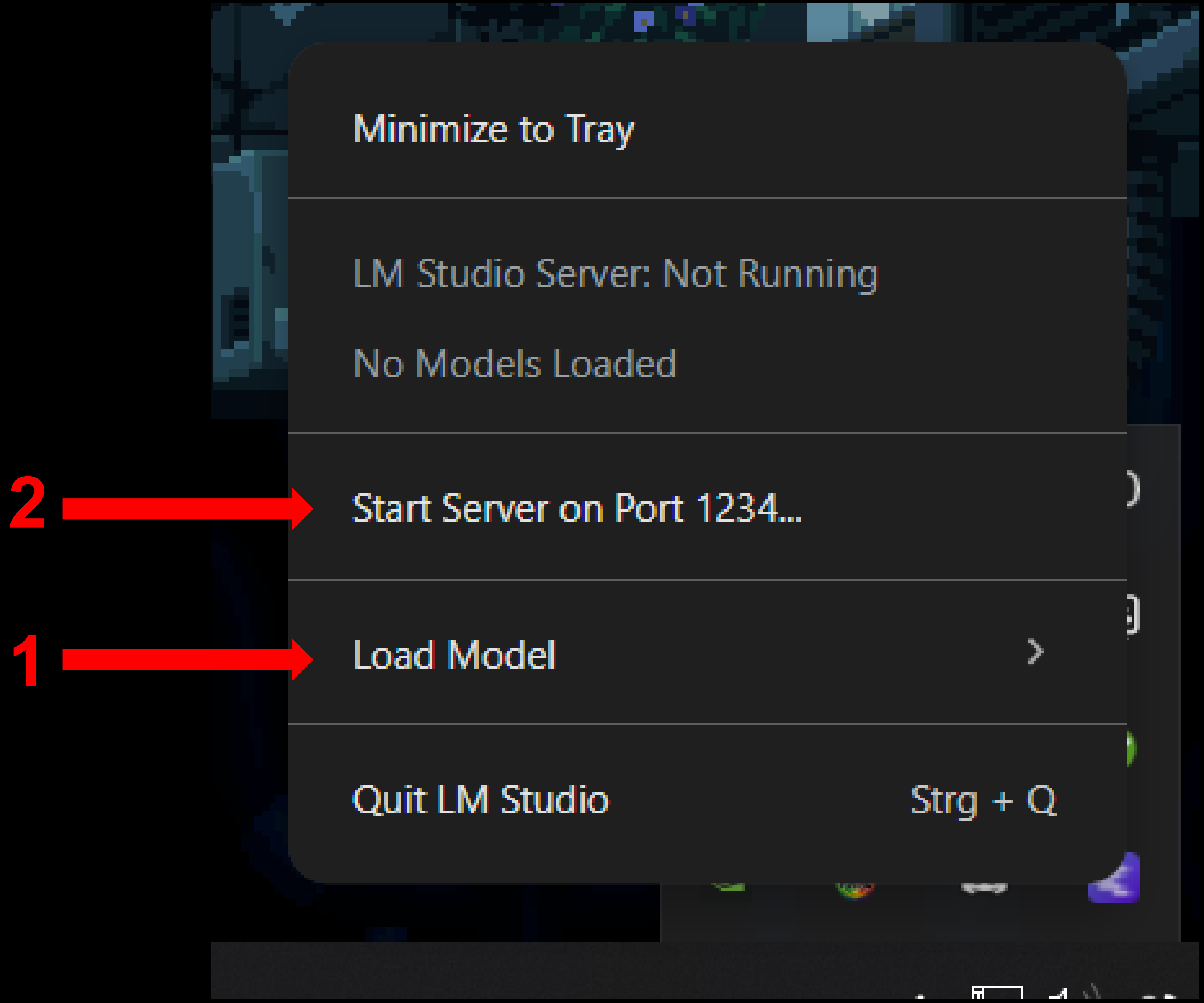
```
GET  /v1/models
POST /v1/chat/completions
POST /v1/completions
POST /v1/embeddings
```

Developer Logs

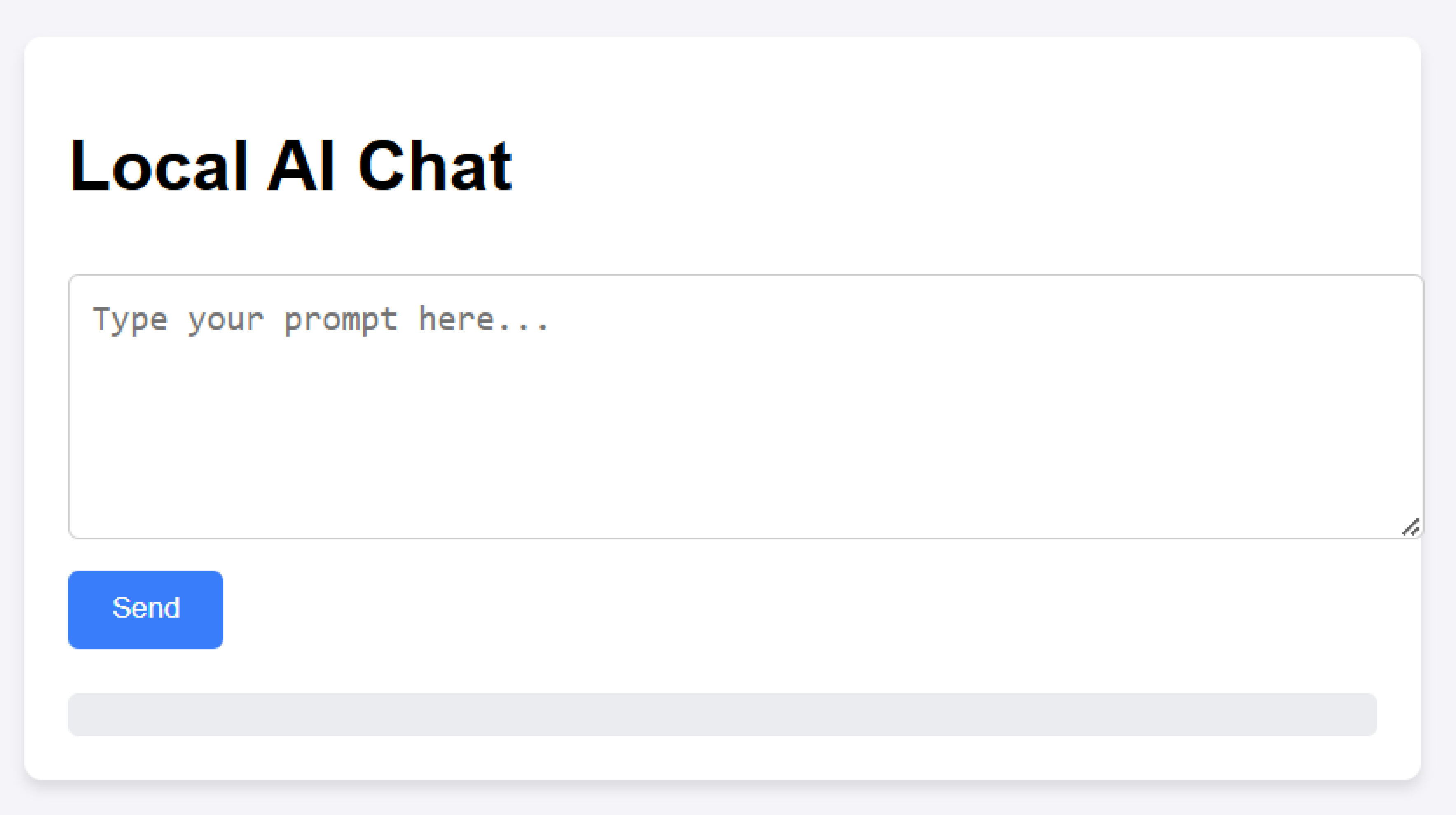
```
          "finish_reason": "length",
          "message": {
            "role": "assistant",
            "content": "A man walked into a library and asked the librarian, \"Do you have any books on Pavlov's dogs and Schrödinger's cat?\" The librarian replied, \"It rings a bell, but I'm not sure if it's"
          }
        }
      ],
      "usage": {
        "prompt_tokens": 40,
        "completion_tokens": 49,
        "total_tokens": 89
      },
      "stats": {},
      "system_fingerprint": "meta-llama-3.1-8b-instruct"
    }
2025-03-28 15:03:06  [INFO] [LM STUDIO SERVER] Client disconnected. Stopping generation... (If the model is busy processing the prompt, it will finish first.)
```

LM Studio 0.3.14 (Build 5)

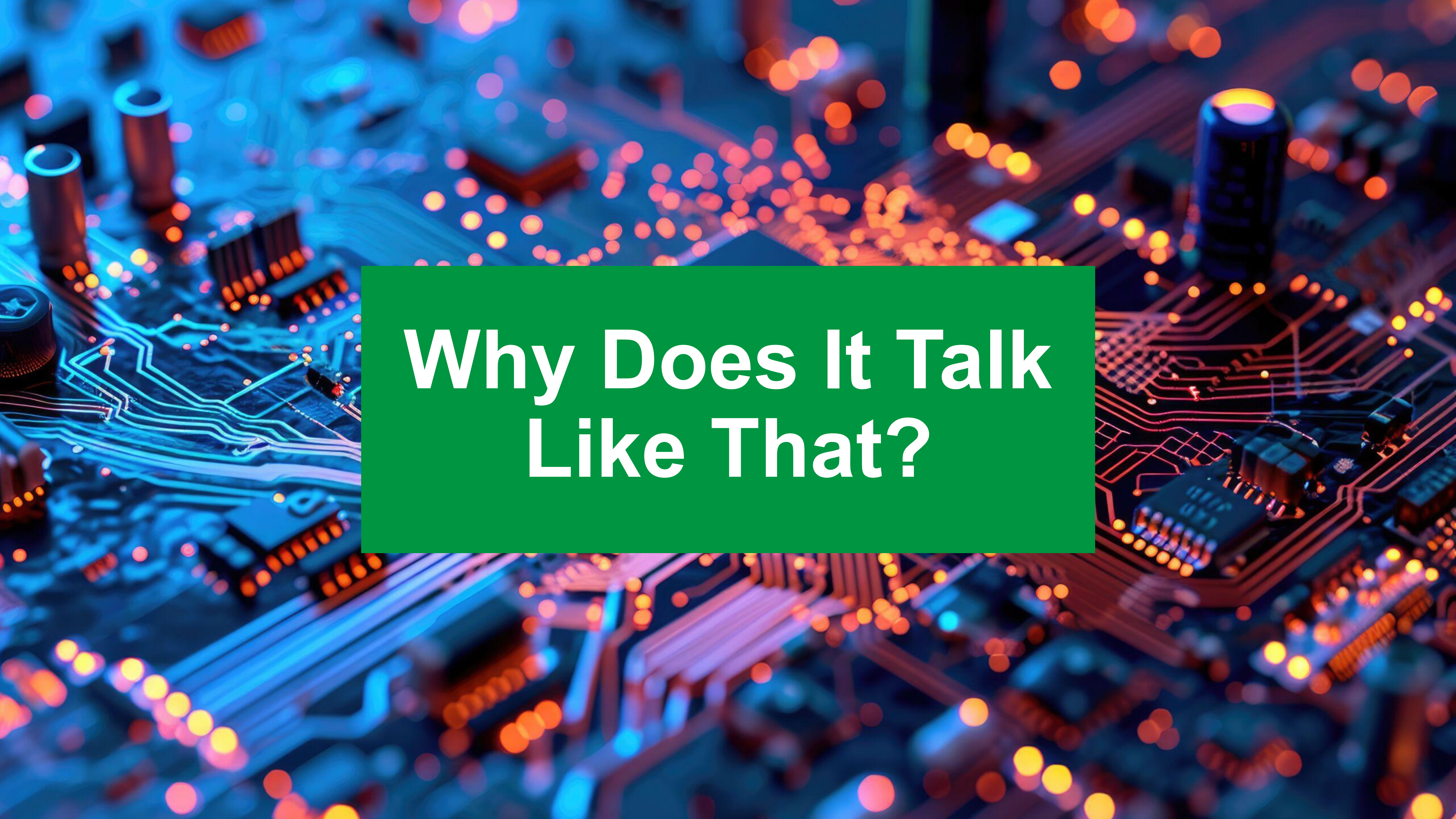
SYSTEM RESOURCES USAGE: RAM: 1.45 GB | CPU: 0.00 %


Minimize to Tray
LM Studio Server: Not Running
No Models Loaded
2
Start Server on Port 1234...
1
Load Model
Quit LM Studio
Strg + Q

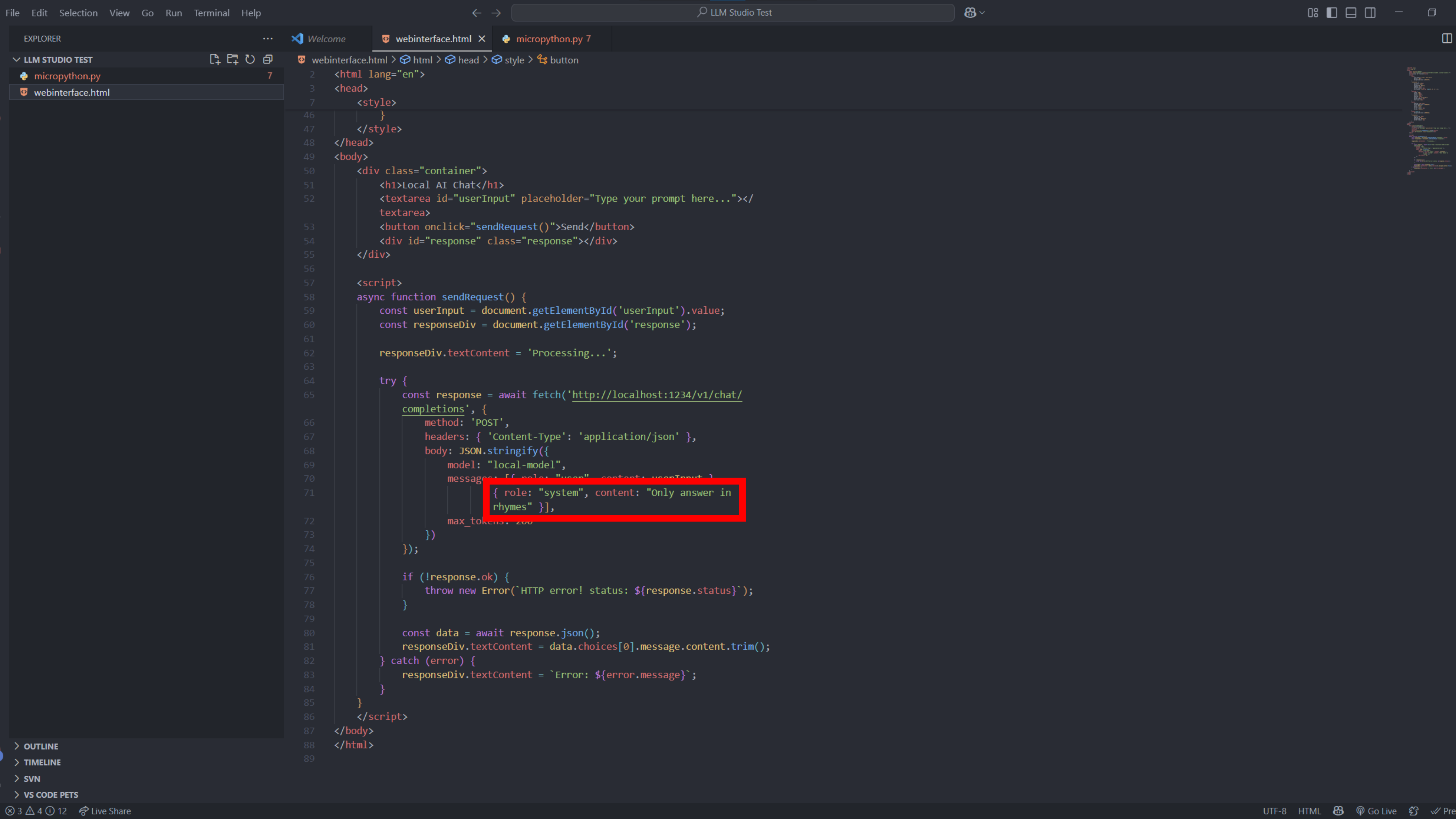
HTML Example



Local AI Chat

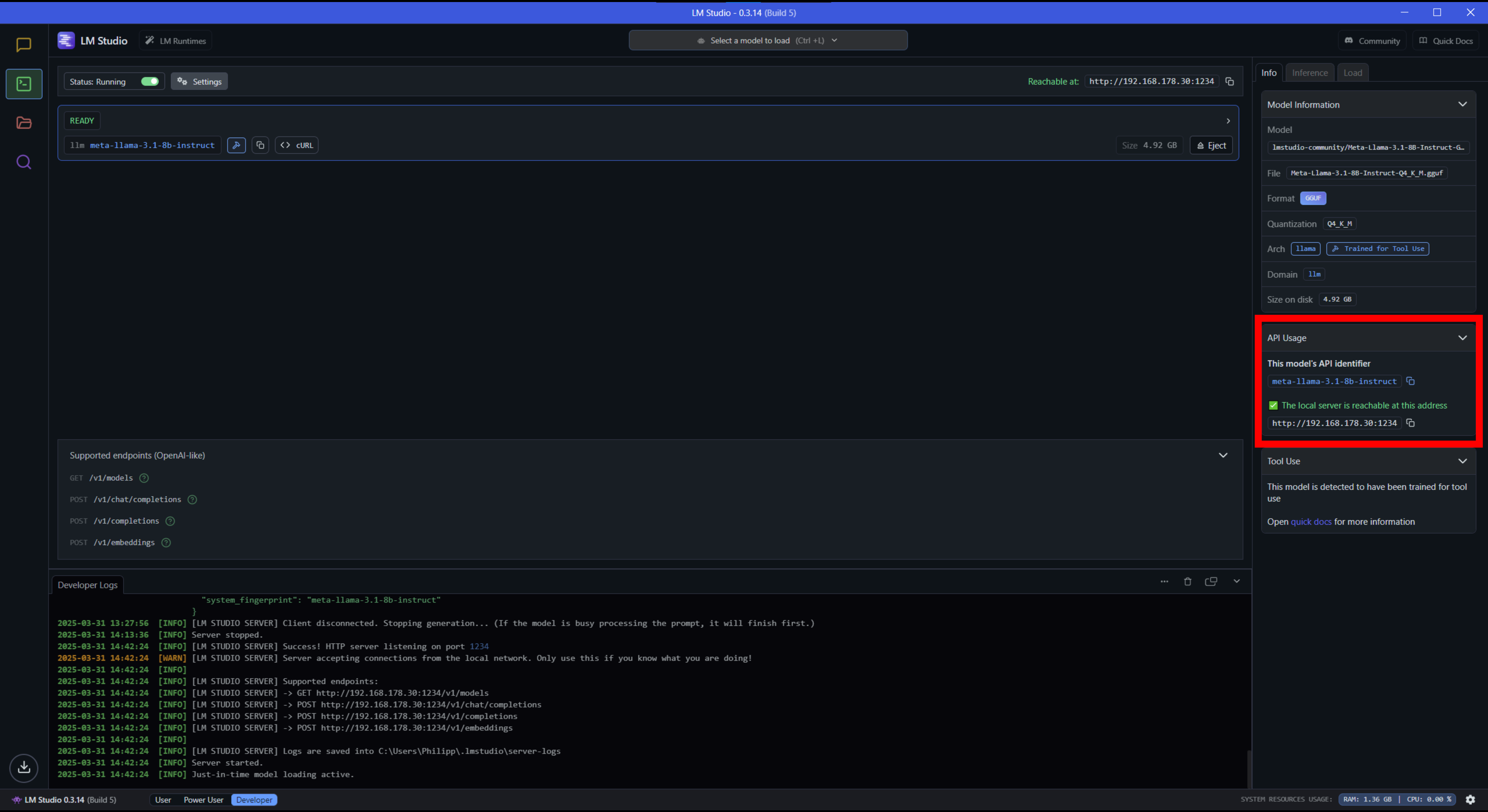


Why Does It Talk Like That?



```html
<html lang="en">
<head>
    <style>
        }
    </style>
</head>
<body>
    <div class="container">
        <h1>Local AI Chat</h1>
        <textarea id="userInput" placeholder="Type your prompt here..."></textarea>
        <button onclick="sendRequest()">Send</button>
        <div id="response" class="response"></div>
    </div>

    <script>
    async function sendRequest() {
        const userInput = document.getElementById('userInput').value;
        const responseDiv = document.getElementById('response');

        responseDiv.textContent = 'Processing...';

        try {
            const response = await fetch('http://localhost:1234/v1/chat/completions', {
                method: 'POST',
                headers: { 'Content-Type': 'application/json' },
                body: JSON.stringify({
                    model: "local-model",
                    messages: [{ role: "user", content: userInput },
                            { role: "system", content: "Only answer in rhymes" }],
                    max_tokens: 200
                })
            });

            if (!response.ok) {
                throw new Error(`HTTP error! status: ${response.status}`);
            }

            const data = await response.json();
            responseDiv.textContent = data.choices[0].message.content.trim();
        } catch (error) {
            responseDiv.textContent = `Error: ${error.message}`;
        }
    }
    </script>
</body>
</html>

```

Now, let’s try it with a microcontroller!



Status: Running

Reachable at: http://192.168.178.30:1234

llm meta-llama-3.1-8b-instruct

Size 4.92 GB

Model Information

Model lmstudio-community/Meta-Llama-3.1-8B-Instruct-G…

File Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf

Format GGUF

Quantization Q4_K_M

Arch llama Trained for Tool Use

Domain llm

Size on disk 4.92 GB

API Usage

This model's API identifier

meta-llama-3.1-8b-instruct

✅ The local server is reachable at this address

http://192.168.178.30:1234

Tool Use

This model is detected to have been trained for tool use

Open quick docs for more information

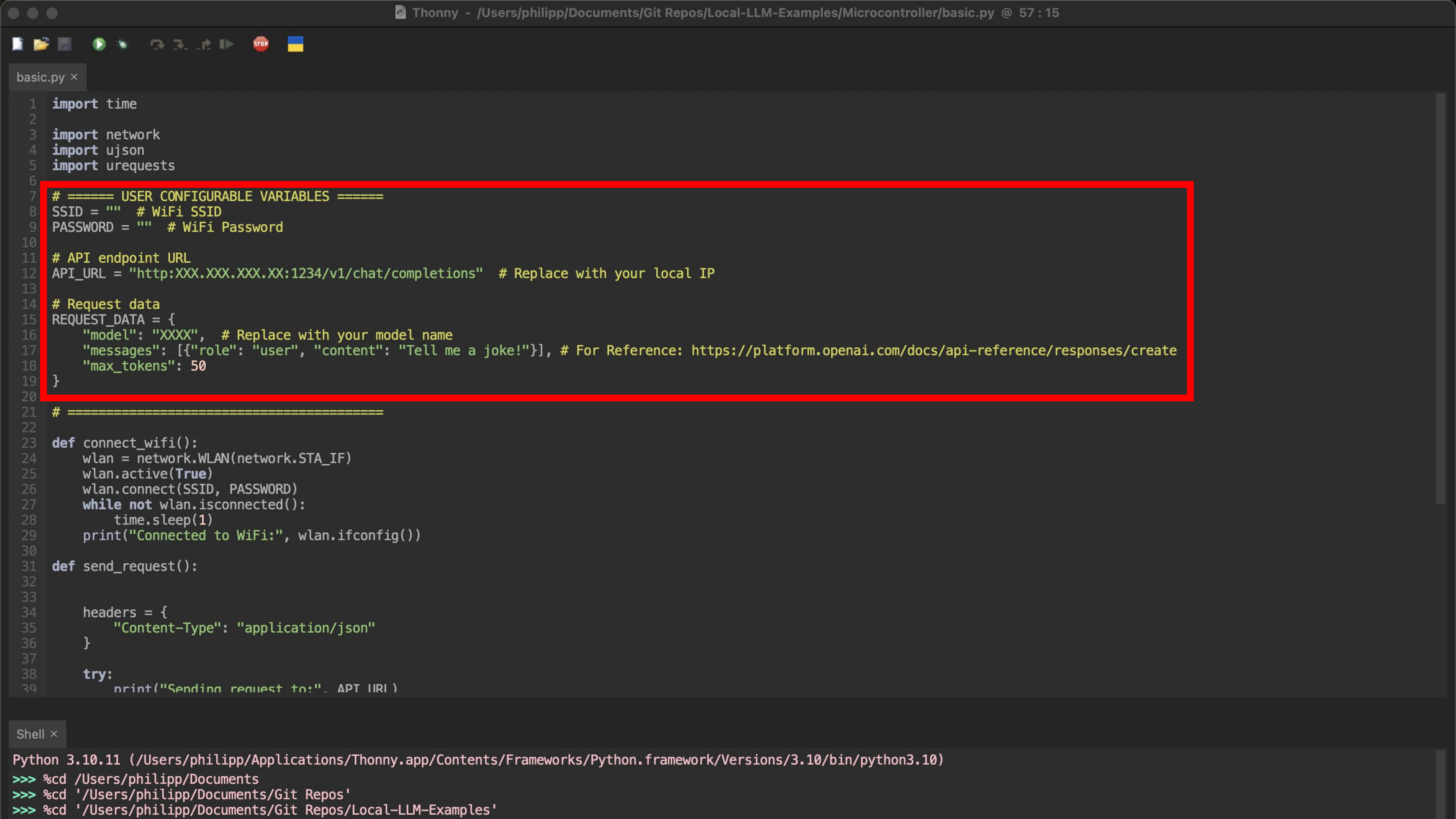
Supported endpoints (OpenAI-like)

```
GET  /v1/models
POST /v1/chat/completions
POST /v1/completions
POST /v1/embeddings
```

```
                            "system_fingerprint": "meta-llama-3.1-8b-instruct"
                          }
2025-03-31 13:27:56  [INFO] [LM STUDIO SERVER] Client disconnected. Stopping generation... (If the model is busy processing the prompt, it will finish first.)
2025-03-31 14:13:36  [INFO] Server stopped.
2025-03-31 14:42:24  [INFO] [LM STUDIO SERVER] Success! HTTP server listening on port 1234
2025-03-31 14:42:24  [WARN] [LM STUDIO SERVER] Server accepting connections from the local network. Only use this if you know what you are doing!
2025-03-31 14:42:24  [INFO]
2025-03-31 14:42:24  [INFO] [LM STUDIO SERVER] Supported endpoints:
2025-03-31 14:42:24  [INFO] [LM STUDIO SERVER] -> GET http://192.168.178.30:1234/v1/models
2025-03-31 14:42:24  [INFO] [LM STUDIO SERVER] -> POST http://192.168.178.30:1234/v1/chat/completions
2025-03-31 14:42:24  [INFO] [LM STUDIO SERVER] -> POST http://192.168.178.30:1234/v1/completions
2025-03-31 14:42:24  [INFO] [LM STUDIO SERVER] -> POST http://192.168.178.30:1234/v1/embeddings
2025-03-31 14:42:24  [INFO]
2025-03-31 14:42:24  [INFO] [LM STUDIO SERVER] Logs are saved into C:\Users\Philipp\.lmstudio\server-logs
2025-03-31 14:42:24  [INFO] Server started.
2025-03-31 14:42:24  [INFO] Just-in-time model loading active.
```

LM Studio 0.3.14 (Build 5)

SYSTEM RESOURCES USAGE: RAM: 1.36 GB | CPU: 0.00 %



Thonny - /Users/philipp/Documents/Git Repos/Local-LLM-Examples/Microcontroller/basic.py @ 57 : 15
basic.py ×
Shell ×

```python
import time

import network
import ujson
import urequests

# ====== USER CONFIGURABLE VARIABLES ======
SSID = ""  # WiFi SSID
PASSWORD = ""  # WiFi Password

# API endpoint URL
API_URL = "http:XXX.XXX.XXX.XX:1234/v1/chat/completions"  # Replace with your local IP

# Request data
REQUEST_DATA = {
    "model": "XXXX",  # Replace with your model name
    "messages": [{"role": "user", "content": "Tell me a joke!"}], # For Reference: https://platform.openai.com/docs/api-reference/responses/create
    "max_tokens": 50
}

# ========================================

def connect_wifi():
    wlan = network.WLAN(network.STA_IF)
    wlan.active(True)
    wlan.connect(SSID, PASSWORD)
    while not wlan.isconnected():
        time.sleep(1)
    print("Connected to WiFi:", wlan.ifconfig())

def send_request():


    headers = {
        "Content-Type": "application/json"
    }

    try:
        print("Sending request to:", API_URL)
```

```
Python 3.10.11 (/Users/philipp/Applications/Thonny.app/Contents/Frameworks/Python.framework/Versions/3.10/bin/python3.10)
>>> %cd /Users/philipp/Documents
>>> %cd '/Users/philipp/Documents/Git Repos'
>>> %cd '/Users/philipp/Documents/Git Repos/Local-LLM-Examples'
```