AI STACK SIGNAL

CHECKED NOV 4, 2026 · NOT TESTED BY US

PLAIN-ENGLISH GUIDE · LOCAL AI WITH COLIBRI

RUN A GIANT AI ON YOUR OWN COMPUTER WITH COLIBRI

THE TRICK, THE SETUP, THE CATCH

Colibri is a free, open-source tool that runs very large AI models on the computer you already have. A graphics card is optional. This guide explains how it works, what you need, how to set it up, which model fits which machine, and what to know before you start.

5%

The trick

For each token, the model only uses about 5% of itself, so the rest can wait on your disk.

1

One question

On Windows: double-click one file, pick a model, done. Mac and Linux need a few typed commands.

429 GB

The catch

The giant model is a huge download, and on a 25 GB laptop it is very slow.

23 GB

Start small

Qwen3.6 is a much smaller download. Try that one first.

SOURCE: COLIBRI'S GITHUB README. WE DID NOT RUN IT.

AISTACKSIGNAL.COM

START HERE

01 WHAT IT IS, WHAT YOU NEED

Colibri is a free, open-source program (Apache 2.0) that runs very large open AI models on your own computer. It dates from July 2026 and had 47,548 stars on GitHub on Nov 4.

744B

That's GLM-5.2, by Colibri's count. 744 billion parameters, the numbers inside the model. Other reports say 753 billion. Colibri says it uses about 40 billion for each token, which is about 5%.

THE TRICK

Models this size are built from many smaller parts called **experts**, and only some of them work on each token. So the model doesn't have to fit in memory all at once. The part that's always needed **stays in memory**. The experts **stay on your disk** and are read when the model asks for them.

Colibri also keeps a cache that learns which experts your work uses. Where a part sits changes how fast the answer comes. Colibri says it doesn't change which parts produce the answer.

THREE WORDS YOU'LL SEE

Token	A piece of a word. AI writes one token at a time, so speed is counted in tokens a second.
Memory (RAM)	Your computer's fast short-term space, for whatever it's working on right now. It's small.
Disk	Long-term storage, where your files live. It's big, and much slower than memory.

8 GB
Memory at the very least. 16 GB or more is better.

22 GB
Free disk space for the smallest model.

Online
An internet connection for the download.

Optional
A graphics card. It helps, but you don't need one.

! **We did not test Colibri.** We read its GitHub page and did not install or run it. Every speed in this guide is Colibri's own or its community's.

02 STEP BY STEP SET IT UP

Windows is the one-double-click route. Mac and Linux need a few typed commands. After that, the setup is the same everywhere.

1 Windows: download it

Go to `github.com/JustVugg/colibri` and click Code, then Download ZIP. Unzip it.

2 Windows: double-click START-HERE.bat

It's in the unzipped folder. If Python is missing, it offers to install it for you.

3 Answer one question

Which model? Press Enter to take the recommendation. The setup has already looked at your memory, disk, processor and graphics card, and picks a model that fits.

4 Wait for the download

Stop whenever you like. Run the same thing again and it carries on where it stopped. Then Colibri starts and opens its dashboard in your browser.

MAC (WITH HOMEBREW)

```
xcode-select --install
brew install libomp git python
git clone https://github.com/JustVugg/colibri
cd colibri
./start-here.sh
```

LINUX (UBUNTU AND DEBIAN)

```
sudo apt install git python3 build-essential
git clone https://github.com/JustVugg/colibri
cd colibri
./start-here.sh
```

NEXT TIME

Run **START-HERE.bat** or **./start-here.sh** again. It starts straight away, with no download. To stop it, press Ctrl+C in its window, or close the window.

HANDY EXTRAS

Add **--list** to see every model against your computer, and why one doesn't fit. Add **--dir** and a folder to put models on a bigger disk.

USING AN AI CODING ASSISTANT? COLIBRI'S README SUGGESTS ASKING IT

```
Set up colibri on this machine following docs/AI_SETUP.md from
https://github.com/JustVugg/colibri
```

03 PICK ONE WHICH MODEL FOR WHICH MACHINE

Memory is two numbers. **Below the first, the model won't start.** From the second, it runs as measured. Speeds come from the machine named, not from yours.

A FEW OF COLIBRI'S MODELS

MODEL	GOOD FOR	DOWNLOAD	MEMORY	COLIBRI'S MEASURED SPEED
Qwen3-Coder	Code	19 GB	8 / 18 GB	8.5 to 9.6 tokens a second with every expert in memory, on a desktop with 61 to 64 GB
Qwen3.6	Chat	23 GB	10 / 20 GB	6 tokens a second on the processor of a desktop with 61 to 64 GB
Qwen-Image-2.1	Pictures, non-commercial	33 GB	12 / 18 GB	One 768x512 picture in 2 min 40 s on an 8-core processor
DeepSeek V4 Flash	Large chat model	167 GB	16 / 32 GB	0.93 tokens a second with 32 GB, processor only
GLM-5.2	The giant one	429 GB	16 / 24 GB	One token every 10 to 20 seconds on a 25 GB laptop, cold. 1.83 a second on a 128 GB machine. 9.0 to 9.2 on six RTX 5090 cards

WHAT THOSE SPEEDS FEEL LIKE OUR SUMS

SPEED	A 60-TOKEN REPLY TAKES
6 tokens a second	10 seconds
One token every 10 to 20 seconds	10 to 20 minutes

OUR ADVICE: START SMALL

Take the setup's recommendation, or try **Qwen3.6** first. It's the smaller download, so you find out how your computer copes before you give up 429 GB of disk.

YOUR MACHINE WILL DIFFER

On a laptop with a Core i7-1355U, Qwen3.6 did 2.1 tokens a second on the processor. On a desktop with an RTX 3070 graphics card, it did 30. So 6 is one machine's number, not a promise.

"Cold" means nothing was cached yet. Colibri says it gets faster the more you use it, because its cache learns which experts your work needs. There are more models in the README tables.

04 KNOW BEFORE YOU START
THE CATCH, AND OUR SOURCES

- [ ] **The downloads are big**
  23 GB for Qwen3.6, 429 GB for GLM-5.2. Check your free disk space first, and expect a long wait.

- [ ] **Your disk sets the speed on large models**
  Large models read from the disk as they write. Colibri says a fast NVMe drive helps most. NVMe is a fast kind of solid-state drive. For small models, memory sets the speed.

- [ ] **Giant models can be painfully slow**
  Colibri calls its 25 GB laptop speed "the honest floor": one token every 10 to 20 seconds.

- [ ] **Check each model's licence**
  Colibri itself is Apache 2.0, but each model keeps its own licence. Colibri says GLM-5.2 is under MIT and Qwen-Image-2.1 is for non-commercial use only.

- [ ] **Answer quality is not covered here**
  We make no claim about how good any of these models are. Try one on your own work and judge it yourself.

- [ ] **We did not test it**
  We read the README on Nov 4, 2026. We did not install or run Colibri, so we can't say how it behaves on your computer.

SOURCES (CHECKED NOV 4, 2026)

- Colibri GitHub README: github.com/JustVugg/colibri
- Model size: 744B is Colibri's figure. Other reports say 753B. We did not read the official model card.

IF SOMETHING GOES WRONG

The download stopped	Run the same thing again. It carries on from what's already on your disk.
"needs N GB free"	Add --dir and a folder on a bigger disk.
Anything else	Colibri asks you to open an issue on GitHub with the last lines the setup printed.

Software, models and speeds can change. AI Stack Signal isn't affiliated with Colibri or any model maker.

aistacksignal.com