

[Open in app](#) ↗[Sign up](#)[Sign in](#)**Medium** Search

How to use a local LLM as a free coding copilot in VS Code

Simon Fraser · [Follow](#)

6 min read · Dec 2, 2023



Listen



Share



Do you want to have an AI powered project without giving any money to Microsoft via paying for GitHub copilot? Do you want a solution that works offline in the middle of the desert just as well as at your house? Want to support open source software? You might be interested in using a local LLM as a coding assistant and all you have to do is follow the instructions below.

Before we start let's make sure this option is a good fit for you by reviewing the pros and cons of a local LLM vs GitHub Copilot.

Pros and Cons of using a Local LLM:

Pros:

- Free — only costs a bit more on your electric bill
- Data security — your data never leaves your device
- No internet required
- Complete control of the system prompt

Cons:

- Hardware requirements — works better if you have a better computer

- Worse models — the best models are proprietary and too big for consumer hardware

Tools we'll use:

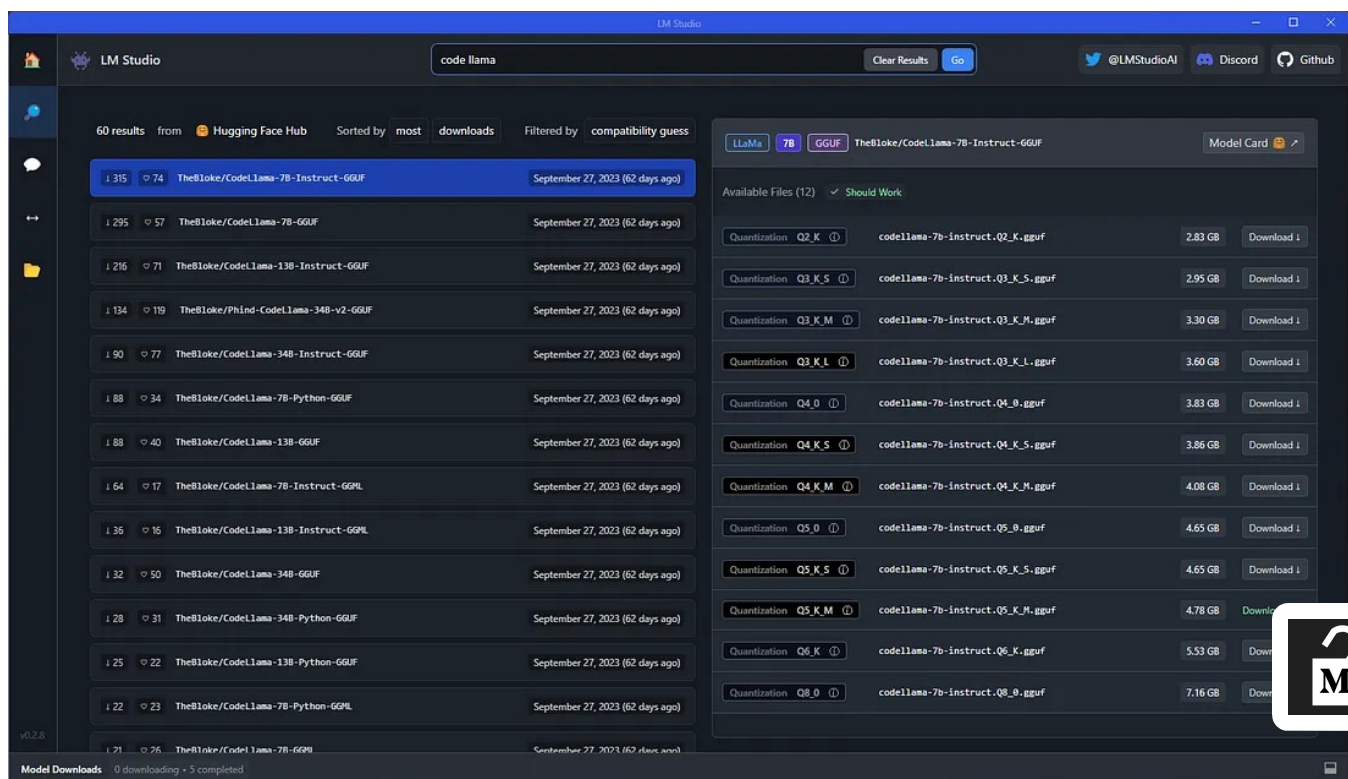
Continue for VS Code

LM Studio (Ollama or llama-cpp-python are alternatives)

Let's Get Started:

First download the LM Studio installer from [here](#) and run the installer that you just downloaded. After installation open LM Studio (if it doesn't open automatically). You should now be on the home page on LM Studio with a search bar in the middle. In this search bar (or the search tab on the left, you end up in the same place) type the name of the model you want to use, I'd recommend starting with Code Llama 7B Instruct, which is the first result if I just search "code llama." The task now is to choose which quantization of the model to download. To save memory, LLMs are quantized down from their original 16-bit representations down to fewer bits, which trades a bit of quality for size. To learn more about the quantizations available check out the model card for the model you've chosen, which should also have information about the maximum RAM required for the various options (which is different than the disk space shown on the download screen). After choosing your quantization click the download button and wait for the model to download.

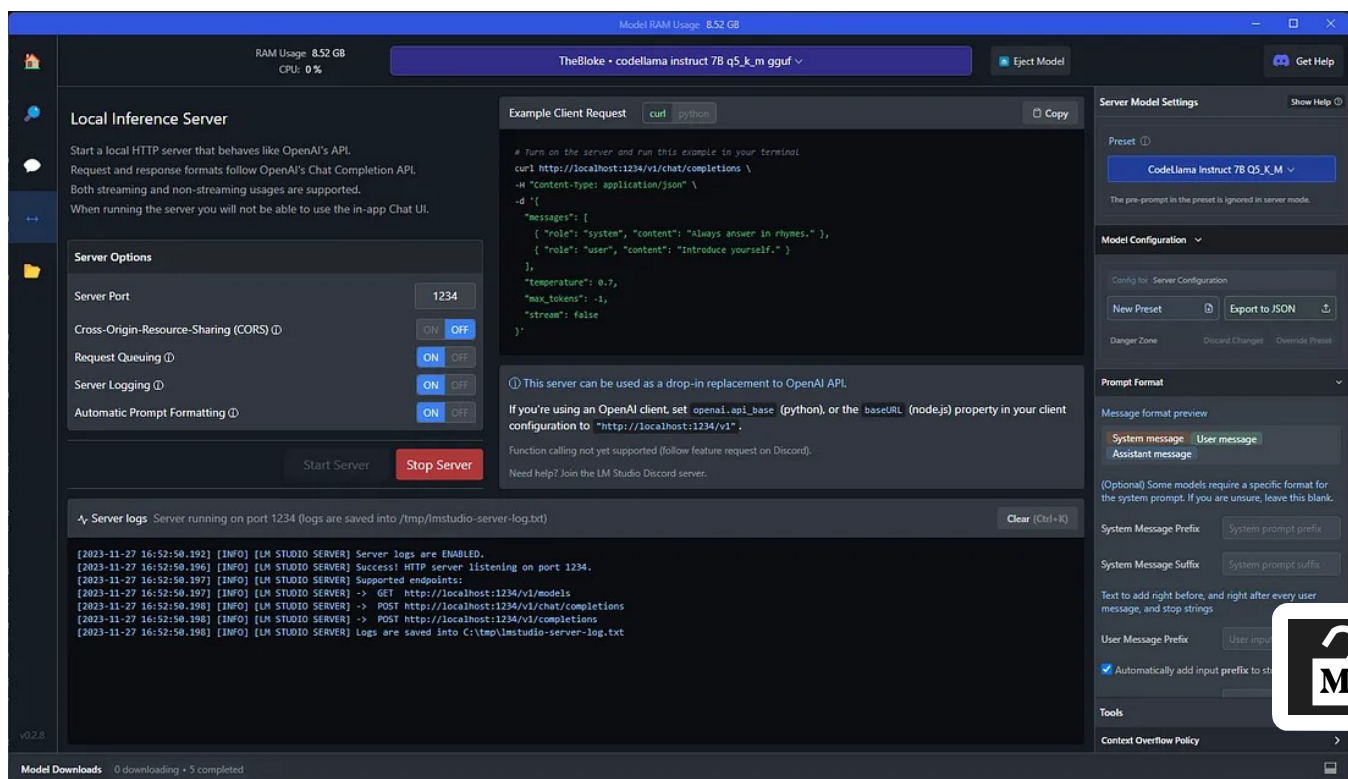




LM Studio search tab after downloading a quantization

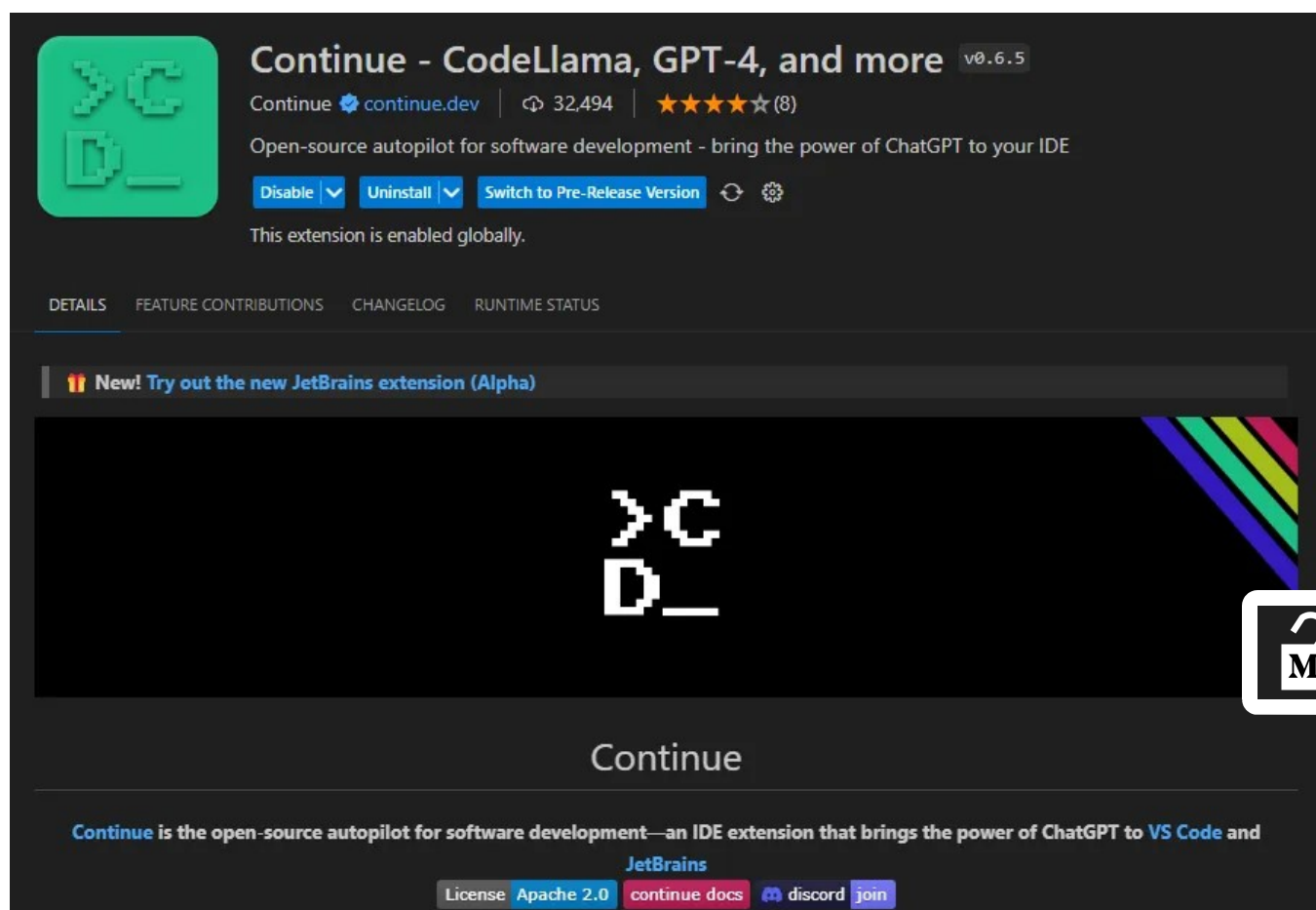
At this point you can switch to the AI Chat tab in LM Studio and chat directly with your model after loading it if you choose. Here we'll focus on the server functionality of LM Studio for use with the Continue VS Code extension.

On the Local Server tab of LM Studio click the "Select a model to load" button at the top (assuming you didn't already do this to chat with the model). If this fails you might have chosen a model/quantization that is too big for your RAM. If you have a GPU you might be able to save it by offloading some layers to the GPU (see below) otherwise you'll have to choose a smaller model/quantization. If the model loaded successfully then you can click the "Start Server" button and you should be hosting your model on a local server ready to go!



LM Studio Local Server tab with a running server

With that sorted let's head over to VS Code and download the open source Continue extension.



Page for the Continue extension after downloading

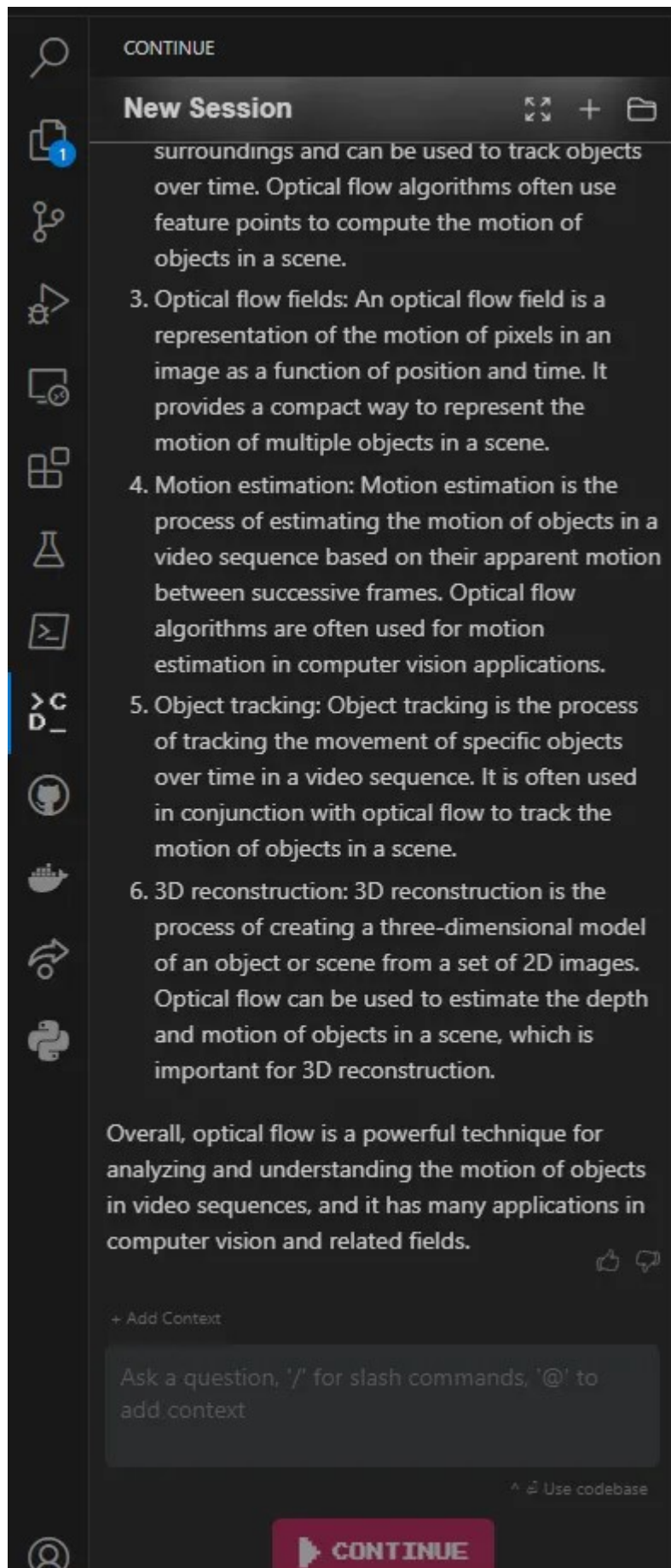
After downloading Continue we just need to hook it up to our LM Studio server. To do this we'll need to edit Continue's config.json file. On Windows this file is located at C:/Users/{user}/.continue/config.json, on Linux or Mac the file should be located at ~/.continue/config.json. In this file we need to tell Continue about the models we plan to use and how we plan on using them. To do this add a JSON object to the "models" array for each model you plan on using. You need to provide the "title", "provider", and "model" fields. You can also provide a "server_url" field if you like but if the "provider" field is set to "lmstudio" it assumes you are using the default address/port of http://localhost:1234. You can also set the model currently in use by changing the "model_roles" object. Finally, you can set the system prompt by setting the "system_message" field.

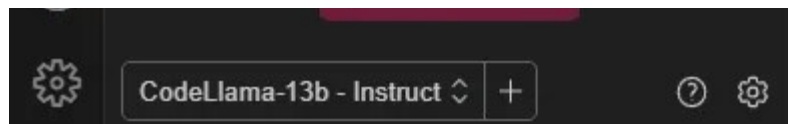
```
+ Add a Model
2  "models": [
3    {
4      "title": "GPT-4",
5      "provider": "openai-free-trial",
6      "model": "gpt-4"
7    },
8    {
9      "title": "GPT-3.5-Turbo",
10     "provider": "openai-free-trial",
11     "model": "gpt-3.5-turbo"
12   },
13   {
14     "title": "CodeLlama-7b - Instruct",
15     "provider": "lmstudio",
16     "model": "codellama-7B"
17   },
18   {
19     "title": "CodeLlama-13b - Instruct",
20     "provider": "lmstudio",
21     "model": "codellama-13b"
22   },
23   {
24     "title": "Mistral",
25     "provider": "lmstudio",
26     "model": "mistral-7b"
27   }
28 ],
29 "model_roles": {
30   "default": "CodeLlama-13b - Instruct",
31   "chat": "CodeLlama-13b - Instruct",
32   "edit": "CodeLlama-13b - Instruct",
33   "summarize": "CodeLlama-13b - Instruct"
34 },
35 "system_message": "",
```



config.json after additions

You can also set the system prompt in Continue's GUI by clicking on the gear icon in the lower left of the Continue tab and editing the "System Message" text box. You can switch the model Continue is using in the GUI by changing the box at the bottom of the Continue tab although be careful that if you set it to another model using the LM Studio server that you also load the corresponding model in LM Studio. Otherwise the Continue GUI will be showing the wrong model (which isn't the worst thing in the world but would bother me, maybe you too).





Continue GUI tab

At this point you should be able to use Continue with your local LLM! Continue can answer questions about your code, edit your code, or generate files from scratch. See the [Continue documentation](#) for full details. If you want to use Continue without any contact with the outside world, let LM Studio use the potential of your hardware to run the models faster or save preset configurations for different models in LM Studio then continue on to the next section.

Fine Tuning Settings:

If you want to use Continue fully locally head back to the config.json file and add this line disallowing anonymous telemetry so Continue doesn't try to interact with the outside world at all.



```
29   "model_roles": {  
30     "default": "CodeLlama-13b - Instruct",  
31     "chat": "CodeLlama-13b - Instruct",  
32     "edit": "CodeLlama-13b - Instruct",  
33     "summarize": "CodeLlama-13b - Instruct"  
34   },  
35   "allow_anonymous_telemetry": false,  
36   "system_message": "",
```

config.json with addition

Now that everything is working let's make it run faster. By default the inference isn't particularly fast since it only uses CPU with 4 threads, which may not be optimal for your system. If you have a GPU you can get significantly faster inference by offloading some layers of the model to it. To do this just check the box by "GPU Acceleration" at the bottom of the right side scrolling menu of LM Studio and set `n_gpu_layers` to something nonzero. As the info icon next to GPU Acceleration notes, start with a small value (10–20) and increase until you're happy with your GPU utilization. The optimal values here are going to require some trial and error to find. I set my number of CPU threads to 6 since I have a 6-core CPU and played around with `n_gpu_layers` until I found a value that gave me good GPU utilization.

Model Initialization

☒ Keep entire model in RAM ⓘ

use_mlock on

Changing the values below will reload the model.

Prompt eval batch size

n_batch512

Context Length

n_ctx100000

Different models support different context sizes.
Consult the model card to verify your chosen value.

Rotary Position Embedding (RoPE) ⓘ

Frequency Scale

rope_freq_scale1

(Experimental) For SuperHOT 8K, set to 0.25. Your results may vary. Join the discussion on Discord.

Frequency Base

rope_freq_base1000000

Hardware Settings

☒ GPU Acceleration ⓘ

n_gpu_layers32

Detected GPU type

Nvidia CUDA

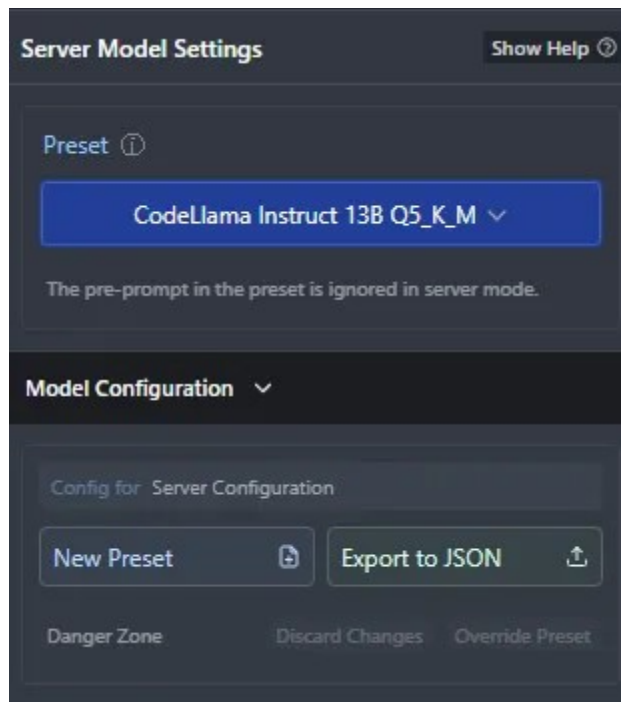
CPU Threads ⓘ

n_threads6



Hardware settings I use for CodeLlama-7B

Once you have a configuration you're happy with you can save it by clicking on the "New Preset" button. This will make it appear as an option in the preset menu and makes switching models a breeze.



A saved preset and button to make more

That's it! We're now running a local LLM coding copilot for free with GPU acceleration. Hope you enjoyed!

Final Notes:

I'm using LM Studio since I'm on Windows (fake dev I know) and Ollama doesn't currently support Windows. In this future I might rewrite this guide to use Ollama or write another guide using Ollama instead of LM Studio to support open source software.

AI

Programming

Vscode





Follow



Written by Simon Fraser

73 Followers

Recommended from Medium



Harendra in Dev Genius

11 Open-Source SaaS Killer — Selfhost With Docker

Selfhost Supabase, Grafana, Uptime Kuma, NocoDB, Dokku, Appwrite, N8N, Redash, Jitsi, Plausible and Nextcloud with docker



Sep 9

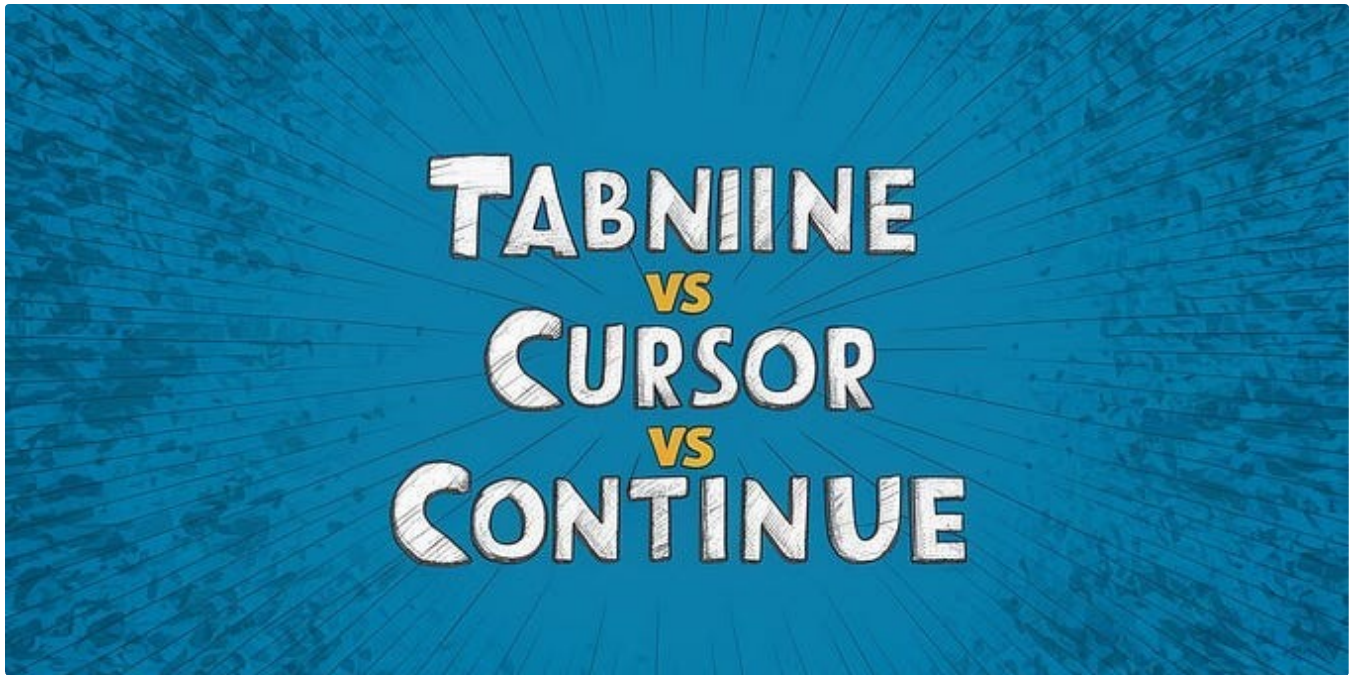


658



7





AI Tool Scan



Comparing Code Assistant Programs: Tabnine, Cursor, and Continue

Code assistant programs help developers write, debug, and optimize code more efficiently. With the advancement of AI-based code assistants...

Jun 16 🖱️ 2



Lists



General Coding Knowledge

20 stories • 1572 saves



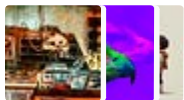
Coding & Development

11 stories • 812 saves



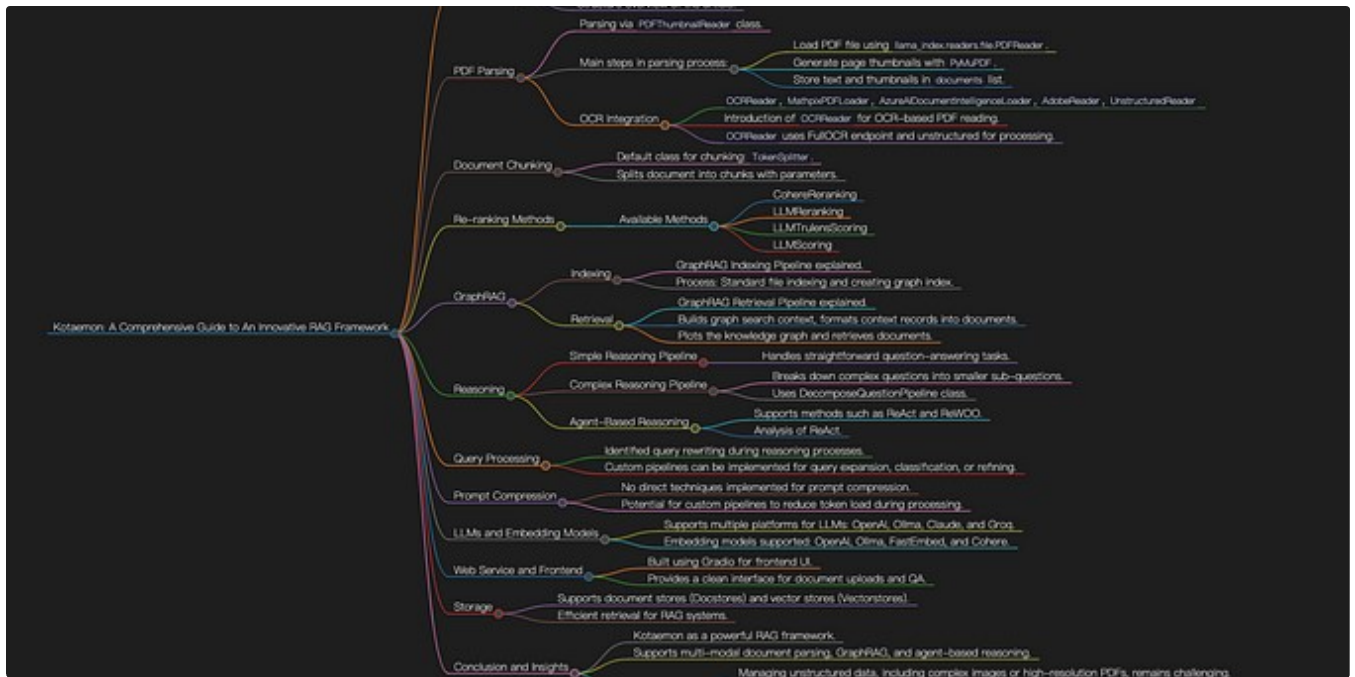
Generative AI Recommended Reading


52 stories • 1369 saves



What is ChatGPT?

9 stories • 433 saves



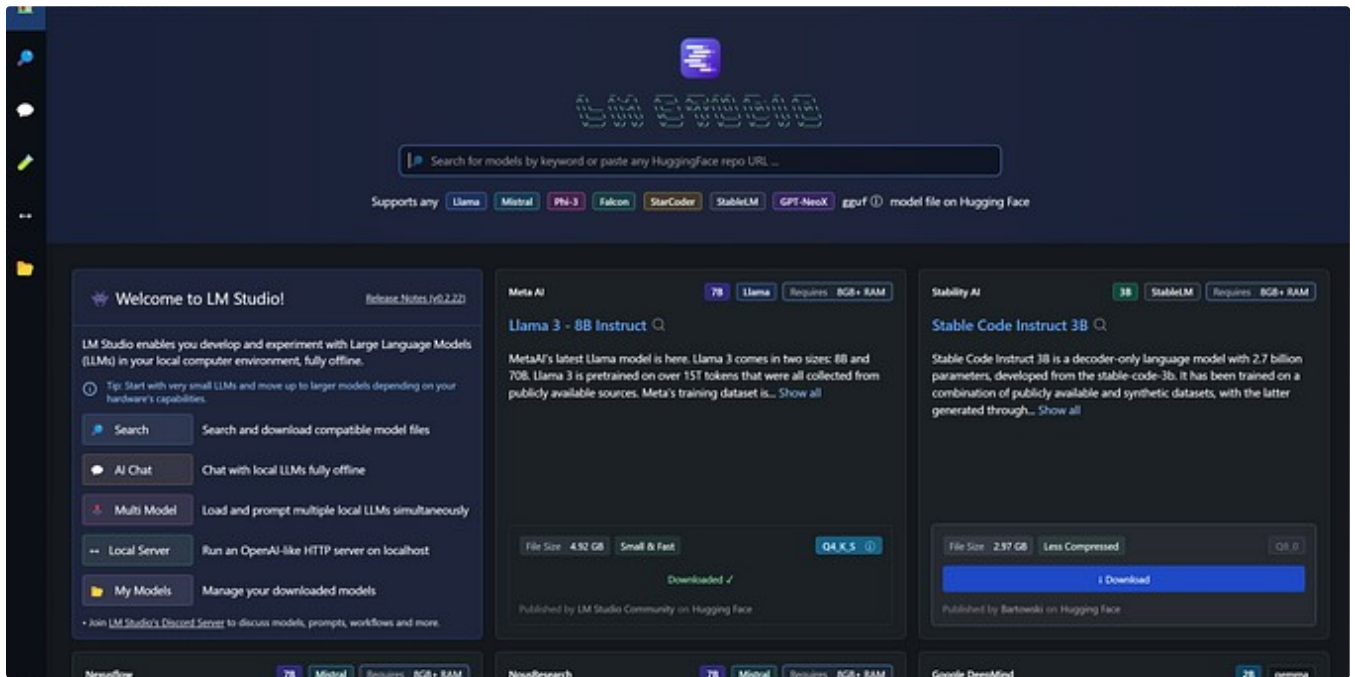
 Florian June in AI Advances




Kotaemon Unveiled: Innovations in RAG Framework for Document QA

PDF Parsing, GraphRAG, Agent-Based Reasoning, and Insights

🌟 5d ago 🤝 405 💬 2




 Emanuele

LM Studio: A One Click Installer to Download and Run LLMs Locally

I'm always on the lookout for the best and easiest way to try out the latest language models on my PC, locally and offline.



 Addison Best in Generative AI

Llama 3.1 405B—How to Use for Free

No Local Install Needed

★ Aug 19 🖱️ 196 💬 2





 howtouselinux in Programming Domain



Meet Cursor AI: The AI Coding Tool That's Redefining Software Development

In the post-AI era, new AI innovations frequently capture the spotlight.

★ Sep 4 👏 23 💬 1



See more recommendations