How to Run Deepseek on VSCode

Step 1: Download Ollama

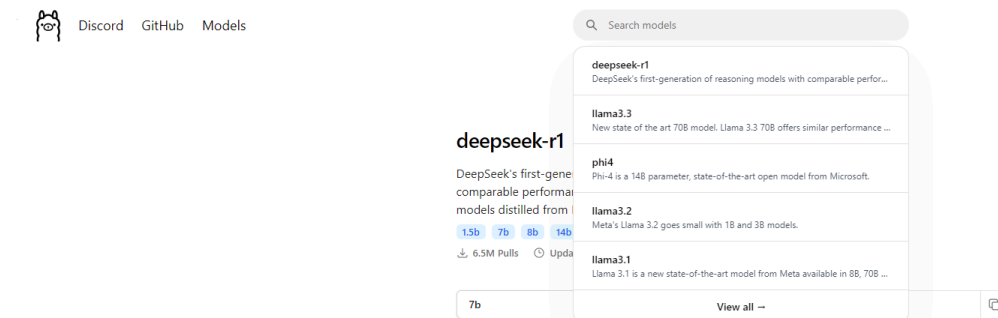Download url: https://ollama.com/download

Step 2: Install and Run Ollama

It does not let you choose where to install it. It automatically installs to the default app directory.

When you see the Ollama in the icon section, it is successfully installed on your computer.

Step 3: Search Deepseek Model on the Ollama Website

Step 4: Find the Best Model for your Computer

Distilled models

DeepSeek team has demonstrated that the reasoning patterns of larger models can be distilled into smaller models, resulting in better performance compared to the reasoning patterns discovered through RL on small models.

Below are the models created via fine-tuning against several dense models widely used in the research community using reasoning data generated by DeepSeek-R1. The evaluation results demonstrate that the distilled smaller dense models perform exceptionally well on benchmarks.

DeepSeek-R1-Distill-Qwen-1.5B

```
ollama run deepseek-r1:1.5b
```

DeepSeek-R1-Distill-Qwen-7B

```
ollama run deepseek-r1:7b
```

DeepSeek-R1-Distill-Llama-8B

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

DeepSeek-R1-Distill-Qwen-14B

```
ollama run deepseek-r1:14b
```

DeepSeek-R1-Distill-Qwen-32B

```
ollama run deepseek-r1:32b
```

DeepSeek-R1-Distill-Llama-70B

```
ollama run deepseek-r1:70b
```

For most personal computers, if you want it to generate response fast, you must choose the 1.5B model.
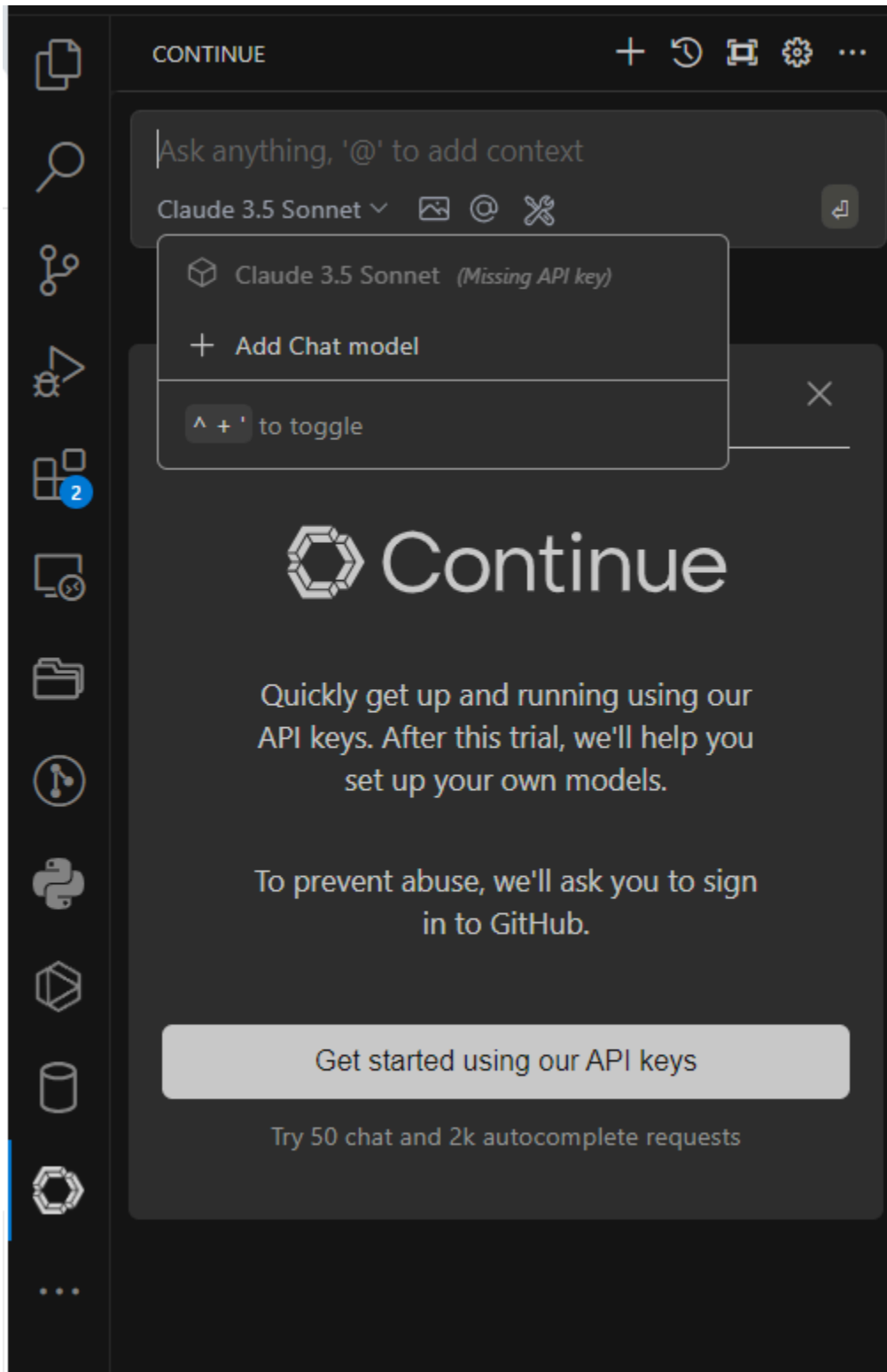
Step 5: Run the above Command in Terminal

```
C:\Users\wykbl>ollama run deepseek-r1:1.5b
pulling manifest
pulling aabd4debf0c8...   6% ▓        ▓  65 MB/1.1 GB   14 MB/s   1m13s
```

Once it is installed, you can talk with it similar to how you will use ChatGPT.

Step 6: Open VSCode and Install Continue from Extension Store
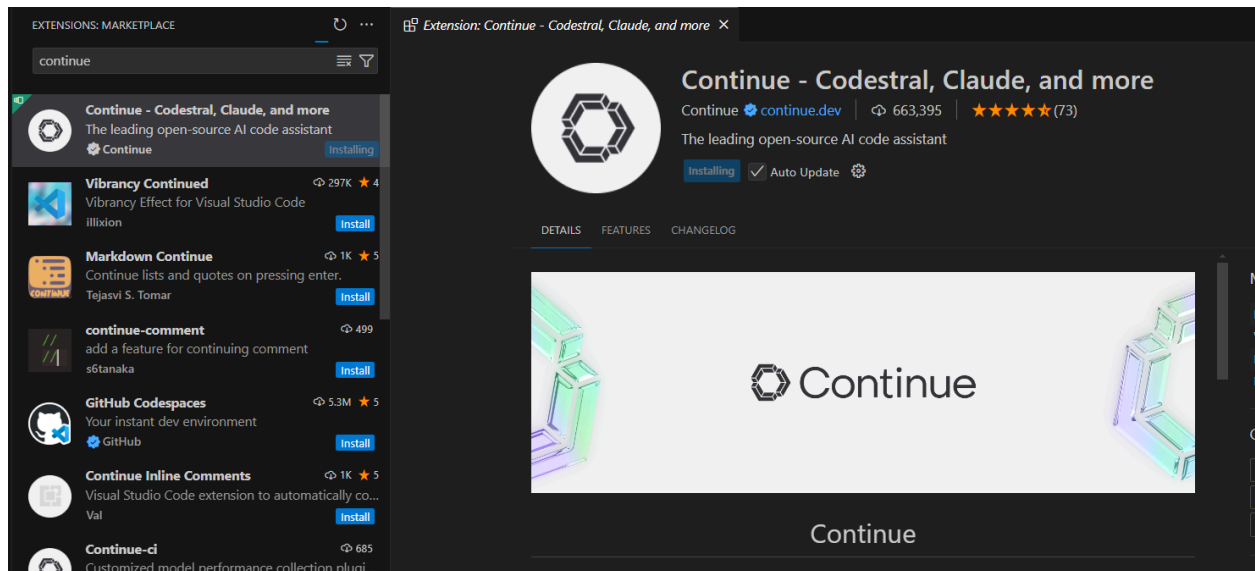
Step 7: Run Continue from the Left Hand Bar and Add a New Chat Model

Step 8: Choose Ollama and Select Autodetect

Step 9: Use it to Write Code Faster