

[Open in app ↗](#)**Medium**

Search



Write

**Member-only story****Featured**

# How To Install And Use DeepSeek R-1 In Your Local PC

Jim Clyde Monge  · [Follow](#)

Published in Generative AI · 7 min read · Jan 23, 2025

2.2K

47



...

Image by [Jim Clyde Monge](#)

**E**veryone seems to be talking about DeepSeek R-1, the new open-source AI language model made by a Chinese AI firm, DeepSeek. Some users claim it's on par with, or even better than, OpenAI's GPT-3 in terms of reasoning capabilities.

Currently, DeepSeek is free to use, which is great news for users, but it does raise some questions. With the surge in user volume, how are they managing the server costs?

Hardware running costs cannot be cheap, right?

The one logical here would be — data. Data is the lifeblood of AI models. They're probably collecting user data in some way that benefits their quant trading model or for some other form of monetization.

So, if you're concerned about data privacy but still want to use R1 without sharing your data, the best way is to run the model locally.

## **What is DeepSeek R-1?**

A couple of days back, Deepseek R-1 was unveiled as a fully open-sourced model, meaning anyone can take the underlying codebase, adapt it, and even fine-tune it to their own needs.

From a technical standpoint, Deepseek R-1 (often abbreviated as R1) stems from a large base model called DeepSeek-V3. The lab then refined this model through a combination of supervised fine-tuning (SFT) on high-quality human-labeled data and reinforcement learning (RL).

The result is a chatbot that can handle intricate prompts, reveal the reasoning steps behind complex questions (sometimes more transparently

than other models), and even render code in the chat interface for quick testing.

It's honestly very impressive, especially for a model that's open-source.

## How To Run It Locally

To run DeepSeek R-1 locally, we'll be using a tool called [Ollama](#).

Ollama is a free, open-source tool that allows users to run large language models (LLMs) locally on their computers. It is available for macOS, Linux, and Windows.

Head to the official Ollama website and click on the “Download” button. Install it on your system.

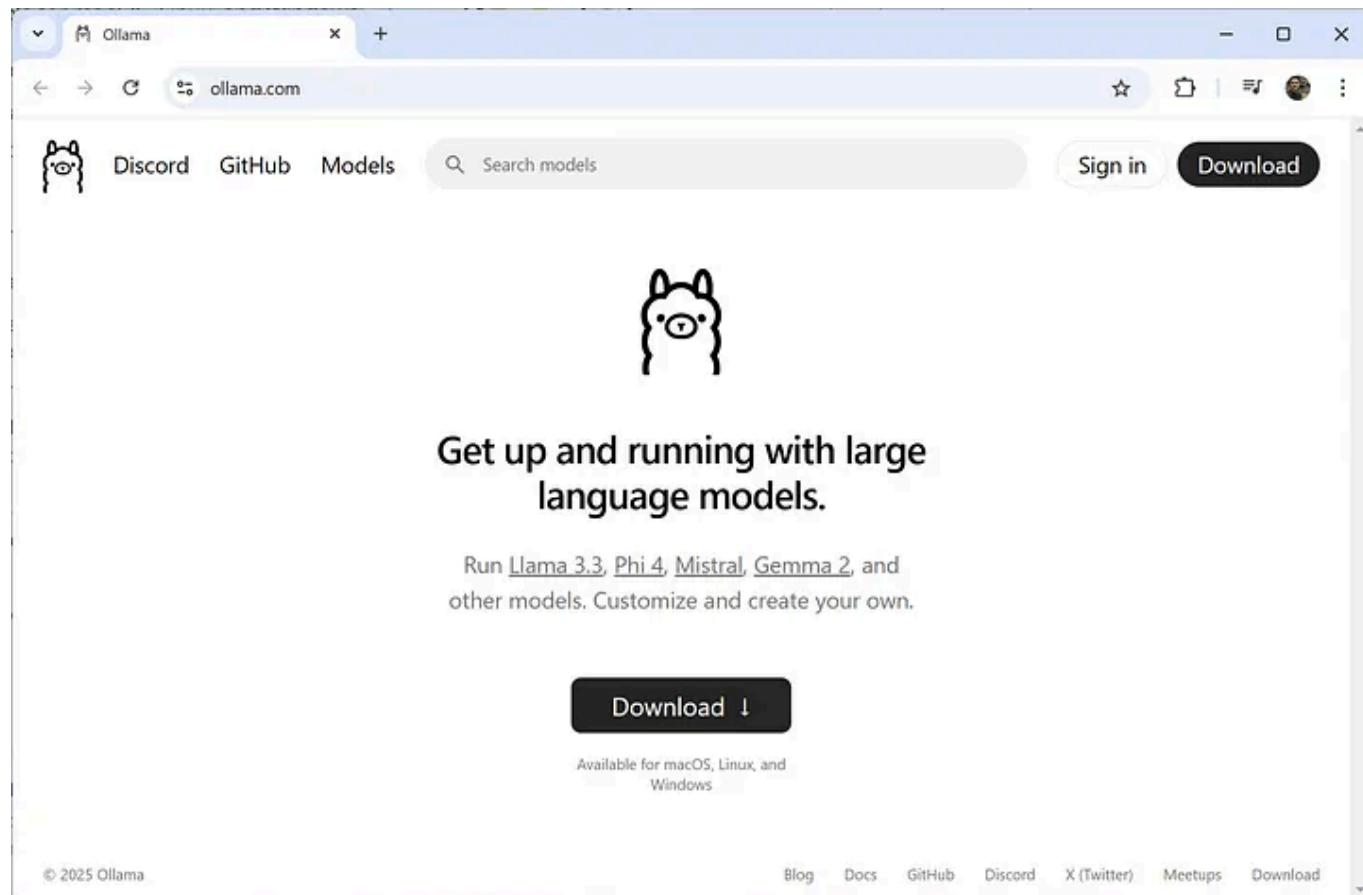
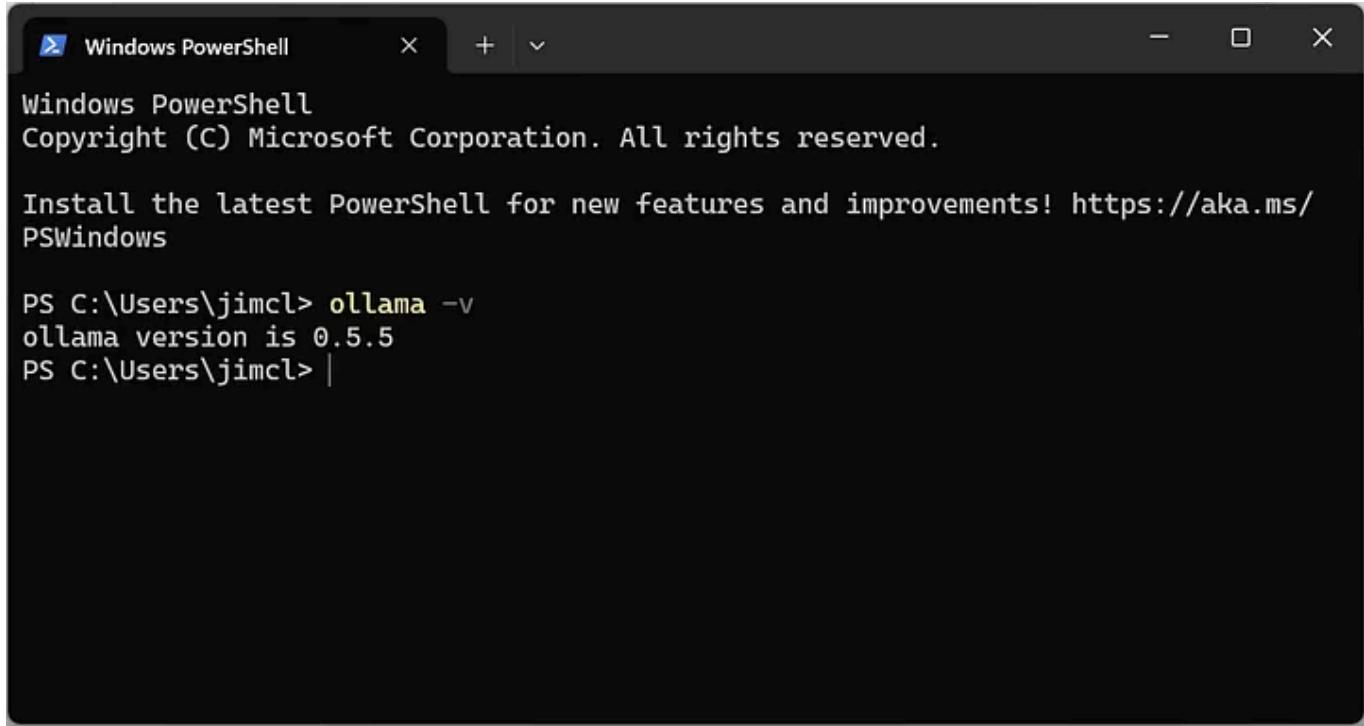


Image by [Jim Clyde Monge](#)

To confirm that the installation is successful, open a terminal and run the following command:

```
ollama -v
```

A screenshot of a Windows PowerShell window titled "Windows PowerShell". The window shows the standard PowerShell welcome message: "Windows PowerShell", "Copyright (C) Microsoft Corporation. All rights reserved.", and a link to install the latest version. Below this, the command "ollama -v" is entered, followed by its output: "ollama version is 0.5.5".

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/
PSWindows

PS C:\Users\jimcl> ollama -v
ollama version is 0.5.5
PS C:\Users\jimcl> |
```

Image by [Jim Clyde Monge](#)

You should see the version number of the Ollama instead of an error.

Under the Models tab, search for the keyword “deepseek” and you should see the “deepseek-r1” on the first item on the search list.

A screenshot of a web browser displaying the Ollama website at ollama.com/search?q=deepseek. The search bar contains 'deepseek'. Below the search bar are tabs for All, Embedding, Vision, Tools, and Popular. A dropdown menu is open over the Popular tab. The main content area shows two model cards. The first card, 'deepseek-r1', is highlighted with a red border. It has a title 'deepseek-r1', a description 'DeepSeek's first generation reasoning models with comparable performance to OpenAI-o1.', and a list of model sizes: 1.5b, 7b, 8b, 14b, 32b, 70b, 671b. Below the sizes are metrics: 366.8K Pulls, 28 Tags, and Updated 2 days ago. The second card, 'deepseek-coder-v2', has a title 'deepseek-coder-v2', a description 'An open-source Mixture-of-Experts code language model that achieves performance comparable to GPT4-Turbo in code-specific tasks.', and a list of model sizes: 16b, 236b. Below the sizes are metrics: 470.9K Pulls, 64 Tags, and Updated 4 months ago.

Image by [Jim Clyde Monge](#)

Click on it and down the Models section, you'll notice that there are multiple model sizes from 5 billion to 671 billion parameters. As a rule of thumb, larger models require more powerful GPUs to run.

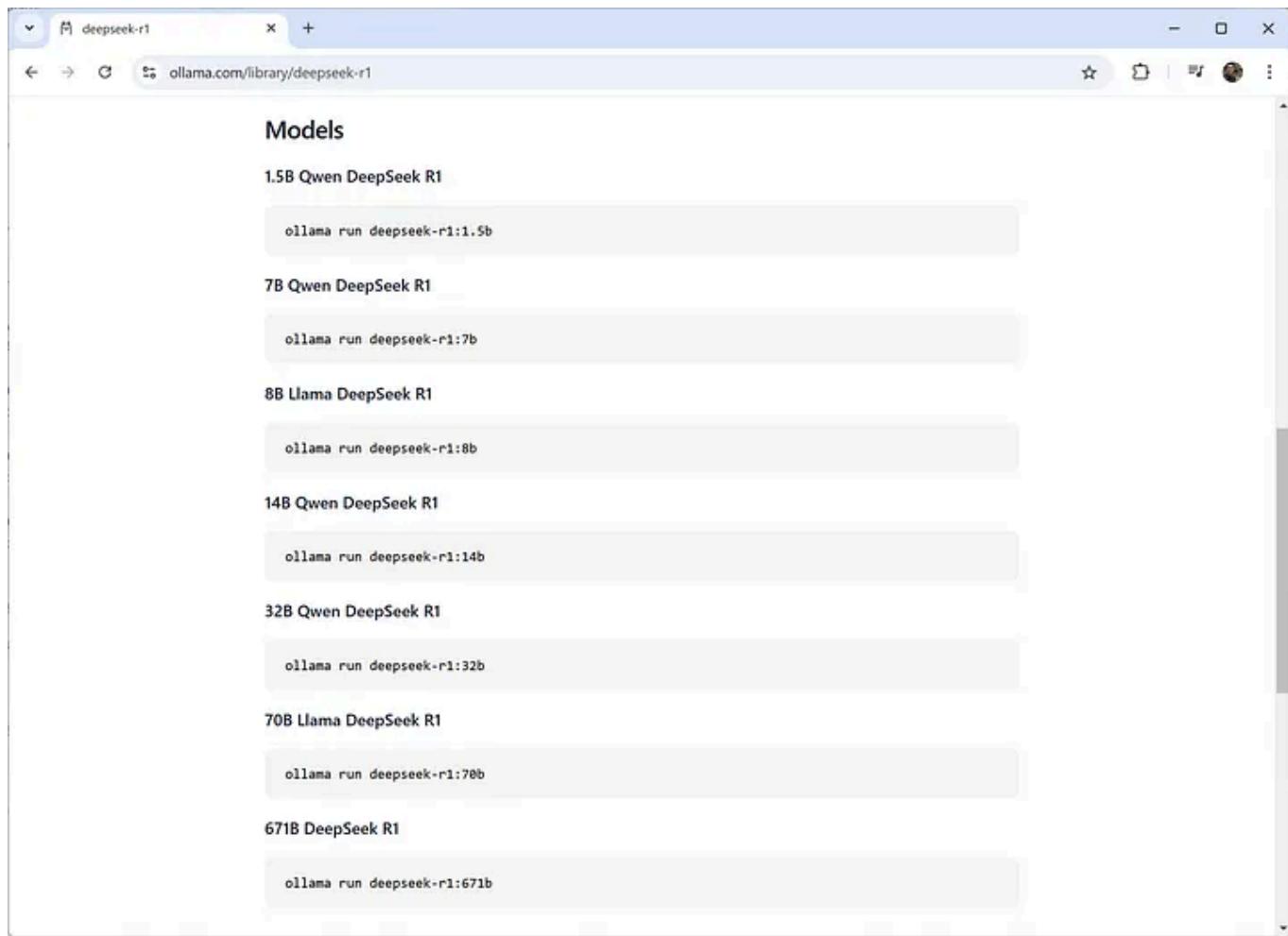


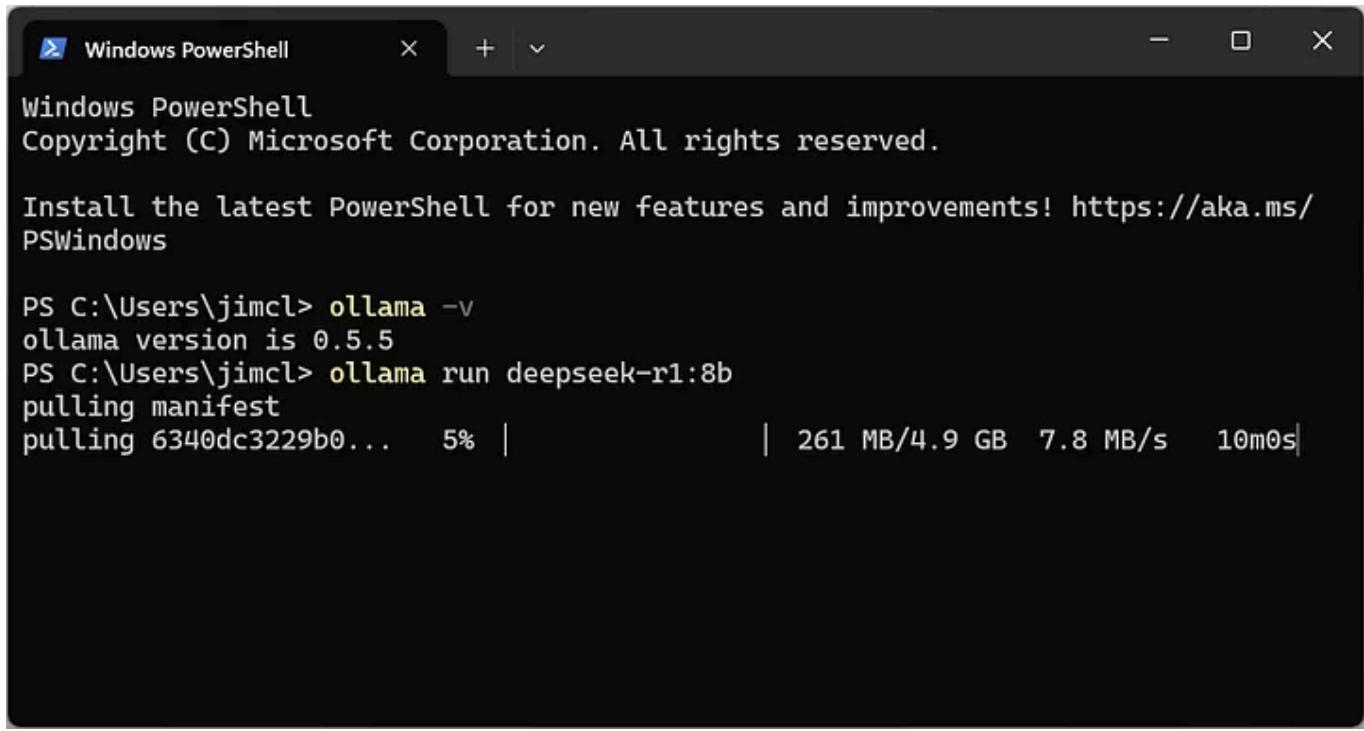
Image by [Jim Clyde Monge](#)

Smaller models like the 8 billion parameter version can run on GPUs with 8GB of VRAM. Larger models need significantly more resources (refer to the VRAM and GPU requirements section below).

To download and run the 8 billion parameter model, use the following command:

```
ollama run deepseek-r1:8b
```

The model will start downloading (around 4.9GB). Ensure you have enough disk space before proceeding.

A screenshot of a Windows PowerShell window titled "Windows PowerShell". The window shows the command "ollama run deepseek-r1:8b" being executed, which is pulling a manifest from a GitHub repository. The progress bar indicates the download is at 5% completion, with a speed of 7.8 MB/s and an estimated time of 10m0s remaining. The background of the window is dark, and the text is white.

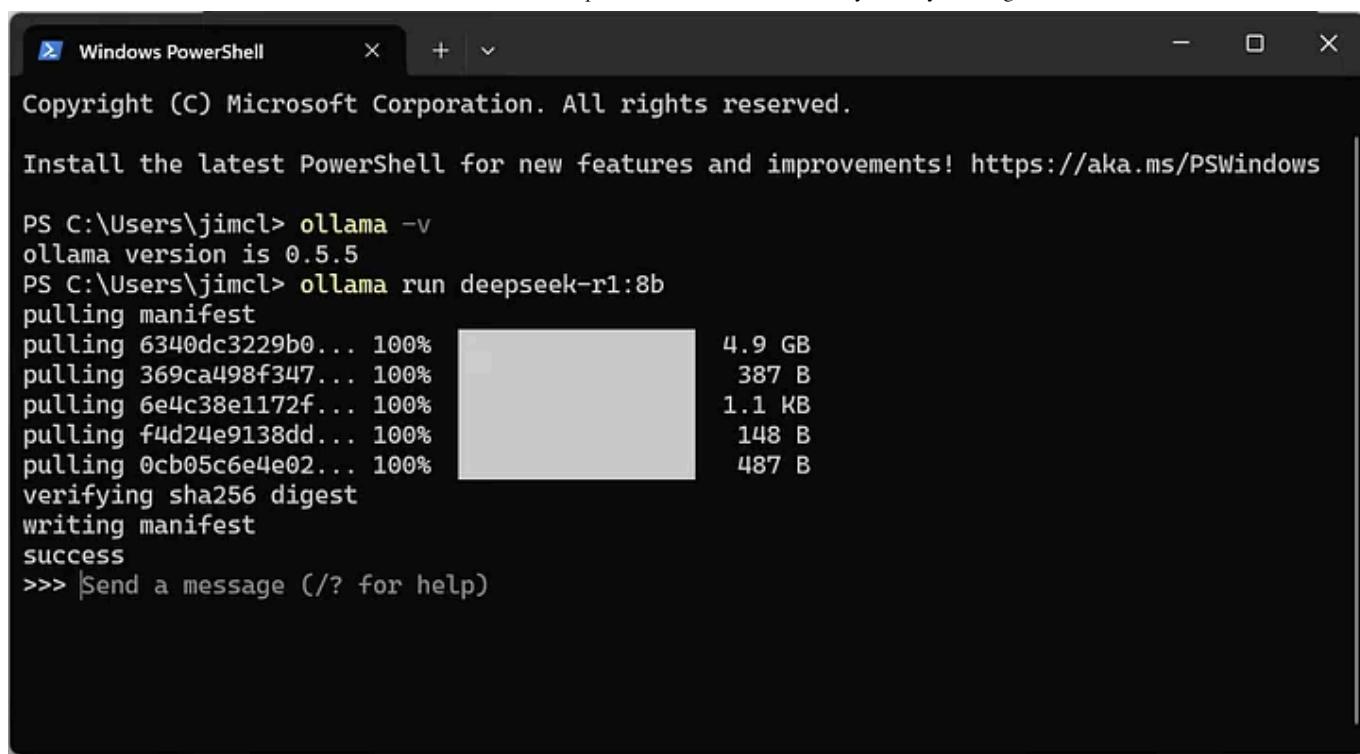
```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/
PSWindows

PS C:\Users\jimcl> ollama -v
ollama version is 0.5.5
PS C:\Users\jimcl> ollama run deepseek-r1:8b
pulling manifest
pulling 6340dc3229b0... 5% | 261 MB/4.9 GB 7.8 MB/s 10m0s|
```

Image by [Jim Clyde Monge](#)

Once downloaded, the model will run locally on your machine. You can chat with it immediately.



```
Windows PowerShell

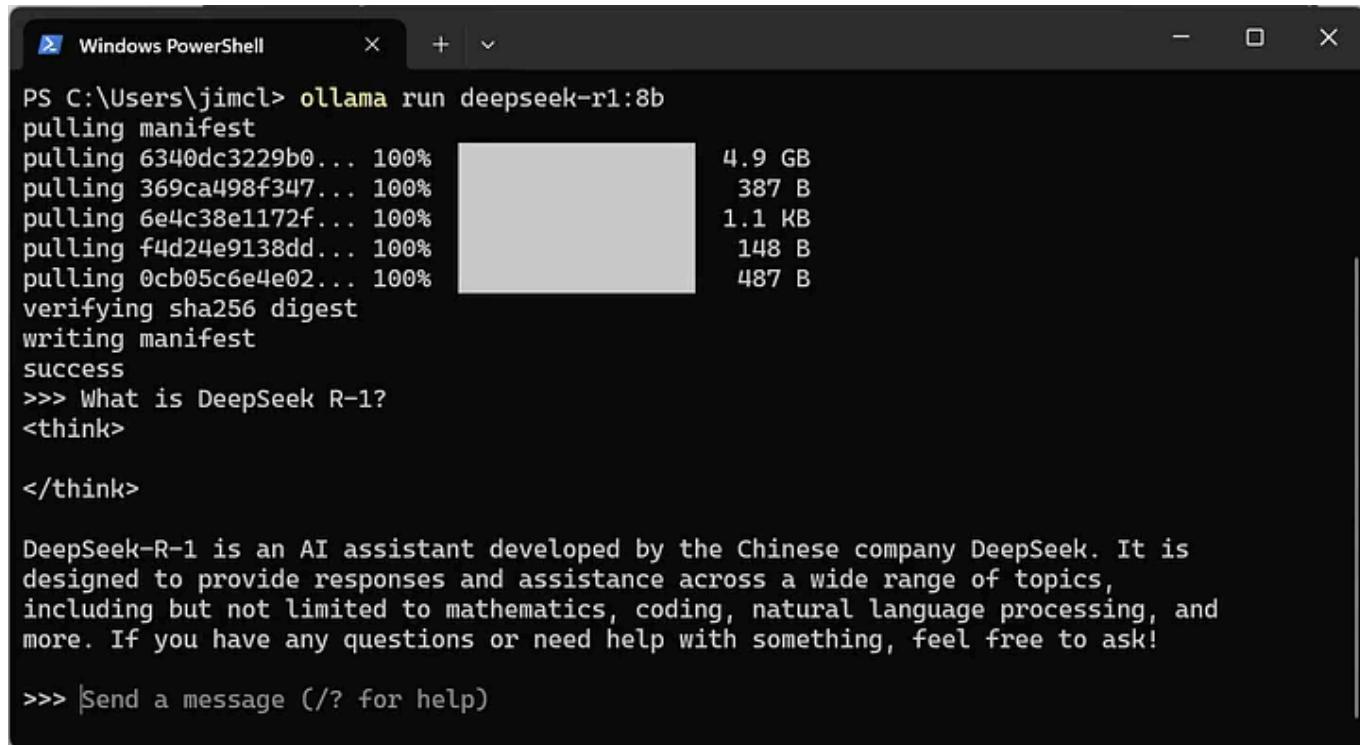
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\jimcl> ollama -v
ollama version is 0.5.5
PS C:\Users\jimcl> ollama run deepseek-r1:8b
pulling manifest
pulling 6340dc3229b0... 100% [redacted] 4.9 GB
pulling 369ca498f347... 100% [redacted] 387 B
pulling 6e4c38e1172f... 100% [redacted] 1.1 KB
pulling f4d24e9138dd... 100% [redacted] 148 B
pulling 0cb05c6e4e02... 100% [redacted] 487 B
verifying sha256 digest
writing manifest
success
>>> |Send a message (/? for help)
```

Image by Jim Clyde Monge

Let's test it with this example prompt:



```
Windows PowerShell

PS C:\Users\jimcl> ollama run deepseek-r1:8b
pulling manifest
pulling 6340dc3229b0... 100% [redacted] 4.9 GB
pulling 369ca498f347... 100% [redacted] 387 B
pulling 6e4c38e1172f... 100% [redacted] 1.1 KB
pulling f4d24e9138dd... 100% [redacted] 148 B
pulling 0cb05c6e4e02... 100% [redacted] 487 B
verifying sha256 digest
writing manifest
success
>>> What is DeepSeek R-1?
<think>

</think>

DeepSeek-R-1 is an AI assistant developed by the Chinese company DeepSeek. It is
designed to provide responses and assistance across a wide range of topics,
including but not limited to mathematics, coding, natural language processing, and
more. If you have any questions or need help with something, feel free to ask!

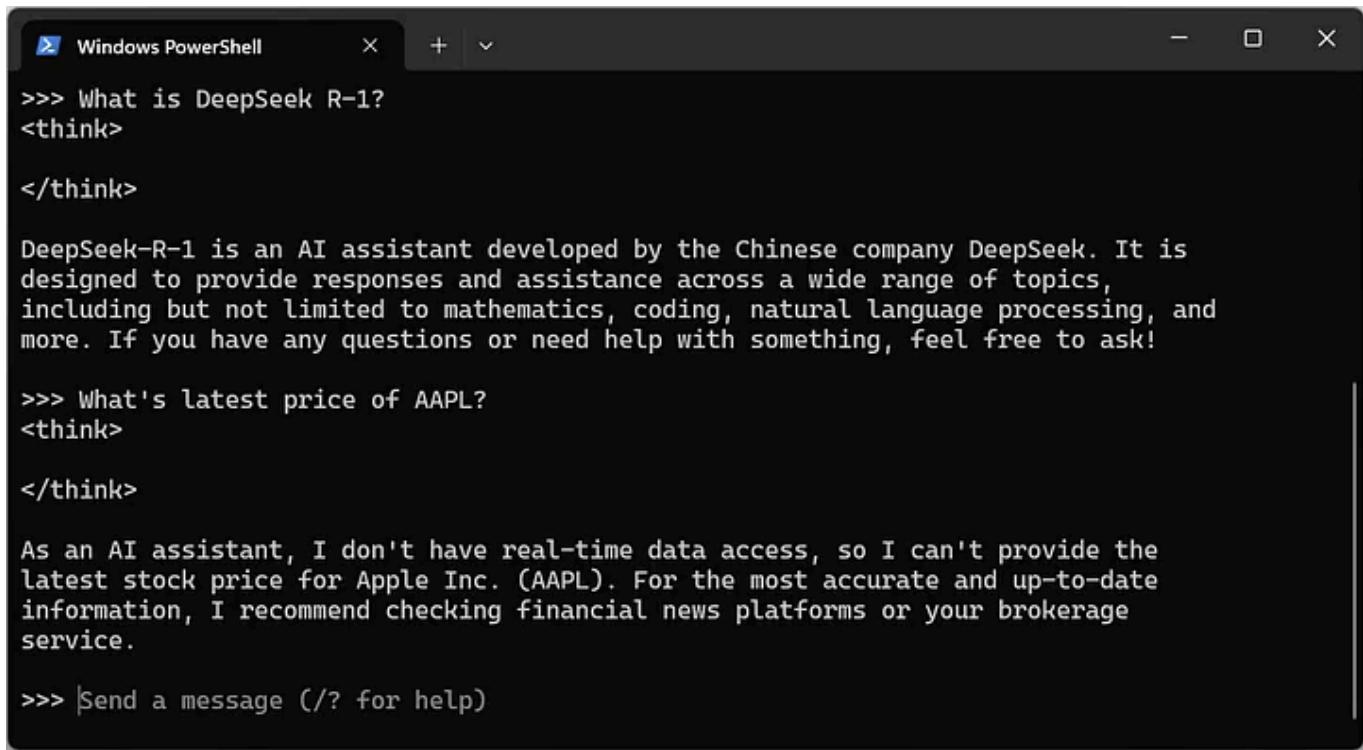
>>> |Send a message (/? for help)
```

Image by Jim Clyde Monge

**Prompt:** What is DeepSeek R-1?

**Response:** DeepSeek-R-1 is an AI assistant developed by the Chinese company DeepSeek. It is designed to provide responses and assistance across a wide range of topics, including but not limited to mathematics, coding, natural language processing, and more. If you have any questions or need help with something, feel free to ask!

Awesome. It's fast and it still works even if I disconnect my laptop from the Wifi. Note that even if you are connected to the internet, it still cannot access the web.



```
Windows PowerShell

>>> What is DeepSeek R-1?
<think>
</think>

DeepSeek-R-1 is an AI assistant developed by the Chinese company DeepSeek. It is
designed to provide responses and assistance across a wide range of topics,
including but not limited to mathematics, coding, natural language processing, and
more. If you have any questions or need help with something, feel free to ask!

>>> What's latest price of AAPL?
<think>
</think>

As an AI assistant, I don't have real-time data access, so I can't provide the
latest stock price for Apple Inc. (AAPL). For the most accurate and up-to-date
information, I recommend checking financial news platforms or your brokerage
service.

>>> |Send a message (/? for help)
```

Image by Jim Clyde Monge

**Prompt:** What's latest price of AAPL?

**Response:** As an AI assistant, I don't have real-time data access, so I can't provide the latest stock price for Apple Inc. (AAPL). For the most accurate and up-to-date

*information, I recommend checking financial news platforms or your brokerage service.*

## Other things Ollama can do:

- Runs LLMs locally, including LLaMA2, Phi 4, Mistral, and Gemma 2
- Allows users to create and share their own LLMs
- Bundles model weights, configuration, and data into a single package
- Optimizes setup and configuration details, including GPU usage

## GPU and VRAM Requirements

The VRAM requirements for DeepSeek-R1 depend on factors like model size, parameter count, and quantization techniques. Below is a detailed overview of the VRAM needs for DeepSeek-R1 and its distilled models, along with recommended GPUs:

Model	Parameters (B)	VRAM Requirement (GB)	Recommended GPU
DeepSeek-R1-Zero	671B	~1,342 GB	Multi-GPU setup (e.g., NVIDIA A100 80GB x16)
DeepSeek-R1	671B	~1,342 GB	Multi-GPU setup (e.g., NVIDIA A100 80GB x16)
DeepSeek-R1-Distill-Qwen-1.5B	1.5B	~0.75 GB	NVIDIA RTX 3060 12GB or higher
DeepSeek-R1-Distill-Qwen-7B	7B	~3.5 GB	NVIDIA RTX 3060 12GB or higher
DeepSeek-R1-Distill-Llama-8B	8B	~4 GB	NVIDIA RTX 3060 12GB or higher
DeepSeek-R1-Distill-Qwen-14B	14B	~7 GB	NVIDIA RTX 3060 12GB or higher
DeepSeek-R1-Distill-Qwen-32B	32B	~16 GB	NVIDIA RTX 4090 24GB
DeepSeek-R1-Distill-Llama-70B	70B	~35 GB	Multi-GPU setup (e.g., NVIDIA RTX 4090 x2)

Image by [APx](#)

Key notes on VRAM usage:

- **Distributed GPU Setup for Larger Models:** Running DeepSeek-R1-Zero and DeepSeek-R1 requires substantial VRAM, necessitating distributed GPU configurations (e.g., NVIDIA A100 or H100 in multi-GPU setups) for optimal performance.
- **Single-GPU Compatibility for Distilled Models:** Distilled models are optimized to run on single GPUs with lower VRAM needs, starting at just 0.7 GB.

- **Additional Memory Usage:** Extra memory may be consumed for activations, buffers, and batch processing tasks.

## Why Run It Locally?

Sure, the web chatbot and mobile app for DeepSeek are free and incredibly convenient. You don't need to set anything up, and features like DeepThink and web search are baked right in. But there are a few reasons why running it locally might be a better choice:

### Privacy

- When you use the web or app versions, your queries — and any file attachments — are sent to DeepSeek's servers for processing. What happens to that data? We don't know. Running the model locally ensures your data stays on your machine, giving you full control over your privacy.

### Offline Access

- Running the model locally means you don't need an internet connection. If you're traveling, dealing with spotty Wi-Fi, or just prefer working offline, local setups let you use DeepSeek anytime, anywhere.

### Future-Proofing

- DeepSeek's services might be free now, but that's unlikely to last forever. At some point, they'll need to monetize, and usage caps or subscription fees could follow. By running the model locally, you can sidestep these limitations entirely.

## Flexibility

- With the local version, you're not limited to the default setup. Want to fine-tune the model? Integrate it with other tools? Build a custom interface? The open-source nature of DeepSeek R-1 gives you endless possibilities.

## Final Thoughts

At this point, it's still unclear how DeepSeek handles user data. If you're not too worried about data privacy, using the web or mobile app might be the way to go since they're easier to use and offer features like DeepThink and web search. But if you're someone who cares about where your data ends up, running the model locally is a good alternative to consider.

DeepSeek models are designed to run well even on hardware that isn't super powerful. While larger models like DeepSeek-R1-Zero need distributed GPU setups, the distilled versions make it possible to run things smoothly on a single GPU with much lower requirements.

If you don't like using the terminal, you can always add a simple UI with tools like [Gradio](#) or [Chatbox AI](#). I'll write a guide on setting that up in the next article. For now, I hope this post helps you get started. Let me know your thoughts or if you've run into any issues, feel free to share them in the comments.

This story is published on [Generative AI](#). Connect with us on [LinkedIn](#) and follow [Zeniteq](#) to stay in the loop with the latest AI stories.

Subscribe to our [newsletter](#) and [YouTube](#) channel to stay updated with the latest news and updates on generative AI. Let's shape the future of AI together!

---

[Technology](#)[Artificial Intelligence](#)[Llm](#)[Deepseek R1](#)[Ollama](#)

## Published in Generative AI

[Follow](#)

34K Followers · Last published 1 day ago

All the latest news and updates on the rapidly evolving field of Generative AI space. From cutting-edge research and developments in LLMs, text-to-image generators, to real-world applications, and the impact of generative AI on various industries.



## Written by Jim Clyde Monge

[Follow](#)

24K Followers · 2.4K Following

4X Top Writer. Chief Editor at [generativeai.pub](#) | [zeniteq.com](#) | Work with me: [jimclydegm@gmail.com](mailto:jimclydegm@gmail.com)

## Responses (47)



What are your thoughts?

Respond



Remsy Schmilinsky he/him

6 days ago (edited)

...

Thanks for sharing. Here's a little web app I created to interact with it locally, including any model pulled with Ollama: <https://github.com/schmitech/ollama-chat>



161



4 replies

[Reply](#)



David Bui

Jan 24

...

WARNING !

Forget TikTok: the US might need to ban DeepSeek-V3.

International users are unknowingly engaging with a chatbot that's designed to support and disseminate the ideological and political views of the Chinese government.

<https://medium.com/the-generator/deepseek-hidden-china-political-bias-5d838bbf3ef9>



120



5 replies

[Reply](#)



Prince K. Gond he/him

Jan 25

...

I tried this model on my local system.

It's better than chatgpt model.

It's free and open source.



85



2 replies

[Reply](#)

[See all responses](#)

## More from Jim Clyde Monge and Generative AI



 In Generative AI by Jim Clyde Monge 

### AI Can Now Replace Any Actor In A Film

Tencent's open-source AI video generator, Hunyuan, has just integrated Low-Rank...

 Jan 22  704  22



 In Generative AI by Fabio Matricardi

### How a Small Language Model can achieve 100% accuracy

In Context Learning is Underrated—ICL is the secret key to reach performance boosting— ...

 Jan 17  515  4



 In Generative AI by Ritvik Nayak

## 10 Shocking Myths About AI That You Probably Believe...

You'll be Amazed to Find That Some of Them Aren't True

4d ago  371  6



...



 In Generative AI by Jim Clyde Monge 

## DeepSeek Releases Its Own AI Image Generator, Janus-Pro

DeepSeek says Janus-Pro 7B outperforms OpenAI's Dall-E 3 and Stable Diffusion in...

 4d ago  910  30



...

[See all from Jim Clyde Monge](#)

[See all from Generative AI](#)

## Recommended from Medium





In Age of Awareness by Cliff Berg



In Level Up Coding by Dr. Ashish Bania



## They Know a Collapse Is Coming

The CIO of Goldman Sachs has said that in the next year, companies at the forefront will...

Jan 21

11.3K

565



...

## DeepSeek-R1 Beats OpenAI's o1, Revealing All Its Training Secrets...

A deep dive into how DeepSeek-R1 was trained from scratch and how this open...

5d ago

1.1K

28



...

## Lists



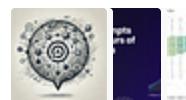
### AI Regulation

6 stories · 679 saves



### ChatGPT

21 stories · 957 saves



### ChatGPT prompts

51 stories · 2527 saves



### Generative AI Recommended Reading

52 stories · 1630 saves



ReplyGuy

### The AI That finds the best places to mention your product online

ReplyGuy finds the perfect conversations to mention your product and drafts suggested replies.

Try it free



Mohit Vaswani

## 6 AI Agents That Are So Good, They Feel Illegal

AI agents are the future because they can replace all the manual work with automation...

Jan 11

2.1K

81



...

In AI Advances by Kenny Vaneetvelde



## Want to Build AI Agents? Tired of LangChain, CrewAI, AutoGen &...

Frameworks like LangChain, CrewAI, and AutoGen have gained popularity by promisin...

Jan 19

1K

22



...



```

195     if (menuItem == null) {
196       handleMenuItemSelection(menuItem);
197     }
198   }
199 }
200
201 child: Dismissible(
202   key: Key(widget.clipId),
203   background: Container(
204     color: Colors.red,
205     alignment: Alignment.centerRight,
206     padding: EdgeInsets.only(right: 16.0),
207     child: Icon(
208       Icons.delete,
209       color: Colors.white,
210     ), // Icon
211   ), // Container
212   direction: DismissDirection.endToStart,
213   onDismissed: (direction) {
214     clipModel.deleteClip(clipId: widget.clipId);
215   },
216   child: Container(
217     color: theme.of(context).primaryColorLight
218   ), // Container
219   padding: EdgeInsets.fromLTRB(0, 0, 0, 0),
220   child: Column(
221     children: [
222       Row(mainAxisSize: MainAxisSize.min, children: [

```

I'll help you add a dismissible widget to make the clip swipeable with a delete action. First, let me check the current implementation of the clip widget.

Analyzed clip\_widget.dart (1-21)

I'll modify the clip widget to make it dismissible with a delete action. I'll wrap the main content in a Dismissible widget and add a delete icon that appears when swiping. Let me update the file:

Edited clip\_widget.dart +19 -2 284 Open diff

I've made the following changes to the clip\_widget.dart file:

- Wrapped the main Container widget with a Dismissible widget
- Added a delete icon using the clip's ID
- Configured the background to show a red container with a white delete icon
- Set the dismiss direction to endToStart (right to left swipe)
- Added an onDismissed callback that calls the existing deleteClip method

Now you can:

In Generative AI by Fabio Matricardi

## DeepSeek-R1? Don't trust the news

Is it really the open-source underdog model that outperform even OpenAI, or is it anothe...

5d ago 524 3



See more recommendations

In Coding Beauty by Tari Ibaba

## This new IDE just destroyed VS Code and Copilot without even...

Wow I never thought the day I stop using VS Code would come so soon...

Jan 17 2.1K 85

